THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Fabon DZOGANG

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Représentation et apprentissage à partir de textes pour des informations émotionnelles et pour des informations dynamiques.

soutenue le 18 juillet 2013

devant le jury composé de :

Maria Rifqi	Directrice de thèse
Eyke Hüllermeier	Rapporteur
Pascal Poncelet	Rapporteur
Bernadette Bouchon-Meunier	Examinatrice
Carl Frélicot	Examinateur
Catherine Gouttas	Examinatrice
Mohamed Nadif	Examinateur
Marie-Jeanne Lesot	Encadrante
Christophe Marsala	Encadrant

Résumé

Les travaux de cette thèse portent sur les problématiques liées à la représentation et à l'apprentissage à partir de textes à la fois pour des informations émotionnelles et pour des informations dynamiques. L'information bas niveau, extraite des documents au travers des mots et des groupes de mots, est mise en correspondance avec des concepts haut niveau par extraction automatique de connaissances. Pour ce faire les choix de représentation sont essentiels dans le contexte de l'apprentissage automatique qui doit de plus tenir compte des particularités des descripteurs textuels.

Dans une première partie, nous étudions un apprentissage pour des informations subjectives puisque émotionnelles. Le *fossé sémantique* entre l'information bas niveau et des concepts émotionnels est plus important que pour des concepts traditionnellement thématiques. Aussi, nous proposons d'étudier de nouveaux descripteurs bas niveau pour enrichir les choix classiques de représentation et mieux modéliser les subtilités du langage comme les changements d'intensité ou la négation. Un enrichissement sémantique des textes permet en outre de guider l'apprentissage réalisé en rapprochant l'information bas niveau de concepts pour lesquels une sémantique émotionnelle est connue a priori. Nous proposons à ce titre un modèle de représentation des émotions qui repose sur la théorie des sous-ensembles flous pour répondre aux imprécisions par la gradualité ainsi que sur les travaux en psychologie pour décrire finement les émotions. De plus, selon les choix de représentation effectués, nous proposons d'étudier une tâche de discrimination de la charge émotionnelle des documents sur un ensemble d'étiquettes pré-définies, ainsi qu'une tâche de caractérisation fine de cette charge. Pour la première nous mettons en œuvre un apprentissage supervisé et nous proposons d'étudier des combinaisons de représentations pour décrire les documents. Pour la seconde nous proposons, dans un cadre d'apprentissage non supervisé, d'exploiter un enrichissement sémantique associant aux mots du langage un ensemble de mesures évaluant notamment leur subjectivité, leur intensité ou leur polarité.

Dans une seconde partie, nous considérons le dynamisme inhérent à l'information : sur Internet, un ensemble de sources publient fréquemment des documents. Lorsqu'elles sont représentées par l'information qu'elles produisent, ces sources s'organisent en communautés qui évoluent dans un espace de représentation parcimonieux et dynamique. Pour identifier ces communautés nous proposons de formuler une tâche de clustering incrémental où les sources, comme l'information étudiée et les communautés extraites évoluent. L'étude successive du dynamisme sur ces trois composantes nous ont conduit à proposer une méthode pour l'identification automatique de sources sur Internet, un algorithme de clustering adapté aux spécificités des représentations textuelles qui extrait de toute l'information celles pour lesquelles les sources s'organisent plus naturellement en communautés homogènes ; et les *threads d'information*, définis comme des périodes durant lesquelles une communauté est associée à une remarquable stabilité sémantique. L'ensemble des méthodes proposées pour analyser ces trois composantes est mis en œuvre sur un corpus de données réelles que nous avons constitué à partir des publications de la presse française.

Mots-clefs : apprentissage automatique, fouille de textes, dynamisme de l'information, informatique émotionnelle, clustering, théorie des sous-ensembles flous, données parcimonieuses, identification de sources, identification de communautés, combinaisons de représentations, sélection de descripteurs.

Summary

This thesis is concerned with the matter of data representation and concept learning from texts for both emotional and dynamical information. Generally speaking automatic knowledge extraction consist in mapping low level information carried through the words and the phrases extracted from documents to higher level concepts. From a machine learning perspective, it is then necessary to give special concern to the choice of the right description for this textual data, it is also important to take account of the many particularities of textual features.

In a first part, we study the learning of emotional hence subjective information. The *semantic gap* between low level information and emotional concepts is deeper than it is for more traditionnal thematic concepts. Also, we propose to study new features aimed at enriching classical representations of the data in order to capture the nuances provided in natural languages as exemplified for instance by intensity modifiers or the negation. Besides, to better lead the learning algorithm we consider integrating semantic content to this representation in order to close the gap by bringing low level information closer to higher level concepts for which an emotional semantic is known beforehand. To that aim we propose a model to describe emotions for their automatic processing in texts, it relies on both the theory of fuzzy sets to match language's imprecision with graduality and classical results in psychology to describe emotions finely. Furthermore, depending on the choices made for describing both emotions and documents, we propose to study a concept discrimination task for the classification of documents' emotional content on a set of predefined emotion labels, and a task aimed at finely characterizing this content. The former one is performed on a supervised setting : we propose to study and compare combined representations for the data in order to better describe documents. Regarding the latter one, we propose, in an unsupervised manner, to exploit semantic resources which bind generic words to emotional scores assessing among others their characteristic subjectivity, intensity or polarity.

In a second part, we consider the dynamism inherent in information : a set of sources publishing documents frequently over the Internet are described by the information they produce over time. As time goes they organize themselves in evolving communities described in sparse and dynamical feature spaces. To detect these communities we propose to formulate an incremental clustering task where sources, as well as the information they produce and the extracted communities evolve. The study of this dynamism over these three components led us to the proposal of three original methods. The first one deals with the automatic detection of homogeneous information sources over the Internet. The second one is a clustering algorithm adapted to the particularities of textual features, aimed at extracting from all the available information, the information over which sources cluster more naturally in homogeneous communities ; and the latter one deals with *information threads*, defined for a community as a time interval during which it is associated with a notably stable semantic. All of theses original methods are put in application on a corpus composed of real data we have collected from French news over the Internet.

Keywords : machine learning, text mining, information dynamism, emotional information, clustering, fuzzy sets theory, sparse data, sources detection, communities detection, features combinations, features selection.

Table des matières

Introduction

Contexte

Une source interprète une information, l'exprime dans un langage et la formule dans un document transmis à un récepteur, qui reconnaît ce langage, extrait du message son information, l'interprète à son tour avant de devenir lui-même source d'une information nouvelle. Pour Shannon (1948), l'information est dénuée de sens, elle constitue une succession de symboles entendue de sa source et de son récepteur, elle peut être quantifiée, compressée, elle est un signal bas niveau.

La problématique d'extraction de connaissances haut niveau à partir d'informations bas niveau a pour objectif de réduire le *fossé sémantique* qui sépare ces deux niveaux.

Elle est au cœur de l'apprentissage automatique qui vise à l'extraction de connaissances à partir de données non structurées : le choix d'une représentation pour décrire les données est essentiel et détermine, dans une grande part, le succès de l'apprentissage réalisé puisque le passage d'une information bas niveau à un concept haut niveau dépend entièrement des descripteurs présentés en entrée. Dans le cas des textes, les documents sont généralement représentés comme des sacs de mots dans des espaces vides en grande dimension. Ce passage constitue alors un défi particulier pour lequel il est nécessaire d'adapter les mesures de comparaison définies, les algorithmes d'apprentissage utilisés ou les stratégies de réduction de bruit considérées.

Il n'en est pas moins que le choix des descripteurs constitue un problème central : dans le cas des textes en particulier, ce passage est rendu difficile du fait de l'ambiguïté et des subtilités inhérentes au langage. Ainsi les choix classiques de représentation posent parfois problème et un enrichissement sémantique des documents est considéré pour associer aux informations bas niveau, un sens particulier, rapprochant du haut niveau et guidant l'apprentissage réalisé.

Cet apprentissage peut prendre deux formes : quand la sémantique recherchée est connue, un apprentissage supervisé vise à déterminer une association entre l'information extraite et des concepts haut niveau explicités par des étiquettes associées aux données. Lorsque la sémantique n'est pas connue, un apprentissage non supervisé vise à organiser cette information pour une interprétation postérieure des structures haut niveau identifiées.

Cette thèse considère deux problèmes particuliers d'extraction de connaissances à partir de textes, d'une part le cas d'informations émotionnelles, ou plus généralement subjectives, et d'autre part le cas d'informations dynamiques. Pour chacune, on considère à la fois le problème de la représentation et celui de l'apprentissage.

Informations émotionnelles

La première partie de ce travail de thèse se place dans le cadre du domaine de l'affective computing (Picard et al., 2001), qui vise de façon générale à prendre en compte les

émotions dans les interactions homme/machines. Il regroupe différentes problématiques, parmi lesquelles, la simulation d'agents affectifs, qui repose sur l'étude des mécanismes émotionnels de par des règles de déclenchement, ou la reconnaissance automatique d'émotions, décrites par un ensemble d'émotions basiques ou de mesures évaluant leur charge. Pour cette seconde tâche, les émotions peuvent être reconnues à partir de différents types de signaux, comme par exemple les signaux physiologiques, les expressions faciales ou les textes rédigés en langue naturelle. Les travaux de thèse présentés dans ce manuscrit se placent dans ce dernier cas : nous étudions plus généralement les problématiques liées à l'identification dans les textes de concepts subjectifs tels que les opinions ou les émotions.

Du fait de la complexité de ces concepts, ainsi que des particularités de l'apprentissage à partir de textes, ces problématiques sont considérées au niveau de la représentation des documents ainsi qu'à celui de la tâche de reconnaissance définie, qui dépend en partie de la représentation faite des émotions.

Représenter une information émotionnelle

Etant donné la subjectivité des émotions, le fossé sémantique entre l'information bas niveau (le vocabulaire d'un document) et les concepts étudiés est plus important que pour des concepts traditionnellement thématiques : ainsi, de nouveaux descripteurs sont considérés pour décrire les textes.

Une première approche consiste à ajouter aux mots simples, employés de manière traditionnelle, des termes plus enclins à constituer des marqueurs d'émotions. La ponctuation pour laquelle le point d'exclamation est le descripteur le plus manifeste en est un exemple. De même, afin de tenir compte des subtilités du langage comme la négation ou l'intensité, les combinaisons de mots apparaissant de manière consécutive dans les documents, comme les p-grammes, sont également exploitées.

Ces nouveaux descripteurs amènent à de nouvelles problématiques comme celle de la fusion des informations qu'ils extraient de manière isolée. Dans ce contexte le passage d'un signal bas niveau à une information haut niveau nécessite, de plus, de revoir les méthodes et les pré-traitements utilisés de manière classique pour éliminer la redondance ou le bruit dans une information textuelle.

Une seconde approche pour représenter les documents repose sur un enrichissement sémantique de ces derniers. Un lexique sémantique peut par exemple associer un état émotionnel à chacun des mots composant un vocabulaire pré-compilé. Ces ressources sont mises en œuvre afin d'extraire des textes une information dont la sémantique émotionnelle est connue.

Nous proposons dans un premier temps l'exploitation isolée de descripteurs bas niveau, puis de descripteurs sémantiques. Dans un second temps, nous envisageons une combinaison de ces deux types de représentation. Plusieurs modes de combinaison peuvent à nouveau être définis : nous considérons une stratégie pour laquelle les descripteurs sont concaténés pour produire une information à la fois fidèle aux documents étudiés et proche des concepts émotionnels.

De plus, étant donné la subtilité du langage, les expressions écrites des émotions peuvent s'avérer complexes puisque nuancées et imprécises, nous proposons dans ce cadre un modèle de représentation des émotions qui repose sur la théorie des sous-ensembles flous pour répondre aux imprécisions par la gradualité, ainsi que sur des travaux en psychologie et en linguistique pour décrire finement les émotions dans les textes.

Extraire une émotion d'un texte

Différentes tâches de reconnaissance peuvent être considérées, celles-ci dépendent à la fois de la représentation faite des documents et de la description faite des émotions.

Lorsqu'une sémantique émotionnelle particulière est recherchée dans les documents, elle est généralement représentée comme une étiquette issue d'une catégorisation des émotions. Reconnaître l'émotion d'un document signifie alors associer à sa charge émotionnelle l'une des émotions du modèle considéré et se formule comme une tâche de classification. Pour ce faire un apprentissage supervisé est mis en œuvre : il identifie les descripteurs les plus discriminants pour décrire chacune des émotions considérés. Traditionnellement, la représentation faite des documents est au plus proche du vocabulaire employé. Etant donné le fossé sémantique important entre ce dernier et les émotions, des enrichissements sémantiques sont également considérés. Ces méthodes nécessitent un corpus de documents préalablement étiquetés à partir duquel est réalisé cet apprentissage : nous exploitons deux corpus étiquetés selon des étiquettes émotionnelles.

Nous considérons également une tâche de caractérisation dans un cadre d'apprentissage non supervisé. Les émotions recherchées ne sont pas connues a priori, la charge émotionnelle d'un document est alors caractérisée finement : un ensemble de mesures évalue les propriétés de cette dernière qui peut alors se distinguer de par son intensité, sa subjectivité, ou encore au travers de sa polarité. Pour cette tâche, la représentation des documents repose généralement sur un enrichissement sémantique basé sur un représentation fine et graduelle des émotions.

Informations dynamiques

Une seconde partie de nos travaux est dédiée au dynamisme inhérent à l'information. Une source qui produit des documents à intervalles de temps fréquents émet de nouvelles informations en continu, parmi celles-ci certaines sont reprises et enrichies pour être retransmises sous une forme nouvelle. Au cours du temps, ces regroupements d'information, vus comme des thématiques, se créent, évoluent selon les informations qui les alimentent puis disparaissent et laissent place à de nouvelles thématiques. Ainsi, un ensemble de sources s'organise en une communauté gouvernée par une thématique fédératrice. Au cours du temps, de nouveaux individus rejoignent cette thématique et intègrent la communauté, d'autres changent d'intérêt et l'abandonnent. L'information est en mouvement constant, elle transite de sources en sources, elles-mêmes en perpétuelle activité, le réseau Internet catalyse ce dynamisme. Nos travaux s'organisent autour de trois axes pour l'étudier : les sources qui produisent l'information étudiée, les communautés qui se forment au gré des intérêts partagés par les sources, et les thématiques qui résultent de l'information produite par les sources et qui expliquent la formation des communautés.

Identifier une source d'information

Nous nous intéressons d'abord à la définition de sources sur Internet. Ces dernières sont comprises comme des producteurs d'information qui publient fréquemment des documents accessibles des autres sources. Un média d'information est par exemple une telle source, elle publie des articles en rapport avec l'actualité et elle couvre de nombreux sujets variés. L'information qu'elle produit est homogène quand elle est fidèle à un ensemble réduit de thématiques, elle est hétérogène et nécessite un raffinement quand il n'est pas possible de situer ses thématiques d'intérêt.

Nous proposons deux méthodes pour effectuer ce raffinement, elles consistent toutes deux à exploiter la structuration des urls associées aux documents publiés sur Internet afin de décomposer une source hétérogène en une hiérarchie de sources plus homogènes. Elles se distinguent de par les propriétés imposées à cette hiérarchie.

Reconnaître des communautés de sources dynamiques

Quand un ensemble de sources émet une information proche, elles créent une communauté autour d'une thématique commune. En représentant les sources par l'information qu'elles produisent, nous proposons de reconnaître ces communautés en formulant un problème de clustering de sources dynamiques.

Par opposition au cadre classique, pour ce problème, les données sont très parcimonieuses car les sources sont décrites d'après le vocabulaire qu'elles emploient pour commenter l'actualité. Elles sont de plus dynamiques puisque l'information émise évolue en continu. Pour cette tâche, l'affectation d'une source à une communauté est décidée à la fois par l'information qu'elle a émise dans le passé et par celle qu'elle émet à l'instant courant : un tel partitionnement encourt de nombreux changements dans le temps et nécessite des méthodes adaptées au dynamisme des données. En particulier, les thématiques les plus fédératrices pour un ensemble de sources ne représentent qu'une infime partie de toute l'information produite par ces dernières. Ainsi, les communautés identifiées par un algorithme de clustering classique ne permettent pas une extraction précise des thématiques quand un fort dynamisme régit les partitions.

Nous proposons un algorithme de clustering adapté à ce contexte, qui consiste à extraire, de toutes ces informations, celles pour lesquelles les sources s'organisent plus naturellement en communautés homogènes.

Extraire une thématique en environnement dynamique

Le dynamisme qui régit les partitions peut à nouveau être observé à différents niveaux, il peut être constaté au travers des changements de thématiques du fait d'une évolution ou d'un renouvellement de l'information globalement émise par la population d'une communauté, il peut également être observé au travers des transitions effectuées par les sources entre les communautés.

Nous proposons de nous intéresser aux *threads d'information* que nous définissons comme des thématiques remarquablement stables durant un intervalle suffisamment long, nous proposons d'extraire des partitions, les communautés de sources associées, pour un temps, à une remarquable stabilité sémantique. En particulier, au travers d'une mise en œuvre expérimentale de l'ensemble des méthodes proposées dans cette seconde partie, nous mettons en évidence les threads associés aux publications de la presse française durant la fin d'année 2012.

Organisation

La thèse est organisée en deux parties portant respectivement sur les informations émotionnelles et sur les informations dynamiques, précédées d'un chapitre introductif.

Le chapitre 1 fait état des méthodes classiques pour la représentation de données textuelles en apprentissage : nous examinons successivement les descripteurs et les mesures de comparaison employés pour les documents, les méthodes de réduction de dimension appliquées dans le cadre de la représentation textuelle, ainsi que trois stratégies de fusion pour combiner des représentations multiples.

La partie I porte sur les informations émotionnelles. Le chapitre 2 présente un état de l'art sur les méthodes en apprentissage pour des informations émotionnelles. Le chapitre 3 décrit une approche ainsi qu'une instanciation de la méthode proposée pour la discrimination de concepts émotionnels à partir de descripteurs bas niveau. Nous proposons une approche différente pour étudier une tâche différente au chapitre 4 : un enrichissement sémantique des documents permet de les décrire comme des nuages de points dans un espace sémantique multidimensionnel et par la suite de caractériser finement leur contenu émotionnel. Enfin, nous décrivons le modèle proposé pour représenter finement les émotions en vue de leur analyse automatique dans les textes au chapitre 5. De plus, nous détaillons sa mise en œuvre expérimentale dans le cadre du projet DoXa présenté dans ce même chapitre.

Dans la partie II, nous considérons les informations dynamiques. Le chapitre 6 introduit une formalisation de la tâche de clustering de sources dynamiques considérée et présente ses problématiques ainsi que ses définitions. Deux méthodes pour l'identification de sources homogènes sur Internet sont présentées au chapitre 7. L'algorithme proposé pour le clustering de données textuelles très parcimonieuses est décrit au chapitre 8. Finalement, l'ensemble des méthodes proposées sont mises en œuvre sur des données réelles au chapitre 9.

Chapitre 1

Représentation des données textuelles en apprentissage

La représentation des données est un problème central en apprentissage automatique : afin de caractériser au mieux les concepts cibles il est nécessaire de modéliser à bien l'information étudiée. Les descripteurs employés de manière classique sont de deux types : d'une part les descripteurs bas niveau représentent une information au plus proche des données étudiées, d'autre part les descripteurs sémantiques visent à tenir compte de connaissances supplémentaires, qui reflètent souvent d'une expertise sur le domaine d'étude. Ces deux types de représentation présentent chacune leurs particularités ; dans certains cas, leur fusion permet de modéliser une information de manière encore plus précise. L'information extraite des données étudiées constitue souvent un signal bruité, une réduction des descripteurs employés vise alors à épurer la représentation faite des données. Un autre axe d'étude intimement lié à celui de la description des données est celui de leur comparaison qui dépend de la nature de l'information extraite. De nombreuses décisions sont ainsi effectuées en amont de tout processus d'apprentissage : la représentation des données est une problématique à la fois dense et riche, à laquelle de nombreux domaines de recherche à part entière, envisagés ici sous l'angle du texte, peuvent être rapportés.

Le contenu de ce chapitre est structuré en trois parties.

La section 1.1 présente les descripteurs utilisés de manière classique pour modéliser une information textuelle. D'une part la représentation en *sac de mots* vise à décrire les documents d'après les mots qui les composent, d'autre part des enrichissements, syntaxiques ou sémantiques, tiennent spécifiquement compte de la structure du langage ou de connaissances expertes sur le domaine d'étude. Une fois les documents décrits se pose le problème de leur comparaison : nous présentons dans un premier temps les mesures de distance traditionnelles, puis nous détaillons le cas des fonctions de similarité et nous étudions en particulier leur compatibilité avec la nature spécifique des descripteurs textuels.

Comme décrit à la section 1.2, un système d'apprentissage est généralement confronté à une information bruitée. Une réduction des dimensions utilisées permet alors d'affiner la description faite des données. Nous présentons des méthodes numériques, non exclusives au cas du texte, étudiées de manière classique pour des représentations en grande dimension. La sélection de descripteurs consiste à identifier puis éliminer les dimensions non pertinentes pour décrire les concepts cibles, tandis que la construction de descripteurs repose sur l'extraction de nouveaux descripteurs à partir de la représentation faite originellement des données.

Enfin dans certains cas une combinaison de différentes représentations permet d'extraire, des données, une information encore plus pertinente. La section 1.3 présente trois

approches classiques pour réaliser cette fusion. En particulier, la fusion anticipée repose sur une concaténation des descripteurs disponibles, la fusion tardive consiste à agréger les décisions prises individuellement sur chacune des représentations, et la fusion intermédiaire constitue un compromis pour lequel ce sont les mesures de comparaison associées à chacune des représentations qui sont combinées.

1.1 Descripteurs et similarités pour le texte

Dans un premier temps nous détaillons la représentation, dite en sacs de mots, reposant sur un dictionnaire lui-même extrait du corpus d'étude : chacun des documents est représenté d'après les mots qui le composent. Dans un second temps nous présentons des enrichissements syntaxiques et sémantiques qui intègrent des connaissances supplémentaires pour la représentation. Les mesures de comparaison employées sont intiment liées à la nature de l'information obtenue. Nous présentons enfin les mesures de distance classiques en apprentissage, puis nous considérons le cas des fonctions de similarité et nous étudions en particulier leur compatibilité avec la nature particulière des descripteurs textuels.

1.1.1 Descripteurs bas niveau

De nombreuses représentations distinctes ont été étudiées dans la littérature, elles diffèrent principalement selon le problème considéré et les hypothèses effectuées sur les données. Pour modéliser la nature séquentielle du langage une approche consiste par exemple à représenter un document comme une séquence de mots. Cette représentation constitue une information riche qui tient compte de l'interdépendance des mots et qui permet de décrire finement les documents étudiés. Néanmoins, dans certains cas cette information s'avère trop complexe pour le problème considéré, une autre approche consiste alors à modéliser ces interactions de manière locale, en représentant une phrase par un arbre de dépendance. Lorsque les interactions sont de nature syntaxique par exemple, il s'agit d'arbres de dépendance syntaxique et l'information obtenue constitue un ensemble ou une succession de groupes syntaxiques.

Ici nous considérons la représentation, dite de sacs de mots, bien établie en apprentissage. Elle consiste à extraire de l'ensemble des documents étudiés, un dictionnaire dont les entrées sont des mots, puis à décrire un document en réalisant un comptage avec pondération éventuelle des entrées obtenues. Selon le mode de constitution du dictionnaire, comme décrit dans la suite, il est possible d'ajuster la complexité de l'information obtenue.

1.1.1.1 Dictionnaire : espace de description

Mots uniques Un document est un ensemble de paragraphes composés de phrases, elles-mêmes composées de mots. Il est d'usage en apprentissage sur les textes de considérer le mot comme l'unité d'information atomique [1]. Etant donné un corpus $\mathcal{D}$ composé de n documents, une représentation naturelle pour décrire un document $d \in \mathcal{D}$ consiste à exploiter l'ensemble des mots uniques composant le corpus. Notons $\mathcal{V}$ cet ensemble, éventuellement enrichi des marqueurs de ponctuation. Le document d peut alors représenté comme un vecteur $\boldsymbol{x}$ dont chacune des composantes correspond à une entrée de $\mathcal{V}$. Autrement dit,

1. Il existe bien sûr de nombreuses exceptions, notamment dans le domaine de l'identification des langues ou dans celui de la traduction où il est alors préférable de modéliser l'information au niveau des groupes de lettres composant les mots.

ordre	$\mathcal{V}$	$\lvert\mathcal{V}\rvert$
$p = 1$	{hier} {pas} {mauvaise} {journée}	4
$p = 2$	{hier pas} {pas mauvaise} {mauvaise journée}	3
$p = 3$	{hier pas mauvaise} {pas mauvaise journée}	2
$p = 4$	{hier pas mauvaise journée}	1

TABLE 1.1 – Unigrammes, bigrammes, trigrammes et quadrigrammes pour la phrase « *hier, pas mauvaise journée !* ». L'effet de la négation n'est modélisé qu'à partir de l'ordre $p = 2$. L'ambiguïté n'est levée qu'à partir de l'ordre $p = 3$.

.

un dictionnaire définit un espace $\mathcal{X}$ dont chacun des axes correspond à un mot unique du corpus : il s'agit de l'*espace de description* des documents, de plus le vecteur $\boldsymbol{x} \in \mathcal{X}$ est la représentation en sacs de mots associée au document d. Dans la suite, nous appelons ce vecteur, le *vecteur de description*, pour un dictionnaire composé de m entrées, nous appelons de plus la matrice X de dimension $n \times m$ qui décrit chacun des n documents du corpus, sa *matrice de représentation*.

Combinaisons de mots : p-grammes Dans certains cas le problème considéré nécessite une information plus riche, des descripteurs plus complexes sont alors employés. Un p-gramme, ou un gramme d'ordre p, est une suite consécutive (dans le corpus) de p mots uniques. A l'ordre $p = 1$, on parle d'unigrammes, à l'ordre $p = 2$ de bigrammes, à l'ordre $p = 3$ de trigrammes.

L'extraction des p-grammes, pour un ordre $p > 1$, sur un corpus représente une méthode simple pour modéliser le contexte d'apparition des mots. En effet, tout mot composant un p-gramme peut être vu comme un unigramme associé à son contexte d'apparition de taille $p - 1$: si deux documents partagent un tel descripteur alors ils partagent un unigramme pourvu du même contexte d'apparition.

Un document de taille n contient $n - p + 1$ p-grammes, le nombre de descripteurs décroît donc en fonction de l'ordre p. Pour de très grands ordres, un descripteur peut être une phrase voire un paragraphe ou même en cas extrême, un document. Il n'est évidemment pas intéressant de travailler sur de tels ordres : dans l'espace de description, chacun des documents ne serait alors similaire qu'à lui même. En pratique, il n'est que très rarement intéressant de dépasser l'ordre $p \approx 10$, et bien souvent $p \leq 3$ est suffisant.

Dans le tableau 1.1, nous avons représenté différents dictionnaires qui ne tiennent pas compte de la ponctuation pour la phrase « *hier, pas mauvaise journée !* ». Pour des unigrammes, le dictionnaire est composé de mots-clefs dépourvus de tout contexte. Les bigrammes introduisent un contexte de taille 1 et permettent de modéliser l'action de la négation. En revanche, ils introduisent aussi une ambiguïté quant à l'information portée : le descripteur *mauvaise journée* prête en effet à confusion. Ce n'est qu'à l'ordre 3 que les trigrammes offrent un contexte de taille suffisante pour modéliser, sans ambiguïté, l'effet de la négation. Cet exemple n'est bien sûr qu'une illustration ; en pratique le choix de l'ordre est principalement motivé par la généricité des descripteurs résultants : sur de petits corpus de documents, il existe peu de chance d'observer des groupes de mots génériques et les unigrammes sont préférés sans quoi l'on s'expose à un risque de sur-apprentissage. Sur de plus grands corpus, il est possible d'introduire une information contextuelle en montant à des ordres plus élevés.

Filtrage des entrées L'espace de représentation associé à un dictionnaire de descripteurs bas niveau est un espace presque vide, en très grande dimension. En effet le vocabulaire utilisé dans chacun des documents ne représente qu'une faible partie du vocabulaire utilisé globalement sur l'ensemble du corpus. Un filtrage consiste généralement à éliminer les entrées non pertinentes du dictionnaire et ainsi à extraire du corpus une information qui présente moins de bruit. Les méthodes de filtrage de dictionnaire sont décrites dans la section 1.2 qui les regroupe avec les méthodes générales de réduction des dimensions de l'espace de description.

1.1.1.2 Méthodes de comptage

Dans l'espace de description $\mathcal{X}$, un document d est décrit par un vecteur $\boldsymbol{x}$ dont chacune des composantes correspond à une entrée du dictionnaire $\mathcal{V}$. Nous présentons trois schémas de comptage utilisés traditionnellement pour caractériser la position de $\boldsymbol{x}$ dans $\mathcal{X}$. Dans la suite nous notons m le nombre d'entrées du dictionnaire.

Représentation binaire Une manière simple de représenter un document dans l'espace de représentation est d'indiquer l'ensemble des descripteurs pour le décrire. Pour la représentation binaire, pour tout j dans l'intervalle $[1..m]$ on a :

$$x_j = \begin{cases} 1 \text{ si le descripteur } \mathcal{V}_j \text{ est présent dans } d \\ 0 \text{ sinon} \end{cases}$$

La représentation binaire consiste ainsi à décrire un document d'après l'ensemble des mots uniques qui le composent.

Méthodes de pondération D'autres schémas de comptage permettent de tenir compte de l'importance des descripteurs pour un document : le schéma fréquentiel consiste par exemple à pondérer la $j^{\text{ème}}$ composante du vecteur de représentation par la fréquence d'apparition du descripteurs correspondant dans le document. Notons $\text{occ}(d, \mathcal{V}_j)$ le nombre d'occurrences du descripteur $\mathcal{V}_j$ dans le document d, pour le schéma fréquentiel on a pour tout j dans $[1..m]$:

$$x_j = \frac{\text{occ}(d, \mathcal{V}_j)}{\sum_{l=0}^{m} \text{occ}(d, \mathcal{V}_l)}$$

La pondération tf/idf (pour *term frequency/inverse document frequency*) consiste, elle, à atténuer l'influence de descripteurs trop fréquents dans le corpus. Elle est calculée en multipliant les scores de fréquence dans le document par une mesure du score de fréquence dans le corpus. Pour tout j dans $[1..m]$, elle se mesure par exemple comme :

$$x_j = \frac{\text{occ}(d, \mathcal{V}_j)}{\sum_{l=0}^{m} \text{occ}(d, \mathcal{V}_l)} \times \frac{|\mathcal{D}|}{\left|\{d' \in \mathcal{D}, \text{occ}(d', \mathcal{V}_j) > 0\}\right|}$$

Le choix de la fonction de pondération varie selon les tâches d'apprentissage. L'intérêt du schéma tf/idf est de pondérer l'importance des descripteurs par leur pertinence, définie en fonction de leur potentiel pouvoir de discrimination dans le corpus $\mathcal{D}$. Ici, un descripteur discriminant est entendu comme une dimension de l'espace sur laquelle les concepts étudiés sont bien séparés.

Il faut noter que lorsque le corpus étudié est de petite taille il existe peu de différence avec le schéma fréquentiel. De même, lorsque les documents étudiés sont courts (par exemple 100 mots), les schémas fréquentiels et binaires sont souvent très similaires. Selon la tâche d'apprentissage, il peut alors s'avérer judicieux de considérer la méthode de pondération la plus simple et la plus rapide à calculer. Par ailleurs, comme décrit au chapitre 2, dans certains cas la pondération des descripteurs ne présente que peu d'intérêt et peut même parfois dégrader l'information extraite des documents pour le problème considéré.

1.1.2 Enrichissements

Pour un algorithme d'apprentissage les descripteurs bas niveau ne sont qu'un ensemble de symboles dénués de toute sémantique, il s'agit du problème du fossé sémantique (*semantic gap*), rencontré dans de nombreuses modalités y compris celle du texte. Les enrichissements, présentés sous la forme de ressources supplémentaires, visent alors à réduire le fossé sémantique en constituant des dictionnaires de plus haut niveau que les mots du langage, dans certains cas, plus proches des concepts cibles. On parle alors de *descripteurs haut niveau* ou encore de *descripteurs sémantiques*.

Cette section est dédiée à l'étude de tels descripteurs. Dans un premier temps, nous présentons les enrichissements syntaxiques qui reposent sur les règles de syntaxe du langage, nous présentons ensuite les enrichissements sémantiques qui intègrent des connaissances supplémentaires sur le domaine d'étude.

1.1.2.1 Enrichissements syntaxiques

Une approche pour pallier le fossé sémantique consiste à prendre en compte les règles de syntaxe du langage. Cette nouvelle information permet entre autres de renseigner les différentes formes d'un mot, par exemple de dissocier les verbes, des noms, des adjectifs. Dans la phrase « *nous avions contrôlé les avions* » par exemple, la première occurrence du mot *avions* fait référence au verbe tandis que la seconde fait référence au nom.

L'étiquetage grammatical (*part of speech tagging*) consiste à associer à chacun des mots du corpus, son type grammatical. Un étiqueteur grammatical est par exemple un classifieur entraîné sur un corpus d'apprentissage pour lequel chacun des mots est préalablement associé à une étiquette grammaticale. Le système *TreeTagger* (Schmid, 1994) est un étiqueteur grammatical qui implémente un arbre de décision entraîné sur plusieurs langues dont le français et l'anglais. Plus récemment, de nouveaux étiqueteurs ont été proposés : parmi les méthodes d'apprentissage utilisées, les CRFs (*Conditional Random Fields*), qui tiennent naturellement compte de la nature séquentielle du texte, ainsi que les machines à vecteurs de support sont les plus étudiés.

Comme nous le verrons au chapitre suivant et dans la partie dédiée à l'analyse des émotions, le typage grammatical des descripteurs peut être exploité à différentes fins. Comme dans l'exemple précédent, il permet d'enrichir les entrées du dictionnaire, il vise également à filtrer les descripteurs bas niveau selon leur famille grammaticale (voir section 1.2.1.1, p. 20).

Il faut noter que les arbres de dépendance présentés en introduction de cette section constituent en eux-mêmes une forme d'enrichissement syntaxique : en effet, pour ces derniers, un document est décrit d'après les relations grammaticales qui existent entre les groupes syntaxiques qui le composent.

1.1.2.2 Enrichissements sémantiques

Une manière d'organiser un ensemble de connaissances sur le domaine d'étude est sous la forme de lexiques qui organisent un certain vocabulaire autour de concepts prédéfinis. Pour les taxinomies, ces concepts sont de plus organisés sous forme de hiérarchies, pour les ontologies des relations supplémentaires sont spécifiées entre les concepts. L'étude et la constitution de telles ressources est étudiée dans le cadre du traitement automatique des langues et constitue un domaine de recherche à part entière. Ici, nous présentons les différentes formes d'enrichissements utilisés de manière classique, à la section 1.2.2, p. 24 nous citons des méthodes de construction automatique pour induire de tels enrichissements directement du corpus d'étude.

Ressources sémantiques Nous proposons d'organiser les ressources sémantiques selon deux catégories : les ressources générales regroupent un vocabulaire qui se veut exhaustif et exploitent un espace des concepts générique ; les ressources contextuelles sont spécialisées pour un domaine d'application spécifique, cette spécialisation pouvant porter tant sur le vocabulaire considéré que sur les concepts définis. La base *wordnet* (Miller, 1995) est un exemple de ressource générale. Il s'agit d'une ontologie [2] qui spécifie entre autres des relations d'héritage et de synonymie entre les mots de la langue anglaise. La base *Bio-Lexicon* est un exemple de lexique spécifique au domaine de la biologie : il organise un ensemble de mots fréquemment utilisés dans le domaine de la biologie autour de concepts grammaticaux mais aussi autour de concepts spécifiques au domaine.

Tandis que les ressources générales définissent un vocabulaire générique, elles sont sujettes aux imprécisions du langage. Par exemple le mot *avocat* peut désigner aussi bien le fruit que l'homme de droit. Les ressources contextuelles ont notamment pour objectif de réaliser une bijection entre le vocabulaire exploité et les concepts définis. Malheureusement ce type de ressources est coûteux à produire et il n'en n'existe pas toujours de disponible pour le domaine étudié.

Par ailleurs, l'organisation de connaissances aussi bien générales que contextuelles reste soumise aux imprécisions du langage. Comme nous le verrons à la partie I, dédiée à l'analyse de concepts affectifs, bien qu'il soit possible de traiter ces ambiguïtés en fonction du contexte d'énonciation, il est également possible de quantifier le degré d'imprécision au niveau des ressources. La théorie des sous-ensembles flous introduite par Zadeh, propose d'exploiter des degrés d'appartenance à des classes d'objets afin de modéliser l'imprécision relative aux différents états d'un objet (Zadeh, 1965).

Descripteurs sémantiques Lorsque les entrées du dictionnaire reposent uniquement sur les enrichissements sémantiques, l'espace de description associé est un espace sémantique formé autour de concepts définis dans les ressources employées.

Plusieurs stratégies sont classiques pour la représentation d'un document dans un espace de description sémantique. La plus simple consiste à définir les entrées du dictionnaire comme l'ensemble du vocabulaire partagé entre les ressources et le corpus d'étude : il s'agit de la détection de mots-clefs (*keyword spotting*).

Pour faire face aux ambiguïtés inhérentes au langage, des stratégies plus avancées mettent en œuvre des grammaires d'extraction de descripteurs et sont étudiées dans le domaine du traitement automatique des langues. D'autres réalisent un apprentissage supervisé pour désambiguiser les différents sens d'un terme. Enfin, lorsque les ressources

2. Ici le terme ontologie est pris au sens large.

définissent un ensemble de concepts flous, les degrés d'appartenance associés au vocabulaire peuvent être exploités pour quantifier l'imprécision associée aux descripteurs correspondants.

1.1.3 Méthodes de comparaison pour les textes

Parallèlement au problème de la sélection de la représentation des données se pose celui de leur comparaison. Le choix d'une mesure de comparaison dépend à la fois de la nature de l'information traitée, de la structure de l'espace de description mais aussi de la sémantique portée par la mesure. A la section 1.1.3.1 nous présentons les mesures de distance utilisées de manière classique en apprentissage. Nous considérons ensuite, à la section 1.1.3.2, le cas des mesures de similarité et nous étudions en particulier leur compatibilité avec les représentations textuelles. Enfin, nous rappelons à la section 1.1.3.3 les fonctions noyaux qui étendent le champ d'étude des mesures de similarité basées sur le calcul d'un produit scalaire.

1.1.3.1 Mesures de distance

Dans un espace de description vectoriel $\mathcal{X}$ formé de m descripteurs, un document est représenté par un vecteur $\boldsymbol{x}$ de dimension m dont les coordonnées synthétisent l'information qu'il contient. Une approche naturelle pour comparer l'information associée à deux documents consiste ainsi à mesurer la distance entre leurs vecteurs de représentation respectifs. Nous présentons ici les mesures de distance utilisées de manière classique en apprentissage, leur emploi n'est pas nécessairement motivé pour des représentations textuelles.

Définition mathématique Une dissimilarité d est une fonction qui associe à deux vecteurs une mesure de la longueur les séparant dans l'espace de description :

$$\begin{aligned} d: \ & \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ & (\boldsymbol{x}, \boldsymbol{z}) \mapsto d(\boldsymbol{x}, \boldsymbol{z}) \end{aligned}$$

Lorsqu'elle possède de plus les propriétés suivantes, d est une *distance* :

identité :	$d(\boldsymbol{x}, \boldsymbol{z}) = 0 \Leftrightarrow \boldsymbol{z} = \boldsymbol{x}$
symétrie :	$d(\boldsymbol{x}, \boldsymbol{z}) = d(\boldsymbol{z}, \boldsymbol{x})$
inégalité triangulaire :	$d(\boldsymbol{x}, \boldsymbol{w}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{w})$

Distance de Minkowski Les distances de Minkowski constituent une famille paramétrée qui généralise le concept naturel de distance. A l'ordre p, l'expression d'une distance de Minkowski est la suivante :

$$d_p(\boldsymbol{x}, \boldsymbol{z}) := \|\boldsymbol{x} - \boldsymbol{z}\|_p = \left(\sum_{j=1}^{m} |x_j - z_j|^p \right)^{1/p}$$

où $\|.\|_p$ représente la p-norme de son vecteur argument. Une distance de Minkowski est de plus invariante aux translations :

$$d_p(\boldsymbol{x} + \boldsymbol{t}, \boldsymbol{z} + \boldsymbol{t}) = \|(\boldsymbol{x} + \boldsymbol{t}) - (\boldsymbol{z} + \boldsymbol{t})\|_p = d_p(\boldsymbol{x}, \boldsymbol{z})$$

Plusieurs éléments de cette famille sont des fonctions de distance bien connues : pour $p = 1$, d_1 est la distance de Manhattan qui mesure la somme des valeurs absolues des différences, sur chacune des dimensions prises indépendamment. En dimension $m = 2$, elle correspond par exemple au chemin parcouru entre deux points lorsque seuls les déplacements non diagonaux sont autorisés. d_2 est la distance euclidienne qui représente la distance « à vol d'oiseau » parcourue entre deux points de l'espace de représentation. A mesure que p croît, $d_p(\boldsymbol{x}, \boldsymbol{z})$ devient plus sensible aux grandes valeurs de $\boldsymbol{u} = |\boldsymbol{x} - \boldsymbol{z}|$. En particulier pour $p = \infty$, d_∞ est la distance de Tchebychev qui correspond à la composante maximum du vecteur $\boldsymbol{u}$.

Distance de Mahalanobis La distance de Mahanalobis est une extension de la distance euclidienne qui pondère la comparaison faite entre deux vecteurs sur chacune des composantes. Ce type de mesure s'avère nécessaire lorsque dans l'espace de représentation, il doit être tenu compte de l'importance relative de chacun des descripteurs, représentée par exemple par leurs effets conjoints. La distance de Mahanalobis, notée d_M, s'exprime de la manière suivante pour deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{z}$:

$$d_M(\boldsymbol{x}, \boldsymbol{z}) = \left[(\boldsymbol{x} - \boldsymbol{z})^\intercal S^{-1} (\boldsymbol{x} - \boldsymbol{z})\right]^{1/2}$$

où S est la matrice de variance-covariance obtenue sur le jeu de données étudié. La comparaison faite sur chacun des descripteurs est ainsi proportionnelle à la concentration des données sur chacun. Lorsque les descripteurs sont supposés indépendants, S est une matrice diagonale qui contient la variance associée à chacun des descripteurs et d_M constitue de plus une version pondérée de la distance euclidienne. Si cette variance est constante et vaut 1 par exemple, alors S est la matrice identité et $d_M = d_2$.

Enfin, tout comme la distance euclidienne, la distance de Mahanalobis est invariante aux translations. Elle est de plus invariante aux dilatations lorsque S est une matrice de variance-covariance.

Divergence de Kullback-Leibler Lorsque le vecteur $\boldsymbol{x} \in \mathcal{X}$ possède les propriétés additionnelles $x_j \geq 0$ pour tout j et $\|\boldsymbol{x}\|_1 = 1$, $\boldsymbol{x}$ décrit une variable aléatoire sur l'univers $\mathcal{V}$. Dans ce cadre, on peut utiliser une mesure de dissimilarité entre distributions de probabilité, $\mathrm{KL}(\boldsymbol{x}\|\boldsymbol{z})$ comme la dissimilarité de Kullback-Leibler également connue sous le nom d'entropie relative (de $\boldsymbol{x}$ par rapport à $\boldsymbol{z}$). Il s'agit d'une mesure asymétrique de la divergence entre deux distributions de probabilité : en général $\mathrm{KL}(\boldsymbol{x}\|\boldsymbol{z}) \neq \mathrm{KL}(\boldsymbol{z}\|\boldsymbol{x})$. Une version symétrisée consiste à calculer la divergence moyenne comme :

$$\begin{aligned}\mathrm{KL}_{\mathrm{avg}}(\boldsymbol{x}\|\boldsymbol{z}) = \frac{1}{2}\left(\mathrm{KL}(\boldsymbol{x}\|\boldsymbol{z}) + \mathrm{KL}(\boldsymbol{z}\|\boldsymbol{x})\right) &= \frac{1}{2}\sum_{j=1}^{m} x_j \log\frac{x_j}{z_j} + \sum_{j=1}^{m} z_j \log\frac{z_j}{x_j} \\ &= \frac{1}{2}\sum_{j=1}^{m}(x_j - z_j)\log\frac{x_j}{z_j}\end{aligned}$$

Il faut noter que $\mathrm{KL}_{\mathrm{avg}}(\boldsymbol{x}\|\boldsymbol{z})$ n'est pas définie quand un descripteur est nul pour la donnée $\boldsymbol{x}$ ou pour la donnée $\boldsymbol{z}$. Il est alors d'usage d'employer une petite quantité ϵ représentant l'absence d'un descripteur pour une donnée.

1.1.3.2 Mesures de similarité

Comme présenté à la section 1.1.1.1, une spécificité des représentations textuelles est que l'espace de description correspondant est presque vide : un document n'est souvent

formé que d'une faible partie du dictionnaire, les vecteurs de représentation correspondants présentent souvent de nombreuses composantes nulles. Ainsi, à moins que le vocabulaire utilisé dans le document $\boldsymbol{z}$ ne soit semblable à celui employé dans le document $\boldsymbol{x}$, l'amplitude des différences $\|\boldsymbol{x}-\boldsymbol{z}\|$ sur lesquelles reposent les fonctions de dissimilarité présentées précédemment n'ont que peu de chances d'approcher le vecteur $\mathbf{0}$. Or comme le témoigne l'insuccès des mesures de comparaisons basées sur une différence vectorielle (Strehl et al., 2000), il semblerait que pour les textes, l'amplitude du vocabulaire partagé soit une information bien plus pertinente que l'amplitude des différences observées.

Ces arguments motivent l'emploi d'autres types de mesures de comparaison pour les textes : dans un premier temps, nous discutons du passage d'une mesure de dissimilarité à une mesure de similarité ; dans un second temps nous considérons le produit scalaire et ses variantes normalisées, et nous étudions sa compatibilité avec les représentations textuelles.

Similarité induite par une dissimilarité Soit d une dissimilarité telle que présentée à la section précédente et $f : \mathbb{R} \to \mathbb{R}$ une fonction décroissante. Alors la fonction :

$$\begin{aligned} \mathrm{sim}_{f,d} : \ & \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ & (\boldsymbol{x}, \boldsymbol{z}) \mapsto f(d(\boldsymbol{x}, \boldsymbol{z})) \end{aligned}$$

est une fonction de similarité, f est appelée *fonction de transfert*. Lorsque d est normalisée dans l'intervalle unité, son complément à 1 réalise par exemple son transfert en une mesure de similarité. Lesot et al. (2009) étudient trois fonctions de transfert classiques en apprentissage :

Cauchy :	$f(d) = 1/(1 + (d/\theta)^\gamma)$
Gaussienne généralisée :	$f(d) = \exp(-(d/\theta)^\gamma)$
Sigmoïde :	$f(d) = 1/\exp(d - \theta/\gamma)$

où les paramètres θ et γ permettent de contrôler le comportement du transfert. Lesot et al. (2009) montrent que tandis que γ ajuste la sensibilité de $\mathrm{sim}_{f,d}$ aux petites valeurs, le paramètre θ permet de contrôler le seuil au dessus duquel les valeurs comparées ont une influence négligeable. Une similarité induite par une dissimilarité possède ainsi un comportement nouveau qui dépend de la fonction de transfert employée ainsi que de son paramétrage.

Produit scalaire Le produit scalaire entre les vecteurs $\boldsymbol{x}$ et $\boldsymbol{z}$ est noté :

$$\mathrm{sim}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^\intercal \boldsymbol{z}$$

Sur l'orthant positif $\mathbb{R}_+^m$ le produit scalaire n'est pas borné, sa valeur est proportionelle aux valeurs, sur chaque composante, de ses vecteurs arguments. Son minimum est nul, il est atteint en $\mathbf{0}$ pour l'un ou l'autre de ses arguments.

Le produit scalaire dispose de propriétés intéressantes pour la comparaison de représentations textuelles : pour chacun des descripteurs, 0 est un élément absorbant, de sorte que seul le vocabulaire partagé entre deux documents est pris en compte pour le calcul de leur similarité. De plus, sur l'intervalle unité, le produit par composantes réalisé par le produit scalaire exhibe des propriétés de renforcement négatif : la similarité entre deux vecteurs, sur chacune des composantes, est d'autant plus faible que les valeurs correspondantes sont faibles. Autrement dit, pour un corpus donné, l'ensemble du vocabulaire peu important pour l'ensemble des documents n'influence que peu les mesures de similarité effectuées entre documents.

Similarité angulaire La similarité angulaire entre deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{z}$ mesure le cosinus de l'angle α qu'ils forment, elle est définie comme :

$$\mathrm{sim}_\alpha(\boldsymbol{x}, \boldsymbol{z}) = \cos\alpha = \frac{\boldsymbol{x}^\intercal \boldsymbol{z}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{z}\|_2}$$

Sur l'orthant positif $\mathbb{R}_+^m$, la similarité angulaire est à valeur sur l'intervalle unité. Cette mesure est maximale et vaut 1 lorsque les vecteurs $\boldsymbol{x}$ et $\boldsymbol{z}$ sont colinéaires, elle est nulle quand ils sont orthogonaux.

La similarité angulaire est égale au produit scalaire si les vecteurs de représentation sont de norme unitaire. Elle mesure ainsi un produit scalaire invariant aux dilatations en normalisant ce dernier par la norme de chacun de ses arguments.

Indice de Jaccard L'indice de Jaccard binaire est une mesure de similarité entre deux ensembles $\mathcal{A}$ et $\mathcal{B}$. Elle consiste à calculer le rapport entre le nombre d'éléments communs aux deux ensembles $|\mathcal{A} \cap \mathcal{B}|$ sur le nombre total de possibilités $|\mathcal{A} \cup \mathcal{B}|$. Dans le cas continu l'extension suivante est généralement employée :

$$\mathrm{J}(\boldsymbol{x}, \boldsymbol{z}) = \frac{\boldsymbol{x}^\intercal \boldsymbol{z}}{\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{z}\|_2^2 - \boldsymbol{x}^\intercal \boldsymbol{z}} = \frac{\boldsymbol{x}^\intercal \boldsymbol{z}}{\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \boldsymbol{x}^\intercal \boldsymbol{z}}$$

Cette mesure est à valeurs dans l'intervalle unité. Elle est maximale lorsque $\boldsymbol{x} = \boldsymbol{z}$ et elle est minimale en $\mathbf{0}$ pour l'un ou l'autre de ses arguments. Entre les deux, elle réalise un compromis entre le produit scalaire et la distance euclidienne. Cette mesure exprime en effet un produit scalaire normalisé par la distance euclidienne de ses arguments augmentée de leur produit scalaire.

Sur l'orthant positif $\mathbb{R}_+^m$, la distance $1 - \mathrm{J}(\boldsymbol{x}, \boldsymbol{z})$ vérifie l'inégalité triangulaire, cette distance porte le nom de *distance de Tanimoto*.

Coefficient de corrélation de Pearson Le coefficient de corrélation de Pearson est une mesure classique de la dépendance linéaire entre deux distributions de probabilité. Comme la divergence de Kullback-Leibler, dans le cas où les documents $\boldsymbol{x}$ et $\boldsymbol{z}$ sont des variables aléatoires, le coefficient de corrélation de Pearson définit une mesure de similarité entre deux distributions de probabilité :

$$\rho(\boldsymbol{x}, \boldsymbol{z}) = \frac{\sigma_{\boldsymbol{x},\boldsymbol{z}}}{\sigma_{\boldsymbol{x}} \sigma_{\boldsymbol{z}}} = \frac{(\boldsymbol{x} - \bar{x})^\intercal (\boldsymbol{z} - \bar{z})}{\|\boldsymbol{x} - \bar{x}\|_2 \|\boldsymbol{z} - \bar{z}\|_2}$$

où $\bar{x} = \dfrac{1}{m} \displaystyle\sum_{j=1}^{m} (\boldsymbol{x})_j$ et $\bar{z} = \dfrac{1}{m} \displaystyle\sum_{j=1}^{m} (\boldsymbol{z})_j$ sont les moyennes respectives de chacun des arguments. Sur l'orthant positif $\mathbb{R}_+^m$, ρ est à valeurs dans l'intervalle unité. Il est maximum lorsque $\boldsymbol{x} = \boldsymbol{z}$ et minimum en $\mathbf{0}$ pour l'un ou l'autre de ses arguments. De plus, il faut noter que ρ est invariant aux translations ainsi qu'aux dilatations, il mesure en effet le cosinus de l'angle $\tilde{\alpha}$, formé entre les vecteurs centrés $\boldsymbol{x} - \bar{x}$ et $\boldsymbol{z} - \bar{z}$.

Tout comme la similarité angulaire, ce coefficient est un produit scalaire normalisé par la norme de ses vecteurs arguments, préalablement centrés.

Comparaison des mesures La figure 1.1, représente les lignes de niveaux pour quelques unes des fonctions de similarité présentées dans cette section. Les vecteurs $\boldsymbol{x} \in [0, 1]$ du carré unité ($m = 2$) sont comparés à un point de référence dont la position varie selon les lignes. Ce point a pour coordonnées $(1, 1)$ pour la première, il vaut $(0.1, 0.1)$ pour la

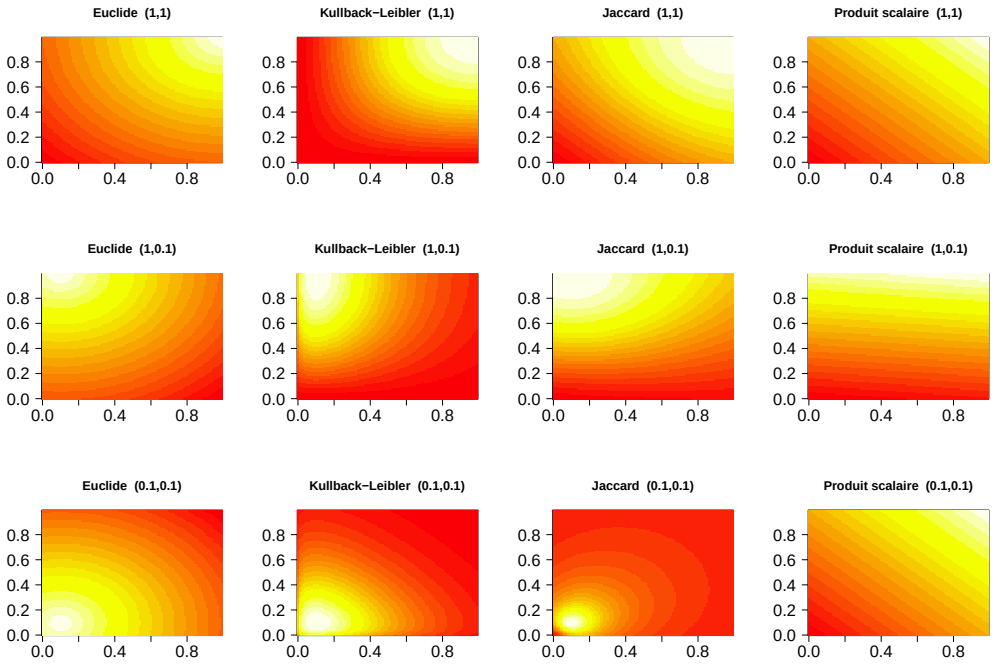


FIGURE 1.1 – Lignes de niveaux des mesures de similarité pour $\boldsymbol{x} \in [0,1]^2$ et un point de référence $\boldsymbol{z}$ placé, selon les lignes, en $(1,1)$, $(0.1,0.1)$, et $(1,0.1)$. Les colonnes décrivent les similarités respectives : $\exp[-d_2(\boldsymbol{x},\boldsymbol{z})]$, $\exp[-\mathrm{KL}_{\mathrm{avg}}(\boldsymbol{x},\boldsymbol{z})]$, $J(\boldsymbol{x},\boldsymbol{z})$, et $\boldsymbol{x}^\intercal\boldsymbol{z}$.

seconde, et $(1,0.1)$ pour la dernière. Les deux premières colonnes décrivent les mesures de similarité obtenues en appliquant, à la distance euclidienne ainsi qu'à la divergence de Kullback-Leibler, la fonction de transfert gaussienne de paramètres $\theta = 1$ et $\gamma = 1$. Les deux dernières colonnes correspondent respectivement à l'indice de Jaccard et au produit scalaire.

Quelle que soit la position du point de référence, quand $\boldsymbol{x}$ approche le vecteur $(1,1)$, c'est le produit scalaire qui produit la plus grande similarité. De plus sur la dernière ligne, pour laquelle la seconde composante du point de référence est presque nulle, le produit scalaire est la seule mesure qui associe une similarité tout aussi élevée à tout point situé en haut du carré. Comme remarqué en introduction de cette section, le produit scalaire est plus adapté aux spécificités des représentations textuelles : lors de leur comparaison, il tient spécifiquement compte du vocabulaire partagé entre deux documents, les descripteurs non pertinents jouant de plus un rôle mineur.

Les autres mesures constituent des variantes normalisées du produit scalaire et présentent toutes un comportement distinct. La similarité induite par la distance euclidienne atteint son maximum au point de référence ; ailleurs les lignes de niveau forment des lignes concentriques autour de ce point. Cette similarité autorise ainsi une prise de décision autour de valeurs centrées ce qui s'avère utile quand les descripteurs décrivent une grandeur quantitative comme une intensité par exemple. L'indice de Jaccard et la similarité induite par la divergence de Kullback-Leibler qui représentent, toutes deux, un compromis entre le produit scalaire et la distance euclidienne, exhibent un comportement très similaire. En effet, tandis que l'indice de Jaccard normalise le produit scalaire par une fonction proche de l'amplitude des différences, la divergence de Kullbak-Leibler réalise le produit scalaire

entre les vecteurs $\boldsymbol{u}$ et $\boldsymbol{u}'$, où le premier est le vecteur des différences entre $\boldsymbol{x}$ et $\boldsymbol{z}$, et le second est le vecteur des différences entre $\log \boldsymbol{x}$ et $\log \boldsymbol{z}$. Ces deux mesures sont d'autant plus élevées que la distance euclidienne entre $\boldsymbol{x}$ et $\boldsymbol{z}$ est élevée et que leur produit scalaire l'est aussi. Quand le point de référence est proche de $\mathbf{0}$, la seconde semble néanmoins plus sensible à la distance euclidienne que la première.

1.1.3.3 Fonctions noyaux

Les fonctions noyaux proposent un cadre d'étude étendu pour le calcul de la similarité par le produit scalaire. Leur emploi est de plus caractéristique du lien étroit qui existe entre espace de représentation et mesure de comparaison. Par exemple, dans le cas où les données sont représentées dans $\mathcal{X} = \mathbb{R}^2$, supposons qu'une mesure de similarité entre les vecteurs $\boldsymbol{x}$ et $\boldsymbol{z}$ soit donnée par $\kappa(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^\intercal \boldsymbol{z})^2$, alors on a :

$$\begin{aligned}\kappa(\boldsymbol{x}, \boldsymbol{z}) = (x_1 z_1 + x_2 z_2)^2 &= (x_1 z_1)^2 + (x_2 z_2)^2 + 2(x_1 x_2)(z_1 z_2) \\ &= (x_1^2, x_2^2, \sqrt{2} x_1 x_2)^\intercal (z_1^2, z_2^2, \sqrt{2} z_1 z_2)\end{aligned}$$

Ainsi $\kappa(\boldsymbol{x}, \boldsymbol{z})$ représente leur produit scalaire mesuré dans l'espace $\mathcal{F} = \mathbb{R}^3 : \kappa(\boldsymbol{x}, \boldsymbol{z})$ transforme implicitement les vecteurs de $\mathcal{X} = \mathbb{R}^2$ par la fonction $\Phi : \boldsymbol{x} \mapsto (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$.

Principe et motivations Un noyau $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ autorise donc le calcul de la similarité entre deux données comme leur produit scalaire dans un espace transformé $\mathcal{F}$, appelé espace des caractéristiques (*feature space*) :

$$\kappa : (\boldsymbol{x}, \boldsymbol{z}) \mapsto \Phi(\boldsymbol{x})^\intercal \Phi(\boldsymbol{z}) \in \mathbb{R}$$

où $\Phi : \mathcal{X} \to \mathcal{F}$ est une fonction appelée projecteur de caractéristiques (*feature map*).

L'intérêt est d'exploiter la géométrie de $\mathcal{F}$ sans avoir à calculer ni sa représentation explicite ni la définition exacte de Φ. Seules la définition de κ et les coordonnées des vecteurs dans l'espace originel doivent être connues : il s'agit de l'astuce du noyau (*kernel trick*). Par ailleurs, un autre attrait des fonctions noyaux est que peu de contraintes sont imposées sur l'espace d'entrée $\mathcal{X}$: il peut même s'agir d'un ensemble quelconque. Tant que l'espace des caractéristiques $\mathcal{F}$ est un espace pré-Hilbertien [3], κ est un noyau valide. Le noyau pour chaînes (*string kernel*) et le noyau pour arbres (*tree kernel*) qui sont notamment employés pour tenir compte de la structure du langage, sont par exemple définis pour des représentations non vectorielles.

Noyau linéaire et variantes normalisées Le produit scalaire classique est aussi connu sous le nom de *noyau linéaire :* pour celui-ci, l'espace des caractéristiques $\mathcal{F}$ est égal à l'espace d'entrée $\mathcal{X}$. Comme c'est le cas pour les représentations textuelles, lorsque l'espace d'entrée est en très grande dimension, le problème d'apprentissage correspondant est généralement linéairement séparable, ou du moins supposé l'être. Dans ce cas, l'intérêt du noyau linéaire est de pouvoir interpréter aisément les décisions prises par le système d'apprentissage : pour une tâche supervisée, ces décisions correspondent par exemple au vocabulaire discriminant pour les concepts cibles.

Par ailleurs comme nous l'avons vu à la section précédente, de nombreuses similarités mesurent un produit scalaire normalisé. A chacune est associé un espace des caractéristiques différent. La similarité angulaire présentée précédemment est un noyau linéaire pour lequel les vecteurs de l'espace d'entrée sont implicitement projetés sur une

3. un espace euclidien dont la dimension peut être infinie

Description	Forme	Matrice de Gram	Caractéristiques
addition de deux noyaux	$\kappa_1 + \kappa_2$	$K_1 + K_2$	$\mathcal{F}_1 \oplus \mathcal{F}_2$
produit par $\lambda \in \mathbb{R}_+$	$\lambda\kappa_1$	λK_1	$\mathcal{F}_1$ dilaté par λ
produit de deux noyaux	$\kappa_1 \times \kappa_2$	$K_1 \circ K_2$	$\mathcal{F}_1 \otimes \mathcal{F}_2$
exponentiation par $s \in \mathbb{R}$	κ_1^s	$K_1K_2(K_1K_2(\ldots))$	$\mathbb{R}[\mathcal{F}_1] \oplus \ldots \oplus \mathbb{R}[\mathcal{F}_1]^s$
exponentielle d'un noyau	$\exp(\kappa_1)$	idem	idem avec $s \to \infty$
noyau gaussien	$\exp\left(\frac{-\|\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$	idem	idem

TABLE 1.2 – Quelques opérations conservant la propriété de noyau : λ et s sont deux réels, κ_1 et κ_2 sont deux noyaux sur les espaces de caractéristiques $\mathcal{F}_1$ et $\mathcal{F}_2$.

hypersphère : seules les directions données par les vecteurs d'entrées sont conservées. De même, lorsque les vecteurs d'entrée sont de plus centrés, le noyau linéaire réalise une mesure du coefficient de corrélation de Pearson. D'autres espaces pourraient de même être mis en évidence pour les autres mesures de similarité : un certain nombre d'opérations sont conservatrices sur le corps des noyaux. Dans le tableau 1.2 nous présentons quelques unes de ces opérations, pour chacune, nous mettons en évidence la matrice de représentation (ou matrice de Gram) associée aux données ainsi que l'espace des caractéristiques implicitement défini. Il faut noter que pour le noyau gaussien présenté ci-dessous, l'espace des caractéristiques est potentiellement infini : il correspond à un espace formé de chacun des vecteurs d'entrée.

Noyau gaussien La fonction de similarité donnée dans l'exemple introductif correspond au *noyau polynomial* de degré 2 ; le *noyau gaussien* est un noyau polynomial dont le degré tend vers l'infini, ce qui s'exprime comme :

$$\kappa : (\boldsymbol{x}, \boldsymbol{z}) \mapsto \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|_2^2}{2\sigma^2}\right)$$

σ est l'écart-type associé à la fonction gaussienne ainsi définie. Ce paramètre ajuste un « rayon d'activation » minimal qui contrôle la sensibilité du noyau : pour de petites valeurs le noyau gaussien approche le noyau linéaire, et pour de très grandes valeurs, la similarité entre $\boldsymbol{x}$ et $\boldsymbol{z}$ est presque nulle pour tout $\boldsymbol{z} \neq \boldsymbol{x}$. Il faut par ailleurs noter que le noyau gaussien justifie l'emploi de la fonction gaussienne généralisée pour le transfert de la distance euclidienne en une mesure de similarité (Lesot et al., 2009).

1.2 Réduction de dimensions

Comme nous l'avons vu jusqu'à présent, dans le cas du texte l'espace de représentation $\mathcal{X}$ contient de nombreuses dimensions et est généralement vide : les documents sont représentés comme des points isolés dans un vaste espace vide. Les représentations textuelles font ainsi face au dilemme du fléau de la dimension (*curse of dimensionality*) : à mesure que le vocabulaire associé à un corpus s'élargit, le nombre de dimensions correspondant augmente de telle manière que la similarité entre toute paire de documents approche une constante. Dans cette section, nous présentons deux approches utilisées de manière classique pour pallier ce dilemme. La première effectue une *sélection des descripteurs* et consiste à éliminer de l'espace de représentation les dimensions non pertinentes au regard du problème considéré. La seconde effectue une *construction de descripteurs* : elle repose sur la constitution d'un nouvel espace dont les dimensions condensent l'information portée dans l'espace d'origine.

Comme décrit à la section 1.2.1, afin d'éliminer l'influence de certains descripteurs, des méthodes exploitent les spécificités liées aux données textuelles pour décider de la pertinence des descripteurs. Nous présentons également des méthodes numériques, utilisées de manière plus générale pour les jeux de données représentés en grande dimension.

Les méthodes qui construisent un nouvel espace de représentation sont présentées à la section 1.2.2. Dans un premier temps nous considérons le cas où des enrichissements sont exploités et les nouvelles dimensions correspondantes constituent des concepts. Dans un second temps, nous détaillons les méthodes qui condensent l'information portée dans l'espace d'entrée et qui forment ainsi un nouvel espace dont les dimensions sont identifiées comme des thèmes (*topics*).

1.2.1 Sélection de descripteurs

Les méthodes présentées dans cette section effectuent une sélection des descripteurs qui composent l'espace de description. Dans un premier temps nous considérons le cas spécifique du texte, dans un second temps nous décrivons des méthodes numériques, utilisées de manière plus générale pour des jeux de données décrits en grande dimension.

1.2.1.1 Méthodes de filtrage textuel

Comme présenté à la section 1.1.1.1, le dictionnaire construit d'après les mots composant un corpus contient généralement de nombreuses entrées non pertinentes. Nous décrivons ici des méthodes de filtrage pour ces entrées.

Filtrage orthographique Un procédé qui permet de réduire le nombre de descripteurs à très faible coût consiste par exemple à normaliser la casse (majuscule ou minuscule) des mots. L'intérêt est ici double puisque les mots dont la casse différait au préalable sont désormais représentés par un descripteur unique, augmentant ainsi la similarité des documents correspondants.

Les jeux de données réelles contiennent par ailleurs de nombreuses fautes de frappe ou d'orthographe qui introduisent de nouvelles entrées aux dictionnaires. Ces derniers étant évidemment non discriminants [4], un procédé répandu est l'emploi de correcteurs automatiques d'orthographe. Le programme libre *GNU Aspell* supporte par exemple de nombreuses langues dont le français et l'anglais.

Filtrage syntaxique Les mots composants le langage naturel possèdent de nombreuses formes comme par exemple les verbes, les adjectifs ou les adverbes. Ces formes présentent elles aussi de nombreuses flexions, pour lesquelles on peut notamment citer les formes conjuguées, les formes infinitives ou encore les pluriels.

Une première méthode repose sur l'observation que pour le problème considéré, certaines catégories grammaticales sont naturellement peu discriminantes. A partir d'un étiquetage grammatical calculé sur le corpus d'étude (voir section 1.1.2.1), les entrées du dictionnaire peuvent alors être filtrées selon la pertinence de leur catégorie correspondante. Il est notamment courant d'éliminer des dictionnaires les caractères de ponctuation, les noms propres ainsi que les chiffres par exemple.

L'étiquetage syntaxique des documents peut également être exploité afin de réduire les mots à leur lemme : pour un verbe il s'agit de sa forme infinitive, pour les autres mots,

4. Ce n'est pas le cas pour les travaux portant sur l'étude de style, par exemple pour la reconnaissance automatique de l'auteur d'un document.

il s'agit de leur forme masculine au singulier. Par exemple, les formes *mangeait*, *manger*, *mangent* ont pour forme canonique le lemme *manger*. De manière similaire, la racinisation (*stemming*) consiste à réduire les mots à leur racine appelée stemme. Par exemple, les formes *mangeait*, *manger*, *mangent* ont pour racine commune *mang*. Ce processus qui est implémenté sous la forme de règles expertes, spécifiques à une langue, est simple et rapide à mettre en œuvre, il de plus efficace d'où son fort succès. La lemmatisation constitue un filtrage plus précis que la racinisation, puisqu'elle différencie généralement les lemmes selon leur catégorie grammaticale : il est par exemple possible de distinguer pour le lemme *dîner*, le verbe du nom commun. Ainsi la lemmatisation est souvent préférée à la racinisation excepté lorsque les mots de $\mathcal{D}$ contiennent de nombreuses fautes, dans quel cas les règles de racinisation donnent de meilleurs résultats. Il faut noter qu'il peut s'agir là d'une alternative à la correction automatique d'orthographe.

Filtrage sémantique Comme présenté à la section 1.1.2.2, p. 12, les enrichissements sémantiques peuvent être exploités afin de constituer un espace de représentation sémantique. Dans ce cas, les entrées des dictionnaires sont réduites aux concepts définis dans les ressources utilisées et l'espace de représentation correspondant n'a plus les caractéristiques d'un espace de représentation textuel.

Filtrage fréquentiel La loi de Zipf énonce que dans un corpus de documents $\mathcal{D}$, la fréquence d'un mot est liée à son rang (dans l'ordre décroissant de fréquence) par une loi exponentielle. Ainsi, si l'on classe les mots selon leur fréquence d'apparition dans $\mathcal{D}$, on observe que le premier mot est bien plus fréquent que le second, qui est lui-même bien plus fréquent que le troisième et ainsi de suite. Il est alors courant d'identifier les mots de plus haut rang (les plus fréquents) ainsi que ceux de plus faible rang (les plus rares) comme un vocabulaire non discriminant pour le problème considéré : en effet les premiers apparaissent dans la majorité des documents de $\mathcal{D}$, les seconds n'apparaissant que de manière isolée, presque accidentellement pour quelques documents. Ce procédé revient donc à lisser la distribution des mots de manière à l'éloigner de la distribution de Zipf, ou de manière à l'approcher d'une distribution uniforme. Les seuils considérés pour le filtrage des mots dépendent entièrement du problème étudié, il est néanmoins d'usage d'éliminer 20% du vocabulaire le plus fréquent ainsi que 20% du vocabulaire le plus rare.

Il faut noter que parmi les mots les plus fréquents, beaucoup sont les mots les plus communs d'une langue comme par exemple *le*, *un*, *elle*, on parle alors de mots vides (*stop words*). Pour des unigrammes il est d'usage d'exploiter une liste de mots vides (*stop words list*) pour filtrer ceux de rang moyen, que l'on ne pourrait identifier au travers de leur fréquence.

Enfin, lorsque la méthode de pondération utilisée est le schéma tf/idf (voir section 1.1.1.1, p. 8), ce type de filtrage est implicitement exploité puisque le poids associé aux mots est alors inversement proportionnel à leur fréquence dans le corpus.

1.2.1.2 Méthodes numériques

De nombreuses méthodes interprètent l'information décrite dans $\mathcal{X}$ comme un signal bruité et adoptent une approche numérique au problème du filtrage des descripteurs. Ces méthodes sont utilisées dans de nombreux domaines, y compris celui du texte : l'apprentissage à partir d'images ou de vidéos (*computer vision*) ou l'apprentissage à partir de puces à ADN rencontrées en bio-informatique (*micro array data*) en sont des exemples d'application.

Lorsque les jeux de données sont représentés en grande dimension, ces méthodes visent à pallier le dilemme du fléau de la dimension présenté ci-avant, mais aussi à situer le pouvoir de description des descripteurs employés. Enfin, le rasoir d'Occam stipule que parmi un ensemble de solutions, il est préférable de choisir celles qui font le moins d'hypothèses. Ici, les hypothèses sont vues comme les descripteurs et le rasoir d'Occam est entendu comme un principe de parcimonie.

Principe général En supposant que soit donné un critère de qualité J, concave, une formulation pour le problème de la sélection de dimensions est la suivante :

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & J(X\boldsymbol{w}, \Theta) \\ \text{s.t} \quad & \boldsymbol{w} \in \{0,1\}^m \\ & \text{card}(\boldsymbol{w}) \leq \delta \end{aligned} \tag{1.1}$$

Le vecteur candidat $\boldsymbol{w}$ est évalué par J sur le jeu de données X étant donné un ensemble de paramètres supplémentaires Θ. Le paramètre δ est donné par l'utilisateur, il ajuste le nombre de dimensions souhaitées : à toute solution $\boldsymbol{w}^*$ correspond un nouvel espace $\mathcal{X}'$ de dimension $\delta < m$ qui contient la nouvelle matrice de représentation $X' = XW$ avec $W = \text{diag}(\boldsymbol{w})$.

Indépendamment de la nature de J, ce problème comporte des contraintes binaires (les vecteurs de poids) et constitue un problème NP-difficile. Les méthodes *enveloppantes* en approchent une solution en organisant l'espace de recherche comme un treillis (Blum & Langley, 1997) : le critère de qualité guide alors l'exploration des 2^m états représentant chacun des vecteurs candidats. Dans la suite nous présentons deux famille d'approches différentes, qui consistent toutes deux, à transformer le problème (1.1) en un problème similaire mais plus facile : pour la première c'est le critère de qualité qui est simplifié, pour la seconde ce sont les contraintes qui le sont. Pour ces deux approches, le vecteur de pondération recherché est à valeurs continues et non binaires. Dans la littérature il est commun de se référer aux problèmes correspondants comme une pondération de descripteurs (*feature weighting*) et non une sélection de descripteurs (*feature selection*). Dans ce document, nous ne marquons pas de différence explicite entre les deux, dans le cas d'une pondération, l'importance des descripteurs non pertinents est en effet souvent hautement négligeable.

Méthodes filtrantes Ces méthodes envisagent une simplification du critère de qualité et l'expriment de manière indépendante pour chacun des descripteurs (Delavallade, 2007) :

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \sum_{j=1}^{m} [w_j J(X_j, \Theta)] \\ \text{s.t} \quad & \boldsymbol{w} \in \{0,1\}^m \\ & \text{card}(\boldsymbol{w}) \leq \delta \end{aligned}$$

où, comme le note Delavallade (2007), la somme peut être remplacée par tout opérateur d'agrégation $\text{Agg} : [0,1]^m \mapsto \mathbb{R}$ qui mesure la qualité globale de $\boldsymbol{w}$ en agrégeant m mesures de qualité données individuellement par J sur chacun des descripteurs X_j.

Une solution pour ce problème consiste alors à produire un vecteur $\boldsymbol{w}$ qui sélectionne les composantes associées aux δ plus grandes mesures de qualité $J(X_j, \Theta)$. Une approche différente, pour laquelle les composantes du vecteur $\boldsymbol{w}$ ne sont plus nécessairement binaires,

consiste à effectuer une pondération des descripteurs en remplaçant les contraintes du problème par :

$$\boldsymbol{w} \in [0,1]^m \text{ et } R(w) \leq s$$

où $R : [0,1]^m \to \mathbb{R}$ est une fonction qui contraint l'espace de recherche et le paramètre s joue un rôle similaire à celui de δ. Par exemple, lorsque $R(w) = \|\boldsymbol{w}\|_1$, les poids associés à chacun des descripteurs sont exprimés de manière relative dans $\boldsymbol{w}$ et la masse totale représentée par s est alors disputée par chacun. Ici le vecteur solution $\boldsymbol{w}^*$ serait donné par :

$$w_j = \frac{s \times J(X_j, \Theta)}{\sum_{l=1}^{m} J(X_l, \Theta)}$$

Ainsi, à mesure que le paramètre s tend vers zéro, de plus en plus de composantes de $\boldsymbol{w}$ approchent la valeur nulle, et les descripteurs de $\mathcal{X}$ correspondants perdent de leur influence. Les différentes méthodes proposées s'organisent alors selon les divers critères de qualité employés.

Quand le problème est supervisé, le vecteur $\Theta = \boldsymbol{y}$ de n composantes décrit les étiquettes de classes associées aux données : pour chacun des descripteurs X_j, un indice de qualité peut être mesuré par alignement avec l'ensemble des étiquettes. A cet effet, certains auteurs proposent d'exploiter une mesure de corrélation (Guyon & Elisseeff, 2003), d'autres se placent dans le cadre de la théorie de l'information et proposent par exemple de mesurer un gain d'information. Au chapitre 3, nous exploitons une méthode similaire qui repose sur l'entropie de Shannon.

En absence de supervision, une approche consiste à induire automatiquement les étiquettes de classes à partir de la densité des données dans l'espace de représentation (Wilbur & Sirotkin, 1992; Liu et al., 2003). Aggarwal et Zhai (2012) proposent une étude détaillée de ces méthodes dans le cas particulier du texte.

Apprentissage parcimonieux Les méthodes qui effectuent un apprentissage parcimonieux reposent sur une simplification des contraintes identique à celle présentée précédemment, le critère de qualité est lui conservé :

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & J(X\boldsymbol{w}, \Theta) \\ \text{s.t} \quad & \boldsymbol{w} \in [0,1]^m \\ & R(\boldsymbol{w}) \leq s \end{aligned}$$

Pour ce problème, J évalue la qualité d'un système d'équations linéaires pour lequel chacun des vecteurs de représentation est comparé au vecteur de pondération $\boldsymbol{w}$, désormais à valeurs continues. Dans un cadre supervisé $\Theta = Y$ et le critère de qualité peut être défini par $-L(Xw, Y)$, avec L une fonction de coût, convexe, qui mesure les erreurs commises par le modèle $\boldsymbol{w}$ sur les données X étant donné leurs étiquettes de classes $\boldsymbol{y}$. Ainsi le problème (1.1) se rapporte à un problème d'apprentissage linéaire. Contrairement aux méthodes filtrantes les méthodes que nous citons ici sont dites directes, elles effectuent une sélection des descripteurs en imposant des contraintes de parcimonie au modèle $\boldsymbol{w}$ ajusté en phase d'apprentissage. De manière équivalente, ce problème peut être formulé comme :

$$\min_{\boldsymbol{w} \in [0,1]^m} \quad L(X\boldsymbol{w}, \Theta) + sR(\boldsymbol{w})$$

cette forme est caractéristique de l'apprentissage régularisé qui fait référence au compromis biais/variance en apprentissage (Hastie et al., 2001). Ici, $R : \mathcal{X} \mapsto \mathbb{R}$ est une fonction de régularisation convexe et s est un paramètre de régularisation qui permet de contrôler le compromis entre les erreurs commises en phase d'apprentissage et la simplicité du modèle $\boldsymbol{w}$.

La famille des normes l_p est traditionnellement utilisée pour régulariser le modèle : à l'ordre $p = 0$, la norme l_0 est définie comme $\|\boldsymbol{w}\|_0 = \text{card}(\boldsymbol{w})$ et le problème précédent est effectivement identique au problème (1.1) (Weston et al., 2003). Pour des ordres $0 \leq p < 1$, le problème précédent n'est plus un problème convexe et comme pour le problème (1.1) la recherche d'un vecteur solution est difficile. Les méthodes considèrent ainsi des ordres plus élevés : tandis que la norme euclidienne $l2$ confère d'intéressantes garanties de généralisation, la norme $l1$ est la plus petite norme convexe qui exhibe des propriétés de parcimonie (Tibshirani, 1996; Wang & Shen, 2009). Néanmoins, cette dernière exhibe une singularité en zéro et la résolution du problème correspondant pose un certain défi (Schmidt et al., 2009; Wang & Shen, 2009). Plusieurs méthodes ont ainsi été étudiées, certaines proposent une solution pour un cadre d'apprentissage particulier (Tibshirani, 1996; Efron et al., 2004; Schmidt et al., 2009), d'autres étudient une solution générale (Park & Hastie, 2007).

Dans un cadre non supervisé, l'apprentissage parcimonieux conduit généralement à une solution triviale pour laquelle un seul descripteur est sélectionné (Witten & Tibshirani, 2010). L'absence de supervision est alors compensé par l'emploi d'une fonction que nous qualifions de *fonction barrière* et dont le rôle est de forcer un partitionnement des données sur un ensemble non trivial de descripteurs. En particulier, une version parcimonieuse de l'algorithme des K-moyennes qui repose sur la norme $l1$, utilise la fonction d'entropie de Shannon comme fonction barrière (Friedman & Meulman, 2004; Jing et al., 2007). Une autre extension propose d'employer pour fonction barrière, la norme $l2$ (Witten & Tibshirani, 2010). Au chapitre 8 nous étudions une approche similaire pour le partitionnement de documents dans un espace en très grande dimensions.

1.2.2 Construction de descripteurs

Une autre stratégie pour réduire le nombre de dimensions de $\mathcal{X}$ consiste à construire de nouveaux descripteurs sur lesquels l'information est naturellement condensée. Tout comme pour la sélection de dimensions, nous notons $\delta < m$ le nombre de dimensions du nouvel espace de description $\mathcal{X}'$. Deux stratégies permettent de le construire : une première approche consiste à exploiter une base de connaissances organisée autour de δ concepts, $\mathcal{X}'$ est alors un espace de concepts tel que présenté à la section 1.1.2, p. 11. Une seconde approche repose sur l'identification de δ thèmes (*topics*) dans l'espace de représentation originel, $\mathcal{X}'$ est alors un espace de thèmes.

1.2.2.1 Espace de concepts

Comme rappelé à la section 1.1.2, les enrichissements sémantiques visent à affiner l'information extraite d'un corpus de données. De nombreuses méthodes ont été proposées pour extraire δ concepts à partir d'un jeu de données $\mathcal{D}$, les plus simples consiste à calculer l'intersection entre le vocabulaire utilisé dans ce dernier et celui défini dans les ressources utilisées. Néanmoins, face aux nombreuses ambiguïtés du langage ainsi qu'à sa grande complexité, ces méthodes apparaissent comme rudimentaires et ne donnent généralement pas de résultats satisfaisants.

Beaucoup de travaux dans le domaine du traitement automatique des langues portent sur l'extraction des concepts dans les textes. Parmi les approches les plus classiques, les grammaires d'extraction proposent de tenir compte des nombreuses subtilités de la langue. Ces dernières sont souvent semi-automatiques et le processus d'extraction tient compte du contexte d'énonciation. Par ailleurs, en tenant compte de la nature séquentielle du langage, certaines méthodes reposent essentiellement sur un apprentissage pour l'extraction des concepts. C'est notamment le cas des outils d'analyse syntaxique présentés à la section 1.1.2. Ces méthodes sortant du cadre de nos travaux, nous ne les approfondissons pas davantage.

1.2.2.2 Espace de thèmes

Lorsque seul est donné le corpus d'étude, une autre approche consiste à construire l'espace $\mathcal{X}'$ à partir de $\mathcal{X}$ au travers de la seule connaissance de la matrice de représentation X. Conceptuellement, ces méthodes consistent toutes en l'apprentissage d'une partition sur $X^\intercal$ composée de δ clusters. En effet, la transposée de la matrice d'entrée décrit les mots du corpus par rapport aux documents du corpus, les clusters produits sont alors vus comme des regroupements de mots expliqués par leurs co-occurrences dans le corpus. Le nouvel espace de description $\mathcal{X}'$ est alors composé de δ dimensions, qui constituent des thèmes décrits par les représentants, aussi appelés *centroïdes*, de chacun des clusters dans l'espace d'origine.

Deux axes de recherche organisent alors les travaux : une première approche trouve racine en algèbre linéaire et plus particulièrement dans le domaine de la factorisation de matrices. L'objectif est alors de trouver une approximation de dimension $n \times \delta$ de la matrice X de dimension $n \times m$. Une seconde approche, probabiliste, consiste à définir et à ajuster un modèle de langage dans lequel il est supposé que les documents sont générés au travers de δ thématiques, elles-mêmes distribuées sur l'espace d'entrée. La méthode d'analyse de sémantique latente (*lsa*) consiste par exemple à calculer la décomposition en valeurs singulières de rang δ de la matrice d'entrée : $X' = U_\delta \Sigma_\delta V_\delta^\intercal$. Cette approximation est alors optimale au sens où elle minimise l'erreur $\|X - X'\|_F$ (Landauer et al., 1998), où $\|.\|_F$ est la norme de Frobenius qui représente le pendant matriciel de la norme euclidienne.

Les nouvelles dimensions de l'espace ainsi formé demeurent cependant difficiles à interpréter, Hofmann (2001) propose une extension probabiliste de l'analyse de sémantique latente (*plsa*) : un modèle de langage est alors obtenu en estimant la distribution de d thématiques sur le vocabulaire d'un corpus, puis celle des documents sur ces thématiques.

Plus récemment, d'autres méthodes ont été étudiées : la factorisation de matrices non négatives (*nmf*) consiste à identifer une approximation plus interprétable que celle fournie par la décomposition en valeurs singulières (Aggarwal & Zhai, 2012). Dans un cadre probabiliste, l'allocation latente de Dirichlet (*dla*) propose à la fois de modéliser la génération des documents et de pallier les problèmes de sur-apprentissage observés sur l'analyse de sémantique latente probabiliste (Blei et al., 2003).

1.3 Représentation multiple : fusion

Comme nous l'avons vu jusqu'à présent, pour le problème considéré, de nombreux choix déterminent la qualité de l'information extraite des données étudiées. Bien que ces choix soient motivés par un savoir expert sur le domaine d'étude, il peut s'avérer qu'aucune forme de représentation individuelle ne soit suffisante pour modéliser à bien les concepts étudiés. Dans ce cadre, exploiter plusieurs espaces de représentations peut améliorer la

description qui est faite des documents.

1.3.1 Motivation et principe

Supposons que, pour une tâche d'apprentissage donnée, il soit fourni deux espaces de description $\mathcal{X}_1$ et $\mathcal{X}_2$: le $i^{\text{ème}}$ document du jeu de données $\mathcal{D}$ possède deux représentations, $\boldsymbol{x}_i^1 \in \mathcal{X}_1$ et $\boldsymbol{x}_i^2 \in \mathcal{X}_2$. Il est possible que $\mathcal{X}_1$ soit plus adapté à décrire les concepts étudiés que $\mathcal{X}_2$ ou inversement, il s'agit alors de choisir le meilleur espace de description pour mener à bien l'apprentissage des concepts cibles. Dans d'autres cas en revanche, il peut être souhaitable d'exploiter ces deux espaces de représentations simultanément, soit parce qu'aucune représentation individuelle ne donne de résultats satisfaisants, soit parce qu'une combinaison permet d'obtenir de meilleurs résultats. Dans le cas général $L > 1$ espaces de représentation décrivent $\mathcal{D} : \mathcal{X}_1, \ldots, \mathcal{X}_L$, où $\boldsymbol{x}_i^l \in \mathcal{X}_l$ est l'un des L vecteurs de représentation du $i^{\text{ème}}$ document de $\mathcal{D}$. Un exemple de représentation multiple consiste alors à construire les $\mathcal{X}_l$ comme des l-grammes d'ordres différents (voir section 1.1.1.1, p. 8), exprimant des contextes de mots de tailles différentes. Un autre exemple est l'exploitation combinée de descripteurs bas niveau ainsi que d'enrichissements structurés (voir section 1.1.2, p. 11). Les espaces de description peuvent aussi être formés des descripteurs multimodaux, ainsi pour un document web, $\mathcal{X}_1$ peut être une représentation textuelle, $\mathcal{X}_2$ une représentation visuelle (e.g. les images figurant sur une page web) et $\mathcal{X}_3$ les hyperliens pointant vers le document considéré.

Dans le cadre classique, chacun des $\mathcal{X}_l$ est soumis à un processus d'évaluation individuel et l'espace de représentation le plus pertinent est ainsi retenu. Selon le contexte de supervision, la notion de pertinence peut être vue comme une mesure des erreurs d'étiquetage commises par un classifieur ou comme une mesure de l'homogénéité associée à un partitionnement des données. Dans la suite, nous désignons par f aussi bien un classifieur qu'une partition ; lorsqu'une distinction doit être faite, nous précisons explicitement la nature de f. Ainsi dans le cadre classique, aucune interaction n'est considérée entre les L espaces de description disponibles.

Au contraire, la fusion d'informations consiste à exploiter une combinaison de ces espaces, l'hypothèse sous-jacente étant qu'il existe une combinaison des $\mathcal{X}_l$ qui facilite l'apprentissage des concepts étudiés. Il faut noter que la sélection d'un espace de description parmi L espaces est également une telle combinaison.

Dans la suite, φ est une fonction d'agrégation sur L espaces de description, comme détaillé dans les sections suivantes, sa définition dépend du type de fusion considéré mais aussi des hypothèses faites pour le problème considéré. Il existe trois grandes approches pour effectuer de la fusion d'information, chacune intervient à un niveau différent de la chaîne d'apprentissage :

1. la fusion anticipée repose sur la construction d'un nouvel espace de description formé par la combinaison de chacun des espaces originaux,
2. la fusion tardive consiste à agréger les réponses de classifieurs entraînés individuellement sur chacune des représentations. Cette approche n'intervient plus réellement au niveau de la représentation des données.
3. Enfin, lorsque le classifieur exploite une mesure de similarité entre documents, la fusion intermédiaire réside en la construction d'une fonction de similarité qui agrège L mesures de similarité, chacune spécifique à l'un des espaces d'origine.

Ces trois approches sont résumées dans le tableau 1.3 et détaillées dans les sections suivantes.

méthode	**niveau de fusion**	**expression**	
pas de fusion	-	$f(\boldsymbol{x}) =$	$f(\boldsymbol{x}^1)$ ou ... ou $f(\boldsymbol{x}^L)$
fusion anticipée	espace de description $\mathcal{X}$	$f(\boldsymbol{x}) =$	$f(\varphi(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^L))$
fusion tardive	classifieurs/partitions f	$f(\boldsymbol{x}) =$	$\varphi(f_1(\boldsymbol{x}^1), \ldots, f_L(\boldsymbol{x}^L))$
fusion intermédiaire	fonction de similarité κ	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) =$	$\varphi(\kappa_1(\boldsymbol{x}_i^1, \boldsymbol{x}_j^1), \ldots, \kappa_L(\boldsymbol{x}_i^L, \boldsymbol{x}_j^L))$

TABLE 1.3 – Méthodes pour l'apprentissage en présence de représentations multiples.

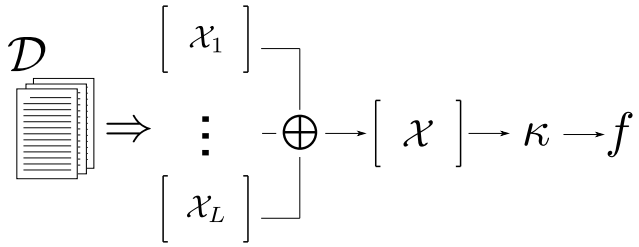


FIGURE 1.2 – Chaîne d'apprentissage pour la fusion anticipée.

1.3.2 Fusion anticipée : concaténation de descripteurs

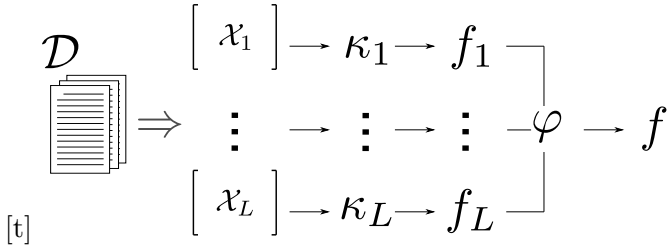
La fusion anticipée consiste à construire l'espace de représentation $\mathcal{X}$ comme une combinaison des L espaces de représentation fournis. Une approche naturelle consiste à définir $\mathcal{X}$ comme la concaténation des espaces d'origine. Aussi pour $L = 2$, φ est l'opération qui consiste à mettre bout à bout les descripteurs de $\mathcal{X}_1$ et ceux de $\mathcal{X}_2$. Dans le cas général, pour $L > 1$ espaces de représentation, la fusion anticipée consiste à construire $\mathcal{X}$ comme :

$$\mathcal{X} = \bigoplus_{i=1}^{l} \mathcal{X}_i$$

Un élément $\boldsymbol{x}$ de $\mathcal{X}$ s'exprime alors comme la concaténation de chacun des L vecteurs de représentation : $\boldsymbol{x} := \boldsymbol{x}^1 \oplus \ldots \oplus \boldsymbol{x}^L$. La chaîne d'apprentissage prend alors la forme représentée sur la figure 1.2.

Cette méthode présente un certain nombre d'avantages : d'abord le calcul de $\mathcal{X}$ est simple à implémenter et rapide à mettre en œuvre, la fusion anticipée permet donc de combiner des informations multiples à moindre coût. De plus, la fusion est implémentée au niveau des espaces de description : dans la chaîne d'apprentissage une seule représentation est exploitée et aucune modification n'est donc apportée aux algorithmes d'apprentissage mis en œuvre. Enfin, la fusion étant implémentée au plus bas niveau de la chaîne d'apprentissage, l'algorithme d'apprentissage peut tenir compte des interactions entre les espaces de représentation d'origine.

En dépit de sa simplicité cette méthode présente des limites, lorsque les $\mathcal{X}_l$ sont de nature très hétérogène, une fois $\mathcal{X}$ construit, la nature originelle des descripteurs est perdue. En particulier, si l'un des espaces de représentation d'origine contient un nombre de dimensions bien supérieure, alors $\mathcal{X}$ se trouve dominé par ce dernier de sorte que les autres espaces d'origine se retrouvent naturellement submergés. C'est par exemple le cas lorsque les $\mathcal{X}_l$ sont des descripteurs bas niveau de contexte de tailles différentes. En effet, comme nous l'avons vu à la section 1.1.1.1, p. 8 les dictionnaires calculés sur les unigrammes, les bigrammes et les trigrammes sont en général de tailles très différentes.



[t]

FIGURE 1.3 – Chaîne d'apprentissage pour la fusion tardive.

Il convient alors d'homogénéiser les espaces d'origine. Une méthode simple consiste alors à définir un nombre de dimensions B identique pour chacun des $\mathcal{X}_l$. En appliquant les méthodes présentées dans la section 1.2, une forme d'homogénéisation consiste par exemple à réduire chacun des espaces de description aux B descripteurs les plus pertinents. En observant par ailleurs que ce problème de déséquilibre se rapporte à un problème de pondération des espaces d'origine, une autre approche consiste à normaliser les vecteurs de représentation dans chacun des espaces originels. Tel que détaillé à la section 1.3.4, p. 31, ce principe motive notamment les méthodes qui réalisent une fusion intermédiaire.

De manière plus générale, les espaces d'origine peuvent contenir des descripteurs de nature intrinsèquement différente. Pour les données textuelles c'est particulièrement le cas lorsque les espaces de concept présentés à la section 1.2.2 sont combinés aux descripteurs bas niveau. En observant que différents types de descripteurs nécessitent différentes notions de similarité, d'autres types de fusions sont préférées à la fusion anticipée dans le cas où les $\mathcal{X}_l$ sont de nature intrinsèquement hétérogène.

1.3.3 Fusion tardive : agrégation de décisions

Un autre type de fusion, dite tardive, consiste à construire la décision $f(\boldsymbol{x})$ prise sur le document $\boldsymbol{x}$ comme une agrégation des décisions $f_l(\boldsymbol{x}^l)$ prises indépendamment sur chacune des représentations $\boldsymbol{x}^l$:

$$f(\boldsymbol{x}) = \varphi(f_1(\boldsymbol{x}^1), \ldots, f_L(\boldsymbol{x}^L))$$

où, f_l est soit un classifieur soit une partition, entraîné ou construite de manière indépendante sur $\mathcal{X}_l$, et φ est une fonction qui agrège les décisions individuelles prises sur chacune des L représentations. Il est possible de faire l'analogie entre la fusion tardive et les méthodes d'ensemble. Contrairement à ces dernières, les f_l sont ici supposés robustes, on suppose de plus que chacun des f_l est obtenu par apprentissage sur l'ensemble des documents disponibles.

1.3.3.1 Motivations

La fusion tardive est particulièrement utilisée lorsque les espaces de description sont de nature hétérogène. En effet, contrairement à la fusion anticipée, il est ici possible de définir des hypothèses propres à chacun des espaces d'origine. Il est par exemple courant d'utiliser la fusion tardive dans le cadre de l'apprentissage multimodal, lorsque chacun des $\mathcal{X}_l$ représente un mode de l'information contenu dans le corpus étudié. Dans ce cas, la représentation faite des données, la mesure de comparaison employée, et le choix ainsi que le paramétrage des systèmes d'apprentissage sont réalisés de manière individuelle.

1.3.3.2 Règles d'agrégation

Contrairement à la fusion anticipée la fusion tardive autorise la prise en compte de règles d'agrégation complexes, nous présentons dans un premier temps le principe de leur emploi pour une tâche de fusion tardive, nous détaillons ensuite quelques opérateurs classiques d'agrégation, pour chacun nous décrivons l'utilisation qui en est faite.

Principe et formalisation Nous supposons que les f_l produisent chacun un ordonnancement des K concepts cibles en leur associant un degré de confiance $\mu_k : f_l : \boldsymbol{x}^l \mapsto \{(k, \mu_k)\}_{k=1}^K$. Ici, φ est une fonction qui agrège les L degrés de confiance $f_l(\boldsymbol{x}^l)$ associés à la décision k pour le document $\boldsymbol{x}$. Selon le problème considéré, cet étiquetage multiple est parfois rapporté à un étiquetage simple : la décision de confiance maximale est alors retenue. Dans ce qui suit, les degrés de confiance sont supposés à valeurs réelles, de plus nous considérons qu'ils sont homogènes entre les différents f_l, autrement dit qu'ils expriment les mêmes quantités. Ainsi, pour chaque f_l, les degrés de confiance sont par exemple des mesures de probabilité, des mesures de distance à un hyperplan séparateur ou encore des mesures d'homogénéité de clusters de documents. Ce principe est illustré sur la figure 1.3.

Dans ce cadre, un opérateur d'agrégation numérique φ est une fonction qui réduit un ensemble de valeurs numériques à une unique valeur représentative ou significative (Detyniecki, 2002), illustré par exemple par les fonctions *maximum*, *minimum* ou *moyenne*. Formellement, un opérateur d'agrégation φ est une fonction définie comme :

$$\begin{aligned} \varphi : \quad & \mathbb{R}^L \to \mathbb{R} \\ & (\mu_1, \ldots, \mu_L) \mapsto \varphi(\mu_1, \ldots, \mu_L) \end{aligned}$$

et qui possède les propriétés suivantes :

$$\begin{aligned} \text{identité :} \quad & L = 1 \Rightarrow \varphi(\mu) = \mu \\ \text{croissance :} \quad & (\mu_1, \ldots, \mu_L) \leq (\sigma_1, \ldots, \sigma_L) \Rightarrow \varphi(\mu_1, \ldots, \mu_L) \leq \varphi(\sigma_1, \ldots, \sigma_L) \end{aligned}$$

Detyniecki (2002) organise cette classe de fonctions selon des propriétés supplémentaires : ainsi un opérateur présentant la propriété de renforcement positif produit un résultat d'autant plus élevé que les valeurs agrégées sont élevées. Inversement un opérateur présentant la propriété de renforcement négatif tend à supporter les valeurs faibles. On parle également d'opérateurs optimistes et pessimistes. Les opérateurs de compromis produisent un résultat toujours compris entre la plus petite et la plus grande valeur agrégée.

Agrégation non paramétrée Les fonctions *maximum* et *minimum* réalisent respectivement une disjonction et une conjonction des degrés de confiances, en ce sens ils représentent des opérateurs respectivement optimiste et pessimiste. Etant donné leur grande sensibilité aux valeurs extrêmes, ces opérateurs sont employés lorsque les f_l sont individuellement très robustes et que les erreurs commises sur chacune des représentations correspondantes ne sont pas de même type.

Les fonctions *somme* et *produit* sont deux autres opérateurs classiques d'agrégation. La somme présente toujours un renforcement positif faible, le comportement du produit change en fonction de l'intervalle de définition des valeurs. Sur l'intervalle unité il exhibe un renforcement négatif. Pour des valeurs toutes supérieures à 1 il réalise un renforcement positif. Par rapport au maximum et au minimum, ces opérateurs sont moins sensibles aux valeurs extrêmes : un espace de description est moins susceptible de dominer le degré

de confiance final. Il faut néanmoins noter que pour le produit, l'élément zéro est absorbant. En pratique, une petite quantité peut être ajoutée aux degrés de confiances pour y remédier.

En modélisant les $f(\boldsymbol{x})$ et les $f_l(\boldsymbol{x}^l)$ comme des distributions de probabilité sur l'univers $[1..K]$, Kittler et al. (1998) montrent que l'exploitation des opérateurs produit et somme revient à émettre une hypothèse d'indépendance entre les f_l. Pour l'opérateur somme une hypothèse supplémentaire est que pour tout espace de description d'origine, la distribution f_l est très proche de la « vraie » distribution des concepts cibles. Malgré l'importance de cette hypothèse, les auteurs justifient théoriquement le succès relatif de l'opérateur somme sur le produit en montrant que le premier est moins sensible aux f_l que le second. Par ailleurs, à partir de la somme est définie la moyenne qui réalise une agrégation équivalente mais qui constitue un opérateur de compromis. Il en découle alors que l'opérateur médiane est d'autant moins sensible que la somme et que le produit.

Au sein de ce document, nous ne couvrons pas l'étendue (très vaste) de l'ensemble des opérateurs d'agrégation numérique et de leurs propriétés, Detyniecki (2002) en propose une analyse théorique approfondie. Kittler et al. (1998) étudient l'utilisation d'opérateurs d'agrégation classiques dans le cadre de la fusion tardive.

Agrégation paramétrée Lorsque les espaces de description d'origine n'ont pas même importance, il est souhaitable de tenir compte de l'influence relative de chacun sur la décision finale. Pour le problème considéré, certains espaces peuvent effectivement contenir plus de bruit que d'autres ou plus simplement moins bien caractériser les concepts cibles. Dans ce cas, l'influence respective des f_l sur la décision finale est représentée par un vecteur de poids $\boldsymbol{\lambda} \in \mathbb{R}^L$ et nous notons $\varphi_{\boldsymbol{\lambda}}$ l'opérateur d'agrégation correspondant. Bien que de nombreux opérateurs puissent être paramétrés ainsi, nous considérons le cas où $\varphi_{\boldsymbol{\lambda}}$ produit une combinaison linéaire des degrés de confiance :

$$\varphi_{\boldsymbol{\lambda}} : (\mu_1, \ldots, \mu_L) \mapsto \sum_{l=1}^{L} \lambda_l \mu_l \in \mathbb{R}$$

Le vecteur de pondération peut être ajusté manuellement : les poids λ_l sont alors issus d'une expertise sur le domaine et sur le problème étudié. Une autre approche, automatique, consiste à exploiter une mesure de performance individuelle associée à chacun des f_l.

Une troisième méthode consiste à mettre en œuvre un apprentissage supplémentaire pour cet ajustement : de manière similaire à chacun des L processus d'apprentissage originaux, l'objectif est de minimiser les erreurs commises par f ou bien de maximiser un critère d'homogénéité pour les clusters de f. Cet apprentissage peut par ailleurs être réalisé sur les mêmes données qui ont servi à construire les f_l, il est néanmoins préférable de réserver un jeu de données supplémentaire à cet effet. Désormais, l'espace de description est formé autour des concepts cibles et les données correspondent aux décisions prises par les f_l sur chacun des documents du corpus considéré. Pour réaliser cet apprentissage l'espace de recherche est restreint en imposant aux poids $\boldsymbol{\lambda}$. Un ensemble de contraintes supplémentaires : lorsque les λ_l sont à valeurs dans l'intervalle unité et qu'ils somment à 1, l'opérateur recherché est par exemple un opérateur de compromis.

1.3.3.3 Limitations

En dépit des avantages que confère une fusion tardive pour des espaces d'origine hétérogènes, cette approche présente deux inconvénients majeurs détaillés ci-dessous.

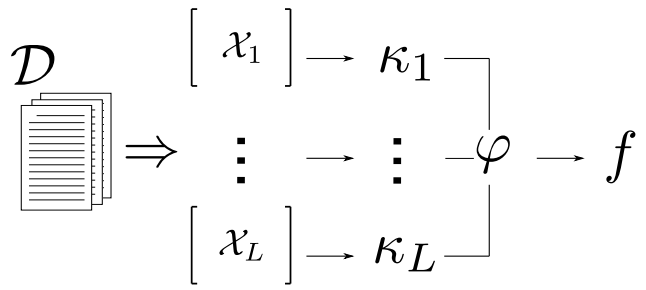


FIGURE 1.4 – Chaîne d'apprentissage pour la fusion intermédiaire.

Elle nécessite d'abord la mise en œuvre individuelle de L processus d'apprentissage : l'effort fourni pour le choix de la mesure de comparaison, celui de l'algorithme d'apprentissage mais aussi celui de son paramétrage est répété en conséquence. De plus, lorsque φ fait lui aussi l'objet d'un apprentissage, un processus supplémentaire et intrinsèquement différent des L processus sous-jacents, doit être réalisé. La fusion tardive constitue ainsi une méthode coûteuse à mettre en œuvre.

D'autre part, l'information contenue dans les $\mathcal{X}_l$ est préalablement compressée au niveau des f_l avant tout processus de fusion et les décisions finales $f(\boldsymbol{x})$ ne tiennent ainsi que peu compte des interactions entre les descripteurs originaux. A ce titre, l'apprentissage multi-vues (ou encore le co-apprentissage lorsque $L = 2$) offre un cadre différent pour réaliser cette fusion (Blum & Mitchell, 1998) : les f_l sont construits de manière conjointe, des contraintes d'accord leur sont imposées.

Enfin, il faut noter que comme représenté sur la figure 1.3, les possibilités de paramétrage intervenant aux trois niveaux de la chaîne d'apprentissage, les décisions finales peuvent devenir très complexes et difficiles à interpréter.

1.3.4 Fusion intermédiaire : agrégation de fonctions de similarité

Lorsque f exploite une mesure de similarité κ entre documents, une autre approche consiste à construire κ comme l'agrégation par φ de L mesures de similarité κ_l, chacune spécifique à chacun des espaces d'origine :

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\kappa_1(\boldsymbol{x}_i^1, \boldsymbol{x}_j^1), \ldots, \kappa_L(\boldsymbol{x}_i^L, \boldsymbol{x}_j^L))$$

L'apprentissage de f est alors réalisé à partir de κ. Ce principe est illustré sur la figure 1.4.

1.3.4.1 Motivations

La fusion intermédiaire offre un compromis entre la fusion anticipée et la fusion tardive. Lorsque φ réalise une combinaison linéaire des κ_l, la fusion anticipée représente en effet un cas particulier pour la fusion intermédiaire. Par exemple si les mesures de similarité sont des noyaux linéaires $\kappa_l(\boldsymbol{x}^l, \boldsymbol{z}^l) = \boldsymbol{x}^{l\intercal}\boldsymbol{z}^l$ et que de plus φ réalise la somme de ses arguments, on a :

$$\begin{aligned}\kappa = \varphi(\kappa_1, \ldots, \kappa_l) = \sum_{l=1}^{L} \kappa_l = \sum_{l=1}^{L} \boldsymbol{x}^{l\intercal}\boldsymbol{z}^l = \sum_{l=1}^{L}\sum_{j=1}^{m} x_j^l z_j^l \\ = (\boldsymbol{x}^1 \oplus \boldsymbol{x}^L)^\intercal(\boldsymbol{z}^1 \oplus \boldsymbol{z}^L)\end{aligned}$$

De plus, contrairement à la fusion tardive, l'information est compressée à un niveau plus proche des données d'origine. Cette fusion permet ainsi la prise en compte d'interactions plus fortes entre les espaces d'origine. En outre, elle est généralement plus aisée à mettre en œuvre que la fusion tardive : f étant unique, une seule chaîne d'apprentissage est nécessaire. Une conséquence immédiate est que la taille de l'espace de recherche pour construire f ainsi que les temps de calcul peuvent être grandement réduits.

En revanche, à l'inverse des deux autres types de fusion, la fusion intermédiaire influence directement le choix de l'algorithme d'apprentissage mis en œuvre pour construire f : f est contraint d'exploiter une notion de similarité entre les documents. Dans la suite, nous nous plaçons dans le cadre où f est construit à partir d'un noyau tel que présenté dans la section 1.1.3.3, p. 18 et nous considérons donc que les κ_l sont tous des noyaux valides. Ce choix est motivé par le succès des noyaux pour les tâches d'apprentissage sur des représentations textuelles, mais aussi par l'adaptabilité des fonctions noyaux qui peuvent aussi bien représenter des interactions simples pour un noyau linéaire que des interactions plus complexes pour un noyau non linéaire comme le noyau gaussien.

1.3.4.2 Opérateurs d'agrégation pour noyaux

Tout comme pour la fusion tardive, le choix d'un opérateur d'agrégation numérique est motivé par les propriétés souhaitées pour le processus de fusion.

La différence est qu'ici φ n'agrège pas des degrés de certitude mais des mesures de similarité et en particulier des noyaux. Parmi les opérations qui conservent la propriété de noyaux (voir tableau 1.2, p. 19), la somme et le produit de L noyaux est par exemple un noyau. Les opérateurs qui conservent la propriété de noyaux sont rappelés à la section 1.1.3.3, p. 18.

1.3.4.3 Ajustement des poids pour opérateurs paramétrés

Parmi l'ensemble des opérateurs qui conservent la propriété de noyaux, certains sont paramétrés par un vecteur de poids $\boldsymbol{\lambda} \in \mathbb{R}_+^L$, nous considérons ici le cas linéaire. Les poids λ_l caractérisent alors l'importance associée à chacune des L représentations originelles, et ont pour effet de dilater ou de contracter les espaces de caractéristique induits par chacun des noyaux. Ce vecteur peut être fixé manuellement, dans la suite nous décrivons deux approches pour son ajustement automatique. Pour la première, un critère de qualité évalue un ensemble de vecteurs candidats, tandis que pour la seconde les poids sont obtenus en phase d'apprentissage, conjointement à f. Pour l'ensemble des méthodes présentées, ce dernier est un classifieur entraîné sur une base d'apprentissage, ces méthodes forment dans leur ensemble un domaine de recherche à part entière : l'apprentissage par noyaux multiples (*multiple kernel learning*). Une présentation plus détaillée de ce domaine est donnée en annexe de ce document à la section A, p. 197.

Approches heuristiques Une première catégorie de méthodes identifie κ en amont de l'apprentissage de f. Pour ce faire, les auteurs proposent d'exploiter le vecteur des étiquettes de classes $\boldsymbol{y}$ et à évaluer de manière indépendante l'influence de chacun des λ_l sur le noyau final (Gönen & Alpaydin, 2011).

Une première approche consiste ainsi à associer au poids λ_l une mesure des performances du classifieur f_l entraîné sur X_l au travers de κ_l (Tanabe et al., 2008). Cette approche présente cependant un inconvénient : comme pour la fusion tardive, L processus d'apprentissage doivent être réalisés de manière indépendante.

Une seconde approche repose uniquement sur les étiquettes de classes : le noyau idéal indique une similarité maximale, 1, pour des paires de données appartenant à une même classe et une similarité minimale, 0, pour des paires de classes différentes (Cristianini et al., 2002). Pour un problème de classification binaire, la matrice de similarité associée au noyau idéal est une matrice à valeurs binaires, indicatrice des regroupements donnés par les étiquettes. Un noyau dont la matrice de similarité produit des regroupements proches se voit alors accorder un poids important : ce poids peut être exprimé de manière proportionnelle à cette proximité (Qiu, 2009), il peut aussi faire l'objet d'un apprentissage (Lanckriet et al., 2004a; Igel et al., 2007).

Approches simultanées Contrairement aux approches heuristiques, les approches simultanées produisent conjointement le vecteur de poids et le classifieur : l'identification des poids est guidée par le problème considéré. Lorsque des contraintes de parcimonie sont prises en compte, l'apprentissage par noyaux multiples offre de plus un cadre bien fondé pour la sélection de noyaux, traditionnellement réalisé empiriquement par validation croisée (Bach et al., 2004).

Cet apprentissage peut être réalisé en une passe : l'algorithme mis en œuvre pour construire f produit également $\boldsymbol{\lambda}$, on parle alors de *méthodes directes* (Lanckriet et al., 2004b; Bach et al., 2004). L'apprentissage peut aussi être réalisé en deux passes itérées : tandis que f est fixé $\boldsymbol{\lambda}$ est mis à jour, puis $\boldsymbol{\lambda}$ est à son tour fixé tandis que f est mis à jour, jusqu'à convergence de l'algorithme. On parle alors de *méthodes enveloppantes* (Sonnenburg et al., 2006; Rakotomamonjy et al., 2008; Kloft et al., 2009; Xu et al., 2010). Pour ces dernières, l'apprentissage de f est communément désigné par *problème maître* tandis que celui du vecteur de poids par *problème esclave*. Les méthodes enveloppantes présentent un intérêt particulier étant donné que l'apprentissage de f pour un $\boldsymbol{\lambda}$ fixé peut bénéficier des développements passés et futurs d'algorithmes d'apprentissage classiques.

1.4 Bilan

Dans ce chapitre nous avons rappelé les différents enjeux liés à la représentation des données en apprentissage et nous avons vu que de nombreux choix sont effectués en amont de tout processus d'apprentissage. Les descripteurs qui composent l'espace de représentation et qui extraient des données étudiées l'information considérée sont de deux type : les descripteurs bas niveau représentent une information au plus proche des données ; pour répondre au problème du fossé sémantique, des descripteurs de plus haut niveau visent la prise en compte d'enrichissements supplémentaires, par exemple syntaxiques ou sémantiques dans le cas du texte.

Nous avons également rappelé l'intime relation qui existe entre la description faite des données et la mesure de comparaison employée pour les traiter. A ce titre nous avons vu, au travers des fonctions noyaux, qu'une mesure de similarité transforme la géométrie de l'espace de description et permet ainsi de remodeler l'information en fonction du problème considéré. Le produit scalaire présente par exemple de bonnes propriétés pour les représentation textuelles ; sa version normalisée, la similarité angulaire, transforme de plus l'espace de représentation en une hypersphère et ne retient de l'information que la direction donnée dans cet espace.

Dans certains cas, plusieurs modes de l'information sont nécessaires pour décrire les concepts considérés. Nous avons rappelé qu'une fusion de représentations multiples peut opérer à différents niveaux de la chaîne d'apprentissage. Dans la partie I nous exploitons

les méthodes présentées afin de décrire au mieux des concepts ambigus, pour lesquels il n'existe pas un vocabulaire naturellement discriminant.

Enfin, le vocabulaire employé au sein des documents étudiés est souvent trop large pour décrire à bien les concepts étudiés, nous avons présenté des méthodes de filtrage spécifiques au texte, ainsi que des méthodes de réduction plus générales, dites numériques. A cet effet, nous avons mis en évidence le lien étroit qui existe entre un tâche d'apprentissage linéaire et une sélection des descripteurs pertinents. Tout au long de ce document nous sommes confrontés au problème de l'identification de descripteurs pertinents. Dans la partie II en particulier, nous proposons une méthode originale pour réaliser un partitionnement ≪ parcimonieux ≫ de données en environnement dynamique.

Première partie

Informations émotionnelles

Dans cette partie nous considérons une information subjective puisque émotionnelle pour laquelle les méthodes d'apprentissage traditionnelles sont mises en difficulté. Les concepts émotionnels sont en effet complexes à décrire dans les documents, par opposition à des concepts thématiques plus classiques : le fossé sémantique, entre l'information bas niveau que représente les mots d'un documents et l'information haut niveau recherchée, est plus important.

Au chapitre 2 nous faisons état des modèles de représentation classiques, issus des travaux en psychologie et en linguistique, pour décrire les émotions. Nous présentons les principales approches considérées ainsi que les méthodes étudiées pour traiter des émotions dans les textes.

Au chapitre 3 nous étudions une tâche de discrimination des émotions sur un corpus récemment proposé dans le cadre d'une compétition. Nous proposons une approche et une méthode reposant sur l'exploitation de descripteurs bas niveau pour associer des étiquettes émotionnelles à un vocabulaire dont les entrées sont constituées de manière plus riche que traditionnellement.

Nous étudions une approche différente au chapitre 4 qui repose sur un enrichissement sémantique des documents. Nous exploitons à cet effet une représentation fine et graduelle des émotions, ainsi qu'un lexique associant aux mots d'un vocabulaire générique, des coordonnées dans un espace sémantique multidimensionnel.

Enfin, au chapitre 5, nous proposons un modèle pour décrire les émotions, adapté à la modalité particulière du texte. Il permet d'une part une caractérisation fine des états affectifs, d'autre part une discrimination des émotions selon des étiquettes émotionnelles pré-définies. Cette proposition s'inscrit dans le cadre du projet DoXa et est par ailleurs motivée par la constitution d'un corpus d'apprentissage étiqueté sur des concepts émotionnels.

Chapitre 2

Méthodes pour l'analyse d'informations émotionnelles

Dans ce chapitre nous considérons un domaine dans lequel les concepts cibles sont subjectifs et imprécis : dans ce contexte, la mise en œuvre d'un apprentissage ne fournit pas toujours les résultats escomptés. Les émotions constituent des concepts complexes, comme le montre la multiplicité des modèles psychologiques proposés pour les représenter et parmi lesquels aucun ne fait consensus. Dans le cadre de l'apprentissage pour des représentations textuelles se posent alors de nombreux défis liés à la fois à la représentation des documents pour des concepts affectifs, à la disponibilité de bases d'apprentissage pour construire des modèles de discrimination ainsi qu'aux ambiguïtés et aux imprécisions inhérentes à ces concepts.

A la section 2.1, nous présentons le domaine de l'*affective computing* qui vise à l'analyse automatique des émotions, et nous considérons plus particulièrement le cas du texte. Nous présentons à la section 2.2 les principaux modèles de représentation, proposés en psychologie ainsi qu'en linguistique, pour décrire les émotions. A la section 2.3, nous faisons état des spécificités liées à la représentation des documents pour une tâche d'analyse des émotions, et nous présentons les approches considérées de manière classique pour l'aborder à la section 2.4. Enfin, les conclusions de ce chapitre sont données à la section 2.5.

Une partie de ces travaux a été publié dans une conférence (Dzogang et al., 2010b).

2.1 Affective computing et textes

Dans cette section nous présentons brièvement le domaine général de l'*affective computing* et nous considérons plus particulièrement le cas du texte : les travaux dans ce domaine peuvent être divisés en deux selon que le modèle de représentation des émotions utilisé consiste en une catégorisation bi-classe, positif/négatif, ou une représentation plus fine autorisant notamment la définition de concepts affectifs complexes, imprécis et ambigus.

2.1.1 Affective computing

L'analyse des émotions et de manière plus générale des états affectifs a connu un développement récent et important. Ce domaine de recherche consiste à étudier les opinions, les sentiments et les émotions à partir de signaux physiologiques, d'expressions faciales ou encore de textes rédigés en langue naturelle. Des exemples d'applications de ce domaine de recherche incluent la robotique, les interfaces intelligentes ou encore le sondage automatique sur Internet (Picard et al., 2001). Plusieurs tâches l'organisent, notamment,

la simulation d'agents émotionnels, la caractérisation fine de contenu émotionnel, ou la discrimination entre des concepts émotionnels.

2.1.2 Cas du texte

L'identification et la classification de concepts positifs ou négatifs sont étudiées dans la tâche de l'*opinion mining* ; l'étude d'états représentés plus finement est abordée dans la tâche de l'*emotion mining*. Pour cette dernière, un concept peut être complexe lorsqu'il est le résultat d'une composition d'états basiques (e.g. une joie mêlée de crainte), imprécis lorsqu'il fait référence de manière indirecte à un état basique (e.g. une grande colère ou un léger énervement), ou ambigu lorsque sa sémantique est très liée à son contexte (e.g. un état de grande excitation). Ces deux tâches se réunissent dans le cadre plus général de l'*affective computing* présenté à la section précédente.

Tandis que l'*opinion mining* a reçu beaucoup d'attention (Wiebe, 2009), moins de travaux ont porté sur le domaine voisin de l'analyse des émotions. L'*emotion mining* consiste en la discrimination des émotions exprimées dans un corpus étiqueté comme par exemple la *colère*, l'*amour*, ou la *tristesse*, cette tâche consiste également en la caractérisation fine de la charge émotionnelle associée aux documents.

Nous pouvons supposer que le manque de consensus sur les modèles de représentation des émotions, la difficulté d'annoter des corpus de documents mais aussi la complexité de l'analyse des émotions dans les textes ont grandement participé à ce phénomène. Au contraire, le succès de l'*opinion mining* peut s'expliquer par la simplicité des modèles de représentation employés (les émotions sont divisées selon un modèle bi-classe positif/négatif) mais aussi par la disponibilité de données étiquetées comme par exemple les commentaires utilisateurs sur Internet (*user ratings*). Pour la modalité particulière du texte, Pang et Lee (2008) organisent les travaux portant sur l'étude des états affectifs en plusieurs sous-tâches, que nous présentons et complétons ci-dessous :

Classification subjectif/objectif La classification des documents d'un corpus en une catégorie *subjectif* et une catégorie *objectif* est très liée aux problèmes posés en analyse des émotions. De nombreux travaux ont été proposés pour identifier automatiquement le contenu subjectif au sein d'un corpus de document : il est classique d'aborder cette tâche comme un problème de classification binaire pour lequel l'enjeu est la construction d'une frontière de décision qui sépare les étiquettes subjectives des étiquettes objectives.

Identification du sujet et de l'objet d'une émotion Un autre enjeu lié à l'analyse des émotions est l'identification de la source associée à un état affectif. Ces travaux trouvent racine en traitement du langage naturel et abordent par exemple le problème de la résolution des anaphores.

De manière similaire, l'association entre un état affectif et l'objet d'affection est un problème très lié à l'analyse des émotions (Duthil et al., 2012). Dans le domaine de l'*opinion mining* par exemple, ces travaux visent par exemple à identifier les attributs d'un objet sur lesquels portent les commentaires positifs ou négatifs d'utilisateurs sur Internet.

Discrimination des émotions L'enjeu principal des méthodes proposées en *affective computing* porte sur la reconnaissance automatique de concepts affectifs à partir de signaux non affectifs. Pour la modalité particulière du texte, ces travaux proposent alors de résoudre une tâche de classification bi-classe pour le domaine de l'*opinion mining*

ou multi-classes pour celui de l'*emotion mining*. Selon les corpus étudiés ces méthodes abordent les problématiques liées à l'apprentissage multi-étiquettes mais aussi ceux liés au déséquilibre des classes. Cette tâche est reconnue comme difficile, de nombreuses pistes sont explorées comme par exemple l'apprentissage semi-supervisé ou l'exploitation conjointe de représentations bas niveau et d'enrichissements sémantiques.

Caractérisation fine des états affectifs Un autre enjeu consiste en la caractérisation automatique de la charge émotionnelle portée par les documents. Ces travaux s'ancrent dans les méthodes d'apprentissage non supervisé et visent à identifier automatiquement des états affectifs associés à un document. Pour pallier le fossé sémantique important entre les mots composants les textes et les émotions qui y sont exprimées, des enrichissements sémantiques sont mis en œuvre.

Positionnement de nos travaux Dans ce chapitre nous concentrons notre étude sur les deux dernières tâches présentées ci-dessus. Comme présenté précédemment, au vue de la difficulté de la tâche, de nombreux auteurs associent aux concepts affectifs une représentation bi-classe des émotions et considèrent une tâche d'*opinion mining*. Nos travaux s'inscrivent dans un cadre d'étude plus complet, dans lequel les modèles de représentation employés permettent de représenter des opinions mais aussi des émotions. Dans la suite de ce document nous traitons ces deux tâches de manière indifférenciée, lorsqu'une spécificité doit être relevée nous le remarquons.

Dans un premier temps nous faisons état des modèles psychologiques pour la représentation des émotions. De nombreux modèles ont été proposés, leur choix est entre autres motivé par la tâche considérée, les données disponibles ou encore le type d'apprentissage mis en œuvre. Nous discutons ensuite des spécificités liées à la construction d'un espace de représentation dédié à l'analyse des émotions dans les corpus de textes. Enfin, nous faisons état des approches classiques pour réaliser cette analyse : une première approche consiste à adapter les méthodes classiques en apprentissage pour des concepts plus généraux. Une deuxième approche exploite des ressources spécifiquement liées aux émotions mais ne réalisent pas nécessairement un apprentissage. Enfin, une dernière approche propose d'étudier un enrichissement sémantique pour des représentations bas niveau.

2.2 Modélisation des états affectifs pour les textes

Dans ce document, nous employons les termes *sentiment* et *émotion* de manière équivalente, bien que la littérature les distingue notamment de par leur durée (Scherer, 2005). Dans la suite, nous parlons d'émotion lorsque nous faisons référence à un concept ou à un état affectif qui présente différents axes de caractérisation (comme détaillé dans la suite la polarité, l'intensité, ou une catégorisation sémantique par exemple). Lorsque les concepts sont uniquement différenciés selon une échelle bipolaire positif/négatif, dans ce document nous préférons également le terme d'émotion à celui d'opinion.

La question de la représentation des émotions est un problème abondamment étudié par les psychologues : bien qu'il n'existe pas un modèle consensuel pour décrire et caractériser les états affectifs, il est possible d'organiser les modèles proposés selon trois familles. Les *modèles événementiels* décrivent le mécanisme de déclenchement émotionnel à l'aide de règles d'évaluation de l'environnement (Scherer, 1981), ils sont principalement employés pour la simulation d'agents affectifs et sortent donc du cadre de nos travaux. Dans la

suite nous présentons deux modèles de représentation utilisés de manière classique pour la détection automatique des émotions : les *modèles catégoriels* et les *modèles dimensionnels*. Nous présentons également les *modèles linguistiques*, qui trouvent racine dans le domaine du traitement automatique des langues.

Modèles catégoriels Ces modèles reposent sur une vision darwinienne de l'évolution selon laquelle les émotions résultent d'un mécanisme de survie et constituent une condition nécessaire pour la préservation des espèces. Selon cette vision le sentiment de *peur* par exemple peut être interprété comme une réponse de survie face à un danger imminent. La vision darwinienne impose que les émotions soient organisées autour d'un ensemble fini d'états affectifs. Les états de cet ensemble sont alors vus comme des étiquettes émotionnelles appelées selon les auteurs *primaires* ou *basiques* (Johnson-Laird & Oatley, 1989; Plutchik, 1990; Ortony & Turner, 1990; Ekman, 1999). Pour certains auteurs, des combinaisons particulières d'émotions primaires forment des *émotions complexes*. Aussi, Plutchik (1990) définit le *mépris* comme le mélange de deux émotions primaires, l'*ennui* et l'*agacement* (voir figure 2.1). Selon les auteurs, d'autres axes peuvent organiser les étiquettes émotionnelles : l'intensité représente par exemple l'ampleur d'un état affectif et la polarité caractérise le caractère positif ou négatif d'une émotion.

La figure 2.1 représente la rosace émotionnelle issue de la catégorisation des émotions proposée par Plutchik (1990) : les états primaires composent le cœur de la rosace, les pétales sont structurés en trois niveaux d'intensité qui spécialisent les émotions primaires, et entre les pétales les émotions complexes sont obtenues en combinant les états primaires correspondants.

Bien que parmi les travaux psychologiques il n'existe pas de consensus sur le nombre et la nature des émotions primaires, il est classique dans le domaine de l'affective computing de faire référence au *big six set* (Ekman, 1999) composé des émotions *peur*, *colère*, *joie*, *tristesse*, *surprise* et *dégoût* (Cowie & Cornelius, 2003).

Modèles dimensionnels Une autre vision considère que tous les objets (par exemple les mots d'une langue ou les sons d'une ville) portent une charge émotionnelle dont la valeur varie selon la culture, l'âge ou encore les expériences personnelles (Scherer, 2005). Ainsi, leur charge émotionnelle est-elle évaluée sur des échelles de mesure continues ; les émotions sont alors représentées comme des vecteurs à valeurs réelles dans un espace sémantique multi-dimensionnel.

Parmi les dimensions les plus utilisées, la *valence* représente le plaisir procuré, l'*activation* mesure l'excitation physique causée et le *contrôle* représente la capacité à surmonter. Bien qu'il n'y ait pas de consensus sur le nombre et la nature des dimensions, pour de nombreux psychologues les états affectifs peuvent être représentés par un modèle bi-dimensionnel composé de la valence et de l'activation (Barrett & Russell, 1999) (voir figure 2.2). Pour d'autres auteurs, la dimensions de contrôle est de plus nécessaire, pour distinguer le concept de *peur* de celui de la *colère* par exemple (Fontaine et al., 2007).

La figure 2.2 représente un modèle bi-dimensionnel pour décrire les émotions, l'espace sémantique est le plan déterminé par les axes de valence et d'activation. Un ensemble d'émotions primaires sont disposées sur ce plan, leur agencement décrit la proximité sémantique des concepts correspondants : deux émotions proches sur l'axe de la valence ont par exemple même caractère positif ou négatif. En dehors des émotions primaires, des états plus imprécis et plus complexes sont représentés comme par exemple les concepts *at ease* ou *annoyed*. Ces derniers ne dénotent pas directement un état basique mais ils les connotent voire ils les combinent.

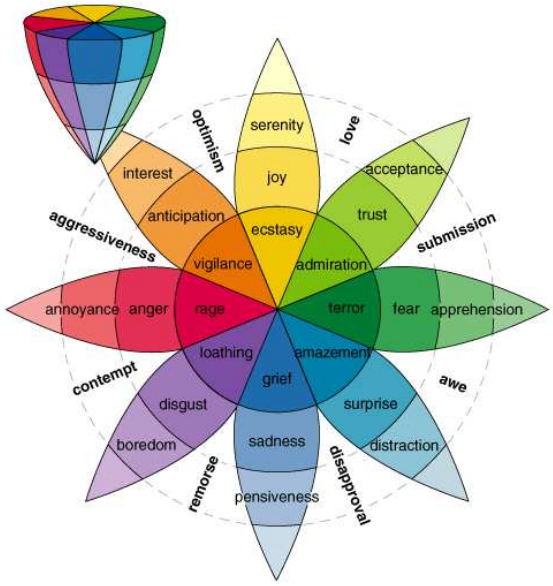


FIGURE 2.1 – Rosace pour décrire une catégorisation émotionnelle : un ensemble d'émotions primaires composent des émotions complexes (entre les pétales), et se déclinent en trois états d'intensité (sur les pétales) (Plutchik, 1990).

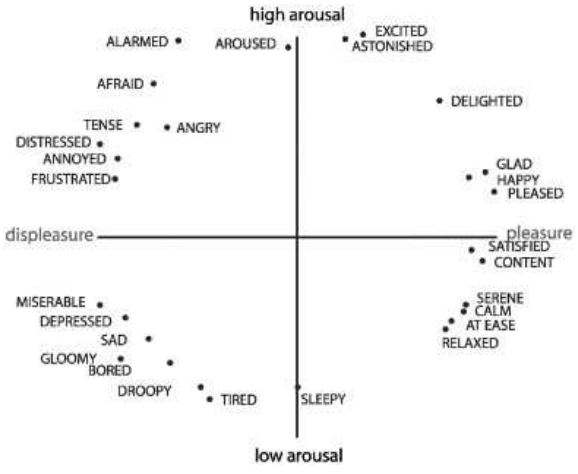


FIGURE 2.2 – Espace émotionnel bi-dimensionnel composé des axes de valence (abscisse) et d'activation (ordonnée) (Russell, 1980).

Modèles linguistiques Reposant sur la sémantique liée aux étiquettes émotionnelles, d'autres modèles ont été proposés dans le domaine du traitement automatique des langues : les émotions primaires sont alors considérées comme des *méta-émotions* spécialisées en *sous-émotions* au sein d'une hiérarchie dont les liens sont entre autres sémantiques. Ces modèles sont souvent appelés *modèles de poupées russes*.

A partir de méthodes automatiques d'analyse de corpus, Mathieu (2006) propose par exemple une classification sémantique des verbes et des noms subjectifs. Ce modèle est structuré comme un graphe implémentant trois types de relations : l'héritage sémantique, l'antinomie et l'intensité.

Piolat et Bannour (2009) proposent une autre forme de représentation en étudiant de manière systématique la sémantique associées aux étiquettes : les émotions sont dans un premier temps réparties selon un axe positif/négatif, puis dans un second temps selon leurs relations sémantiques. Dans ce modèle les émotions sont donc organisées en une hiérarchie dont les premiers niveaux spécialisent les concepts *positif* et *négatif* et les suivants précisent la sémantique des niveaux supérieurs.

Enjeux pour le texte Cowie et Cornelius (2003) étudient les représentations catégorielles et dimensionnelles des émotions ainsi que leurs emplois pour le traitement automatique de la parole et de la langue. Il ressort de leur étude que les états affectifs, souvent ressentis et exprimés de manière imprécise, nécessitent des modèles de représentation adaptés, qui présentent entre autres des propriétés de gradualité.

Dans (Dzogang et al., 2010b) nous proposons une étude comparative des principaux modèles pour lesquels nous analysons le rôle de la gradualité. En particulier, nous organisons ces modèles autour de trois composantes de gradualité :

- la *composition* vise à représenter un état affectif comme un mélange d'états primaires (issus d'une catégorisation des émotions). Elle permet de caractériser des transitions d'états affectifs ou de modéliser des états complexes dont la nature est ambiguë.
- l'*intensité* permet de décrire les états affectifs sur une échelle allant des états platoniques aux états passionnés.
- l'*héritage* exploite une hiérarchisation de concepts affectifs qui spécialise ces concepts selon la sémantique qu'ils portent.

Pour l'analyse automatique de textes, nous montrons que l'intensité peut être utilisée afin de caractériser l'action des modificateurs linguistiques d'intensité comme par exemple *très* ou *peu*. L'héritage permet de spécialiser une émotion selon son contexte d'énonciation. La composition permet de modéliser des expressions subtiles comme par exemple *à la fois en colère et triste*. Dans leur ensemble, ces trois composantes issues des modèles de représentations des émotions peuvent être exploitées afin de tenir compte de la nature complexe, imprécise et ambiguë des émotions, en vue de leur analyse automatique dans les documents. Bien sûr, l'utilisation pratique de ces composantes constitue un défi et demande des résultats expérimentaux : dans le chapitre 5 nous étudions une modélisation des émotions destinée à leur étude automatique dans le texte.

2.3 Spécificité des descripteurs

La complexité d'une tâche d'apprentissage pour des concepts affectifs peut en grande partie s'expliquer par la difficulté à décrire des concepts subjectifs, subtils et ambigus. Dans cette section nous reprenons la structure adoptée à la section 1.1 pour présenter les différentes approches étudiées dans la littérature : pour chacune des méthodes présentées, nous mettons en évidence les particularités liées à l'étude de concepts affectifs. Ainsi, dans

un premier temps nous considérons le cas des descripteurs bas niveau, nous détaillons à ce titre les entrées considérées pour constituer un espace de représentation fidèle aux concepts étudiés. Dans un second temps nous présentons les enrichissements sémantiques adoptés pour répondre au problème du fossé sémantique entre le vocabulaire d'un corpus et les émotions qui le composent. L'enjeu de ces méthodes est alors la constitution d'un espace de représentation au plus proche des concepts étudiés.

2.3.1 Descripteurs bas niveau

Comme nous l'avons rappelé à la section 1.2.1.1 p. 20, la constitution d'un espace de représentation bas niveau consiste à extraire, du corpus d'étude, un vocabulaire qui décrit une information fidèle au problème considéré. Pour ce faire, de nombreux choix sont à effectuer, dans leur ensemble, ils constituent un processus de décision en plusieurs étapes. Initialement, un dictionnaire dont les entrées représentent le vocabulaire employé dans les documents est extrait du corpus d'étude. Ce dernier décrit alors l'ensemble des mots qui apparaissent de manière unique dans le corpus. Il peut par exemple s'agir des termes, des marqueurs de ponctuation ou de toute autre entrée jugée pertinente pour le problème considéré. Une information plus riche peut par ailleurs être obtenue en considérant des combinaisons de telles entrées. Ensuite, un filtrage du vocabulaire vise à éliminer le bruit de l'information ainsi extraite : les filtres employés dépendent des concepts étudiés, nous présentons les méthodes employées dans le cas des émotions. Enfin, une fois le dictionnaire filtré, ses entrées définissent l'espace de représentation pour décrire l'information extraite des documents. Un schéma de comptage permet alors d'associer aux entrées une mesure de leur importance pour les documents.

Dans la suite nous présentons les particularités liées à l'étude des émotions pour chacune de ces étapes.

Mots uniques Le vocabulaire d'un corpus véhicule l'information qu'il contient, il est composé de l'ensemble des termes qui apparaissent de manière unique dans les documents à l'exception des mots jugés non porteurs de sens pour le problème considéré. Dans le cas des émotions, le sens accordé aux mots diffère cependant du cadre classique et de nouveaux descripteurs sont considérés.

Inspirée par les travaux sur les études de style, une approche consiste par exemple à intégrer au vocabulaire l'ensemble de la ponctuation employée (Mishne, 2005; Dzogang et al., 2010a). En effet, cette dernière contient des marqueurs naturellement pertinents pour décrire des concepts affectifs : le point d'exclamation en est l'exemple le plus manifeste. Avec l'essor des conversations Internet, de nouveaux mots ont par ailleurs été adoptés pour expliciter les émotions qui nuancent les textes. Les *émoticones* sont par exemple des successions de marqueurs de ponctuation, dont l'arrangement exprime un état précis comme la joie, la colère ou l'ennui. Certains auteurs proposent d'intégrer ces nouveaux mots au vocabulaire extrait (Mishne, 2005; Neviarouskaya et al., 2007; Go et al., 2009).

De plus, de nombreux travaux ont montré qu'une grande part des émotions contenues dans les textes sont exprimées par les verbes (e.g. *apprécier*), les adverbes (e.g. *heureusement*), ainsi que les adjectifs (e.g. *extraordinaire*) (Hatzivassiloglou & Wiebe, 2000; Dray et al., 2009). Ces descripteurs traditionnellement écartés du vocabulaire sont alors explicitement extraits pour représenter des concepts affectifs.

Enfin, contrairement au cadre classique, la négation joue un rôle déterminant pour les émotions : les expressions *ravie* et *pas ravie* représentent une information évidemment différente. Bien que la détection des marqueurs de négation soit une tâche réputée difficile,

certaines approches reposent sur une identification simple de ces derniers, et proposent de doubler la taille du vocabulaire afin d'inclure le préfixe *not* aux descripteurs qui subissent l'effet d'une négation (Das & Chen, 2007).

Combinaisons de mots L'emploi de mots uniques repose sur l'hypothèse qu'il existe un vocabulaire simple pour bien décrire les concepts cibles. Cependant, pour des concepts affectifs, de nombreux résultats (Pang & Lee, 2008; Mejova & Srinivasan, 2011) contrastent avec cette hypothèse. Une approche consiste alors à constituer un vocabulaire plus riche en considérant une combinaison des descripteurs présentés précédemment. Pour ce faire, il est d'usage d'exploiter des p-grammes pour des ordres $p \geq 2$ mots : comme rappelé à la section 1.1.1.1, p. 8 un descripteur décrit alors un mot pris dans son contexte d'énonciation.

Cette approche permet notamment de tenir naturellement compte de l'effet de la négation sur le sens porté par les mots. En utilisant la ponctuation il est de plus possible d'utiliser les points d'exclamation pour extraire les interjections des documents (e.g. *oh !*), ou encore d'exploiter les points qui terminent les phrases pour tenir compte de la position des descripteurs. En exploitant un grand corpus de documents, Cui et al. (2006) montrent par exemple une nette amélioration des résultats lorsque les descripteurs décrivent des morceaux de phrases pour des ordres allant jusqu'à $p = 6$ mots. Une approche similaire dont l'efficacité a été montrée consiste à utiliser simultanément des combinaisons de différents ordres (Mejova & Srinivasan, 2011).

Enfin, de manière indépendante à la représentation en sacs de mots, d'autres types de description ont été étudiés afin de tenir compte du contexte d'apparition des mots, c'est notamment le cas des arbres de dépendances syntaxiques (Pak & Paroubek, 2010).

Filtrage Comme rappelé à la section 1.2, p. 19, de nombreuses méthodes sont employées pour éliminer le bruit de l'information ainsi extraite. En revanche, afin de décrire au mieux les émotions, l'ensemble des descripteurs considérés est souvent retenu : selon le problème considéré il est en effet difficile de juger de leur pertinence a priori, la rareté de leurs occurences dans les corpus pose de plus obstacle aux méthodes de sélection numériques (Rafrafi et al., 2012).

Certains auteurs effectuent néanmoins un filtrage sévère des descripteurs et ne retiennent de l'information que certaines classes grammaticales comme les verbes, les adverbes ou les adjectifs (Chesley, 2006; Benamara et al., 2007). La description faite des documents n'a alors plus les propriétés des représentations textuelles et comme nous le présentons dans la suite (voir section 2.4.2), les analyses effectuées ne mettent généralement pas en œuvre un apprentissage mais une agrégation des concepts identifiés.

Méthodes de comptage Dans le cadre classique, le schéma de comptage tf/idf est le plus employé, il permet en effet de rendre compte de l'importance des mots pour un document tout en tenant compte de son pouvoir de discrimination dans un corpus. Pour des concepts affectifs en revanche, certaines études montrent que seule l'information de présence ou d'absence des mots est nécessaire (Ng et al., 2006). Cependant d'autres études montrent qu'un schéma de comptage fréquentiel comme celui du tf/idf est équivalent, dans certains cas meilleur (Mejova & Srinivasan, 2011). Pour des concepts affectifs, le choix d'une méthode de comptage semble ainsi dépendre de la nature des concepts étudiés mais de manière plus générale de la nature des données étudiées.

Vers des règles de construction générales ? Il est difficile d'établir des règles générales pour la construction d'un espace de représentation bas niveau. Tandis que la ponctua-

tion, les verbes, les adverbes ainsi que les adjectifs semblent généralement offrir une représentation discriminante pour ces concepts, les meilleures stratégies de construction de l'espace de représentation varient souvent d'une étude à l'autre. Des indices semblent indiquer que la taille du corpus d'étude, la nature des concepts affectifs et la nature du vocabulaire employé ont un rôle déterminant sur le choix des descripteurs (Mejova & Srinivasan, 2011). De même, tandis que certaines études montrent l'intérêt de combiner ces descripteurs au travers de p-grammes d'ordres élevés par exemple, comme nous le rappelons à la section 1.1.1.1, à des ordres élevés les p-grammes se font plus rares dans les documents et le succès de telles combinaisons dépend grandement de la taille du corpus considéré.

Par ailleurs, il faut noter que de nombreuses méthodes qui sortent du cadre de nos travaux sont proposées dans le domaine du traitement automatique des langues. Ces dernières reposent par exemple sur l'exploitation de grammaires semi-automatiques. Bien que ces méthodes permettent d'extraire des descripteurs plus complexes de manière plus subtile, elles nécessitent un effort important de mise en œuvre et souffrent de plus d'un problème de généralisation à d'autres langues ou à d'autres thématiques.

2.3.2 Enrichissements sémantiques

Comme nous l'avons observé à la section 1.1.1.1, p. 8, la construction d'un espace de représentation bas niveau fait face au problème du fossé sémantique : un descripteur est souvent un symbole dénué de sémantique pour les concept étudiés. Ici, l'emploi d'enrichissements sémantiques peut pallier ces limitations et en particulier agir sur trois points : 1) obtenir des descripteurs plus proches des concepts affectifs étudiés pour remédier à l'ambiguïté, à la subtilité ainsi qu'à la complexité de l'information nécessaire à les décrire, 2) proposer une alternative au manque de données étiquetées selon des concepts affectifs, 3) combler la rareté des descripteurs reconnus comme porteurs d'émotions dans les corpus d'étude. Ces enrichissements prennent la forme de lexiques affectifs qui rassemblent un vocabulaire autour de concepts définis dans le modèle employé pour représenter les émotions.

Dans un premier temps nous présentons le cas des lexiques génériques qui regroupent un vocabulaire qui se veut exhaustif pour une langue. Dans un second temps, nous présentons différentes méthodes visant à une spécialisation des lexiques pour le corpus d'étude.

2.3.2.1 Lexiques génériques

Il est courant d'employer des listes de mots dont la charge émotionnelle est importante (*sentiment words*) pour guider la construction de l'espace de représentation. Un résumé des ressources disponibles dans la littérature est rappelé dans le tableau 2.1, p. 53. Dans un premier temps nous présentons les lexiques issus des travaux précurseurs en psychologie ou en linguistique, nous détaillons ensuite des méthodes pour étendre automatiquement ces derniers à un vocabulaire plus important.

Travaux psychologiques et linguistiques Les premières listes de mots structurées selon des concepts affectifs s'inscrivent dans le cadre de travaux psychologiques portant sur la validité des modèles de représentation des émotions. Ces lexiques sont le résultat d'expériences dans lesquelles il est demandé à des sujets humains d'évaluer l'association entre des mots et des états affectifs. Des taux d'accords inter-annotateurs sont utilisés pour mesurer l'ambiguïté associée aux mots d'une part, et attester de la validité du modèle d'autre part. Le choix du vocabulaire employé est par ailleurs motivé

par le modèle de représentation considéré : une catégorisation des émotions imposant qu'il existe un vocabulaire qui discrimine chacune des émotions, seuls les termes chargés émotionnellement comme *joie*, *colère* ou *crise*, sont considérés (Strapparava & Valitutti, 2004). Pour un modèle dimensionnel, l'hypothèse étant que tous les mots de la langue dénotent un état affectif, parmi les termes considérés figurent des mots plus communs dont la charge émotionnelle est variable. En particulier, Leleu (1987) demande à 10 sujets d'évaluer 3 000 mots fréquents de la langue française dans un espace tri-dimensionnel composé des axes de *valence*, d'*activation* et d'*émotionalité* où l'émotionalité mesure le caractère subjectif d'un mot. A l'exception de cette dernière, les lexiques constitués dans le cadre d'expériences psychologiques sont rarement de grande taille, ils constituent une compilation de mots racines (*seed words*) dont l'exploitation nécessite une expansion à un vocabulaire plus large.

Les modèles de représentation issus de travaux linguistiques sont eux-mêmes le résultat d'une telle expansion. Comme décrit à la section 2.2, p. 41, les étiquettes issues d'une catégorisation des émotions sont en effet utilisées pour organiser un ensemble de mots en une hiérarchie de concepts dont les liens sont entre autres sémantiques. Chacun des niveaux rassemble alors un ensemble de termes spécialisant ceux des niveaux supérieurs. Dans les travaux linguistiques cette propagation peut être manuelle (Piolat & Bannour, 2009), ou être semi-automatique (Mathieu, 2006; Mohammad et al., 2009).

Expansion automatique à un vocabulaire plus large Une approche désormais classique pour étendre automatiquement un ensemble de mots racines à un vocabulaire plus important consiste à exploiter les liens, notamment syntaxiques, encodés dans des ressources générales comme la base *wordnet* (Miller, 1995). Cette méthode présente deux avantages : d'une part de telles bases regroupent de très nombreuses entrées et les organisent autour de liaisons de types variés ; d'autre part, les mots y sont encodés sous une forme canonique qui permet de factoriser leurs différentes formes de manière similaire aux stems et aux lemmes présentés à la section 1.2, p. 19. La base *wordnet* par exemple, définit la notion de *synsets* pour regrouper un ensemble de mots associés à une même sémantique sous une même entrée.

La base *senti-wordnet* (Esuli & Fabrizio, 2006) est un exemple de lexique affectif constitué au travers d'une propagation de mots racines dans le réseau *wordnet* : il associe à 3 000 *synsets* un marqueur de polarité (positif ou négatif) ainsi qu'un indice d'intensité sur une échelle discrète. Comme c'est ici le cas, les lexiques ainsi constitués exploitent souvent une modélisation bi-polaire des états affectifs. La base *wordnet-affect* (Strapparava & Valitutti, 2004) est une exception qui repose sur la propagation des étiquettes issues d'une catégorisation des émotions à un vocabulaire plus large.

Enfin, une nouvelle voie s'inscrit dans le domaine du *web sémantique* : ce récent domaine repose sur le *crowd sourcing* pour organiser et structurer des connaissances générales sur le monde. A partir des résultats d'un projet fondateur pour ce domaine, l'*Open Mind Common Sense* (Singh, 2002), Singh (2004) propose par exemple d'organiser les connaissances recueillies en un graphe sémantique dont les noeuds sont des concepts et les arcs (orientés) sont des assertions de sens commun (par exemple x *possède* y ou bien x *est nécessaire à* y). Cambria et al. (2010) proposent d'exploiter ce graphe pour constituer un lexique affectif organisé autour de concepts plus fins. Ils exploitent notamment la base *wordnet-affect* comme liste de mots racines.

2.3.2.2 Lexiques spécialisés sur le corpus d'étude

Un problème inhérent aux lexiques génériques relève des nombreuses ambiguïtés associées à leurs entrées : dans un cadre d'utilisation pratique, de nombreux choix sont à faire sur le sens porté par les mots et généralement peu d'éléments sont disponibles pour les mener à bien. D'autres méthodes pour la constitution de lexiques consistent alors à propager une liste de mots racines à un vocabulaire spécifique au corpus étudié. Comme présenté à la section 1.1.2.2, p. 12, l'intérêt d'une spécialisation des ressources est d'établir une bijection entre les concepts employés et le vocabulaire utilisé.

Ces méthodes s'inscrivent dans le paradigme de l'*apprentissage transductif* pour lequel l'enjeu porte moins sur les propriétés de généralisation des modèles obtenus que sur la prise en compte de toute l'information disponible dans un corpus d'étude. Pour ce faire, ces méthodes extraient de nouvelles connaissances à partir des documents non étiquetées du corpus, à cet effet les algorithmes de construction de descripteurs présentés à la section 1.2.2, p. 24 sont largement utilisées. Il faut cependant noter que les descripteurs obtenus reposent sur un partitionnement naturel des mots dans les corpus. Pour les corpus étudiés dans le cadre des émotions, une catégorisation du vocabulaire selon des états affectifs est rarement observée de manière naturelle (Mei et al., 2007). Une approche classique pour traiter des émotions consiste alors à introduire une forme de supervision nouvelle qui repose sur un étiquetage de mots racines tel que présenté précédemment. Ces derniers peuvent être extraits d'un corpus étiqueté selon les concepts étudiés (Mei et al., 2007; Petersen & Butkus, 2008) ou, comme décrit précédemment, issus des travaux psychologiques ou linguistiques (Strapparava & Mihalcea, 2008). Les concepts initialement définis sont alors automatiquement spécialisés et augmentés sur le corpus d'étude.

D'autres méthodes ont été proposées pour traiter spécifiquement des émotions, Turney (2002) et Read (2004) considèrent par exemple une mesure d'association entre les mots d'un corpus et un vocabulaire plus général au travers de leur *information mutuelle par points*.

2.4 Méthodes pour l'analyse d'états émotionnels dans les textes

De nombreuses approches ont été étudiées pour analyser automatiquement la charge émotionnelle portée par des documents. Elles ne considèrent pas toutes une même tâche : les méthodes proposées varient selon la nature des concepts étudiés mais aussi selon le type de descripteurs employés. Dans cette section nous organisons les travaux réalisés dans le domaine de l'*affective computing* pour les textes selon la représentation qui est faite des documents. Nous distinguons trois approches selon que les descripteurs employés sont au plus proche des corpus étudiés, reposent sur un enrichissement sémantique de ces derniers ou intègrent ces deux formes de représentation.

2.4.1 Apprentissage à partir de descripteurs bas niveau

En apprentissage sur les textes, les concepts étudiés sont traditionnellement des thématiques pour lesquelles il est supposé qu'un vocabulaire discrimine naturellement les documents d'un corpus. Une tâche de classification pour ces thématiques consiste alors en l'identification du vocabulaire le plus pertinent pour décrire les regroupements considérés dans un corpus étiqueté. Comme rappelé à la section 1.2.1, p. 20, pour ce faire les méthodes mises en œuvre réalisent un apprentissage linéaire des concepts étudiés.

Pour identifier la charge émotionnelle des documents, une approche consiste ainsi à exploiter une catégorisation des émotions afin d'identifier un vocabulaire discriminant pour les étiquettes émotionnelles considérées (par exemple *joie*, *tristesse* ou *colère*). Cependant la plupart des méthodes qui réalisent un apprentissage à partir de descripteurs bas niveau s'inscrivent dans le domaine de l'*opinion mining* : elles consistent à adopter une représentation bi-polaire des émotions et à identifier un vocabulaire soit positif soit négatif dans les documents étiquetés (Pang & Lee, 2008). Bien que dans certains cas il soit de plus tenu compte d'une échelle d'intensité, les modèles de représentation plus fins proposés par les psychologues et les linguistes ne sont que rarement voire jamais exploités.

L'absence de corpus étiqueté selon une représentation des émotions plus fine ainsi que les faibles taux d'accords inter-annotateurs observés sur des concepts subjectifs tels que les émotions sont à l'origine du manque d'intérêt porté par ces méthodes à cette tâche (Ovesdotter-Alm et al., 2005; Mishne, 2005).

Pour une discrimination des émotions selon un modèle de représentation bi-classes positif/négatif, Pang et al. (2002) proposent une étude détaillée de trois méthodes d'apprentissage : *naïve Bayes*, machines à vecteurs de support linéaires, et classification par maximum d'entropie. Les auteurs étudient de plus différents schémas de pondération pour différents types de descripteurs bas niveau : ils considèrent notamment des enrichissements syntaxiques ainsi que des combinaisons de tels descripteurs. Les auteurs obtiennent de meilleurs résultats pour les machines à vecteurs de support linéaires, dans un espace de représentation composé des unigrammes codés selon le schéma binaire. Les performances n'égalant pas celles obtenues pour des concepts cibles plus classiques, les auteurs concluent à la difficulté de la tâche et insistent sur la difficulté d'identifier un vocabulaire discriminant pour des concepts affectifs.

Plus récemment, Ng et al. (2006) mettent en œuvre une fusion anticipée de p-grammes pour différents ordres et montrent l'intérêt de tenir compte des spécificités liées aux descripteurs employés. Les auteurs obtiennent en effet une nette amélioration des résultats lorsque les espaces de représentation d'origine sont maintenus à un nombre égal de dimensions. Enfin Mejova et Srinivasan (2011) proposent une étude comparative de plus grande ampleur sur des corpus de différentes tailles : ils obtiennent des résultats variables selon les corpus et montrent la difficulté d'établir des règles générales pour l'identification d'un vocabulaire discriminant (voir section 2.3.1), p. 45. Rafrafi et al. (2012) soulèvent un problème plus général lié à la rareté des mots chargés émotionnellement dans les corpus d'étude : pour des concepts affectifs, ce sont souvent les mots les plus fréquents d'un corpus qui dominent le vocabulaire identifié, et ce, indépendamment de leur charge émotionnelle. Pour y remédier, les auteurs proposent de tenir explicitement compte de la fréquence des mots lors de l'identification du vocabulaire discriminant.

2.4.2 Caractérisation dans un espace sémantique

De nombreux travaux proposent de caractériser plus finement les émotions et mettent en œuvre une modélisation de ces dernières. Ces méthodes constituent une approche différente de celle présentée à la section précédente : un document est décrit par un ensemble d'états affectifs extraits au niveau des phrases, une agrégation de ces concepts permet de caractériser finement sa charge émotionnelle.

Pour ce faire ces méthodes exploitent un enrichissement sémantique des documents : elles extraient de ces derniers des descripteurs sémantiques à partir des mots ou des groupes de mots définis dans les ressources utilisées et leur associent l'un des états affectifs du modèle employé pour décrire les émotions. Pour un modèle catégoriel, un descripteur est

par exemple une émotion primaire ou une émotion complexe, pour un modèle dimensionnel c'est un vecteur dans un espace multi-dimensionnel (voir section 2.2, p. 41). Ainsi, une agrégation de descripteurs sémantiques autorise pour le premier, un étiquetage émotionnel selon une catégorisation fine des émotions (Salway & Graham, 2003; Bestgen et al., 2004; Piolat & Bannour, 2009), pour le second, une caractérisation fine de l'état affectif associé à un document (Cowie et al., 1999).

L'extraction des descripteurs correspond souvent à une annotation des documents (*keyword spotting*, voir section 1.1.2.2, p. 12). Des stratégies plus avancées ont été considérées notamment pour tenir compte des modificateurs d'intensité linguistiques ou de la négation. Lorsque ces dernières sont utilisées de pair avec des ressources mettant explicitement en œuvre des composantes de gradualité (Subasic & Huettner, 2000; Fitrianie & Rothkrantz, 2008), il est tenu compte, au niveau de la phrase des ambiguïtés et des imprécisions du langage. En particulier, lorsque la description faite des émotions le permet, les modificateurs linguistiques comme *très* et *peu* sont exploités pour préciser la charge émotionnelle des concepts annotés, la négation est modélisée comme un opérateur d'inversion de polarité (Boucouvalas, 2003; Neviarouskaya et al., 2007; Balahur & Montoyo, 2008).

2.4.3 Apprentissage à partir de descripteurs enrichis

Les deux approches précédentes souffrent chacune d'inconvénients qu'il convient de relever : les descripteurs bas niveau n'offrent que peu de précision et de finesse pour la représentation faite des documents ; les descripteurs sémantiques souffrent d'un manque de représentativité. Il est difficile d'établir des lexiques couvrant un vocabulaire exhaustif et il s'avère que le contexte, thématique par exemple, influence grandement le sens des mots employés.

Une troisième approche combine ces deux formes de représentation afin de tirer parti de leurs avantages respectifs. D'une part, l'utilisation de descripteurs bas niveau permet d'extraire une information proche des données étudiées, d'autre part, l'emploi d'enrichissements sémantiques permet de combler le fossé sémantique en exploitant une représentation fine des émotions. Comme rappelé à la section 1.3, p. 25, cette fusion peut opérer à différents niveaux de la chaîne d'apprentissage, une première stratégie l'effectue à l'échelle des descripteurs (Mullen & Collier, 2004; Whitelaw et al., 2005; Ng et al., 2006), une autre à celle des classifieurs (Das & Chen, 2007; Melville et al., 2009; Prabowo & Thelwall, 2009).

2.5 Conclusions

L'étude des émotions dans les textes est une tâche dont la difficulté est intimement liée à celle de décrire des concepts complexes, imprécis et ambigus.

Dans ce chapitre nous avons présenté les principaux modèles de représentation des émotions, proposés dans le cadre de travaux psychologiques ou linguistiques, qui offrent différents niveaux de gradualité pour décrire les concepts affectifs. L'approche catégorielle des émotions segmente ces dernières en un ensemble fini d'états primaires, autour desquels sont organisés des états plus complexes ou plus intenses ; les modèles dimensionnels décrivent une émotion comme un état affectif dans un espace sémantique multidimensionel, et permettent ainsi d'évaluer la charge émotionnelle d'un document sur différents axes. Parmi les plus classiques figurent notamment la valence, l'activation et le contrôle.

Pour le cas où les concepts étudiés sont de type émotionnel, nous avons également présentés les difficultés liées à la représentation des documents pour une tâche d'appren-

tissage. Dans ce cadre, de nouveaux descripteurs ont été considérés : il ressort des études réalisées dans la littérature que les verbes, les adverbes, les adjectifs ainsi que la ponctuation, traditionnellement écartés des documents pour des concepts plus classiques, apportent une information nécessaire à la discrimination de concepts affectifs. Néanmoins, cet apport semble conditionné par la nature du problème considéré et plus particulièrement par le vocabulaire utilisé dans les corpus, les concepts considérés, et le nombre de documents disponibles en phase d'apprentissage. Une autre approche repose sur un enrichissement sémantique des corpus. Nous avons présenté des ressources qui associent aux mots du vocabulaire l'un des états affectifs du modèle utilisé pour décrire les émotions. Nous avons divisé ces lexiques selon que le vocabulaire employé est générique ou spécialisé sur le corpus d'étude.

De plus, nous avons fait état de trois approches classiques pour l'analyse des émotions dans les textes : dans un premier temps nous avons présenté des méthodes qui réalisent un apprentissage de concepts affectifs dans un espace de représentation bas niveau. Les émotions sont représentées selon un axe bi-polaire positif/négatif et ces méthodes s'inscrivent dans le domaine de l'*opinion mining*. Dans un second temps nous avons détaillé une tâche d'*emotion mining* pour laquelle la charge émotionnelle des documents est caractérisée sur des états affectifs fins. Les méthodes présentées exploitent un espace de représentation non textuel dont la constitution repose pleinement sur un enrichissement sémantique des corpus. Enfin, nous avons décrit une approche consistant à intégrer au vocabulaire extrait des corpus, des enrichissements sémantiques pour réaliser un apprentissage des concepts affectifs.

Auteurs	**Nom**	**Modèle**	**Taille**	**Expansion**	**Type**	**Langue**
Leleu (1987)	-	valence/activation/contrôle	3000	-	mots/déclinaisons	Francais
Russell et Mehrabian (1977)	-	valence/activation	300	-	mots	Anglais
Wilson et al. (2005)	MPQA	subjectif/objectif/polarité	8000	-	mots	Anglais
Stone et al. (1962)	Harvard GI	positif/négatif	11788	-	mots	Anglais
LIWC inc.	LIWC$^{(*)}$	positif/négatif/étiquettes	8000	-	mots	Anglais
Mathieu (2006)	-	intensité/polarité/héritage	950	auto	mots/verbes	Francais
Mohammad et al. (2009)	Semantic Orientation Lexicon	positif/négatif	76400	auto	mots	Anglais
Piolat et Bannour (2009)	Emotaix	polarité/héritage	2014	manuel	mots	Francais
Esuli et Fabrizio (2006)	Senti-wordnet	polarité/subjectivité	2000	auto	synsets	Anglais
Nielsen (2011)	ANEW	étiquettes émotionnelles	900	auto	synsets	Anglais
Strapparava et Valitutti (2004)	Wordnet-Affect	8 étiquettes émotionnelles	900	auto	synsets	Anglais
Cambria et al. (2010)	Sentic-Net	polarité	4000	auto	synsets	Anglais

TABLE 2.1 – Exemples de ressources sémantiques générales pour l'analyse automatique de concepts affectifs. $^{(*)}$ LIWC n'est pas librement distribué.

Chapitre 3

Apprentissage de concepts affectifs à partir de descripteurs bas niveau

Nous étudions un système de discrimination de concepts affectifs issus d'une catégorisation des émotions : nous envisageons le cas où un document est associé à une parmi M étiquettes émotionnelles ou une étiquette neutre dans le cas où il est dépourvu de charge émotionnelle. Dans ce cadre, nous proposons de réaliser une discrimination de concepts affectifs représentés finement dans un espace de représentation bas niveau.

La méthode que nous proposons consiste à combiner, dans une approche de fusion anticipée, des descripteurs bas niveau définis comme des p-grammes de plusieurs ordres. Nous proposons de plus d'effectuer une extraction automatique de dictionnaires spécialisés pour chacune des émotions considérées. Nous appliquons un processus de décision de type « un contre tous » à deux niveaux, utilisant des classifieurs linéaires. L'approche proposée est appliquée à un corpus de textes réels dont chaque phrase est étiquetée selon 15 émotions.

Ce chapitre est organisé comme suit : l'approche proposée est présentée à la section 3.1, l'espace de description pour représenter les documents est détaillé à la section 3.2. Pour un problème de discrimination des émotions, les bases d'apprentissage diffèrent du cadre classique, les particularités de l'algorithme d'apprentissage mis en œuvre sont présentés à la section 3.3. Nous avons implémenté l'approche considérée dans le cadre d'une compétition, le système résultant est détaillé à la section 3.4. Nous présentons également une étude de ses performances et des marqueurs d'émotions identifiés dans les documents. Enfin, des pistes d'enrichissement sont fournies à la section 3.5, et les conclusions et les perspectives de ce chapitre sont présentées à la section 3.6.

Ces travaux ont fait l'objet d'un article de conférence et de journal : (Dzogang et al., 2012), la méthode proposée a été mis en œuvre dans le cadre de la compétition I2B2 (*track2*) (Pestian et al., 2012).

3.1 Architecture générale

L'approche que nous considérons consiste à représenter un document d'après les mélanges de p-grammes qui le composent. La méthode proposée repose sur trois caractéristiques décrites successivement.

Elle met en œuvre un processus de fusion anticipée pour l'agrégation de descripteurs bas niveau, calculés comme des p-grammes de différents ordres.

Afin d'homogénéiser la représentation d'un document dans chacun des espaces $\mathcal{X}_{p\in[1..L]}$ associés aux combinaisons de $p \in [1..L]$ mots, les combinaisons les plus rares sont éliminées

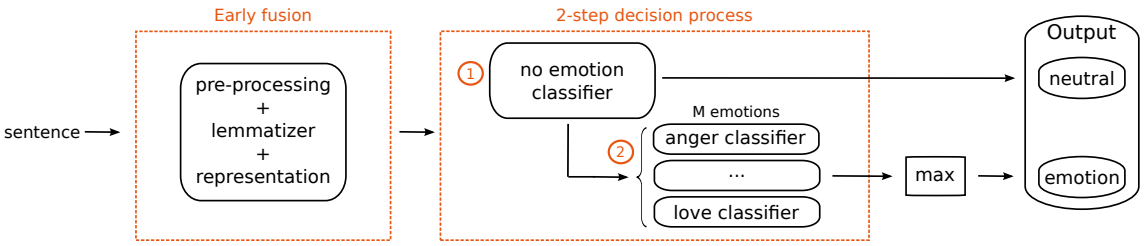


FIGURE 3.1 – Architecture globale de l'approche proposée

des espaces correspondants. Nous proposons de plus un second filtre spécifique à chacun des M concepts considérés qui repose sur la mesure d'entropie de Shannon pour identifier les descripteurs les moins discriminants pour chacun.

L'apprentissage « un contre tous » consiste à décomposer une tâche de classification multiclasses en multiples sous-tâches de classification binaires, chacune discriminant une des classes par opposition à toutes les autres. La stratégie de classification proposée, schématisée sur la figure 3.1, s'inscrit dans ce cadre, en considérant toutefois deux niveaux : un premier classifieur vise à distinguer si un document est porteur d'émotions ou non, c'est-à-dire considère une classe neutre par opposition à toutes les émotions. Dans un second temps, si le texte a été classé comme exprimant une émotion, il est soumis à une seconde chaîne de traitement, dans laquelle M systèmes de décision discriminent chacun une émotion e contre toutes les autres. Le classifieur prédisant une émotion avec la confiance maximale détermine la classe globalement prédite.

3.2 Espace de description bas niveau

Etant donné un corpus composé de n documents, nous proposons de construire un espace de représentation bas niveau à partir de L espaces de représentation $\mathcal{X}_{p\in[1..L]}$, chacun composé des grammes d'ordre p extraits du corpus. Dans un premier temps nous détaillons l'emploi de mélanges de p-grammes pour la représentation des documents, à cet effet nous motivons l'extraction d'un vocabulaire qui tient compte du contexte d'apparition des mots.

La taille des dictionnaires associés à de tels descripteurs croît en fonction de l'ordre considéré, nous présentons dans un second une méthode simple pour équilibrer le nombre d'entrées composant chacun. De plus, nous proposons d'éliminer des dictionnaires, les entrées non pertinentes pour discriminer chacune des émotions : nous détaillons une méthode de filtrage effectuant une spécialisation des dictionnaires pour chacune des M émotions considérées.

Enfin nous discutons de l'espace de représentation final pour décrire les documents : nous adoptons une stratégie de fusion anticipée (voir section 1.3.2, p. 27) et nous formons cet espace à partir de la concaténation de l'ensemble des dictionnaires obtenus. Ce dernier contient aussi bien des descripteurs simples et génériques comme les unigrammes que des descripteurs plus riches qui tiennent compte du contexte d'énonciation des mots.

3.2.1 Descripteurs considérés : p-grammes

Nous proposons de décrire les documents d'après le vocabulaire employé dans le corpus étudié. Pour ce faire, il est classique d'exploiter les mots apparaissant de manière unique

dans ce dernier, cependant dans le cas des émotions, nous faisons l'hypothèse qu'ils ne suffisent pas, à eux seuls, à discriminer des concepts affectifs intrinsèquement subtils et complexes.

Nous proposons ainsi d'exploiter des combinaisons de tels descripteurs et nous considérons en particulier l'extraction de p-grammes pour des ordres $p \geq 1$ mot(s). En effet, tandis que les unigrammes constituent des descripteurs simples et génériques qui assurent une bonne couverture des documents, à des ordres plus élevés, les mots sont associés à leurs contextes d'apparition : les descripteurs résultants sont plus à même de représenter une information riche et complexe, nécessaire à l'étude des émotions. Ainsi, comme rappelé à la section 2.3, p. 44, à des ordres élevés les p-grammes offrent un cadre de représentation naturel pour les constructions linguistiques complexes : à l'ordre $p = 2$ mots un bigramme permet par exemple de modéliser l'effet d'une négation sur les mots (e.g. *pas mauvais*), à l'ordre $p = 3$ mots des constructions plus subtiles (e.g. *vraiment pas mauvais*) sont associées à un unique descripteur, et des ordres plus grands permettent de décrire des expressions encore plus subtiles et complexes.

Néanmoins, comme décrit à la section 1.1.1.1, p. 8, la fréquence d'un p-gramme dans un corpus est directement liée à l'ordre considéré. Dans un scénario extrême les descripteurs décrivent des documents entiers du corpus et l'information résultante ne permet plus de décrire les concepts étudiés. En effet, dans l'espace de représentation correspondant il n'est alors plus possible d'identifier une frontière de décision exhibant de bonnes propriétés de généralisation. L'ordre maximal p doit donc être estimé selon le corpus étudié.

3.2.2 Spécialisation des dictionnaires selon les émotions

Etant donné un dictionnaire dont les entrées contiennent les p-grammes extraits du corpus d'étude, nous proposons une méthode pour sa spécialisation sur chacune des M émotions considérés.

Motivations Comme remarqué à la section 2.3.1, p. 45, pour une tâche de discrimination des émotions il est souvent souhaitable d'extraire toute l'information disponible sur un corpus et de conserver telles quelles les entrées des dictionnaires. C'est ensuite l'algorithme d'apprentissage mis en œuvre qui détermine la pertinence des descripteurs correspondants. Cependant, étant donné la rareté des marqueurs d'émotions dans les documents, il est également souhaitable d'écarter des dictionnaires les entrées les moins à même de décrire les émotions étudiées. Traditionnellement, des listes de mots vides (*stop words*) qui regroupent les mots les plus communs d'une langue (par exemple *le*, *un*, *il*) permettent d'éliminer ce vocabulaire des dictionnaires dans le cas de combinaisons de mots, leur utilisation est moins naturelle. Il est en effet plus délicat de constituer des listes d'expressions vides, et selon le problème considéré, l'élimination de mots vides au sein de combinaisons de mots nécessite une connaissance profonde des interactions entre les termes d'un corpus.

Principe Nous proposons plutôt de faire usage de l'information portée par les étiquettes émotionnelles et nous proposons un procédé permettant à la fois d'écarter les descripteurs vides et de filtrer des dictionnaires les entrées les moins pertinentes pour décrire les concepts cibles.

Soit une étiquette émotionnelle e et un dictionnaire $\mathcal{D}^p$ contenant les p-grammes du corpus. Le critère de filtrage proposé est basé sur une mesure de la quantité d'information apportée par un descripteur f issu de $\mathcal{D}^p$ pour la prédiction de l'émotion e. Des mesures

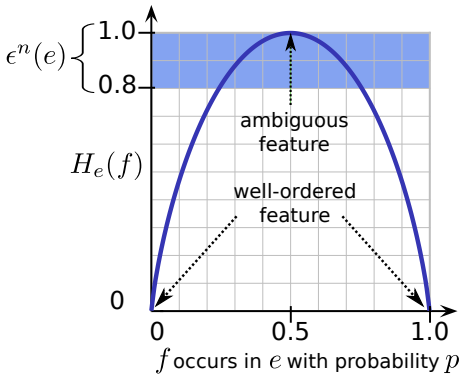


FIGURE 3.2 – Sélection des descripteurs à partir du seuillage d'un critère de qualité basé sur la mesure d'entropie de Shannon. Selon les émotions, nous avons empiriquement déterminé que le seuil ϵ varie entre 0.8 et 1.

basées sur la log-vraisemblance pondérée ou le score de χ^2 on été exploitées, nous proposons d'utiliser la mesure d'entropie de Shannon : formellement, en notant p_e la proportion de documents étiquetés e parmi les documents contenant le descripteur f, le critère proposé est défini comme :

$$H_e(f) = -\left[(1-p_e)\log_2(1-p_e) + p_e \log_2(p_e)\right] \tag{3.1}$$

où $(1-p_e)$ représente la proportion de documents non étiquetés par l'émotion e parmi ceux contenant le descripteur f. Il faut noter qu'un filtrage plus fin serait réalisé par un critère tenant simultanément compte de l'ensemble des M étiquettes. Comme nous le présentons à la section 3.3, le critère H_e est motivé par la discrimination d'une émotion contre toutes les autres.

Comme illustré sur la figure 3.2, H_e est maximal si f est uniformément distribué, c'est-à-dire autant présent dans les documents étiquetés e que les autres et minimal si tous les documents contenant f, $(p_e = 1)$ ou aucun $(p_e = 0)$, sont étiquetés e, c'est-a-dire la présence ou l'absence de f est totalement discriminante.

Méthode Nous proposons donc de filtrer un dictionnaire en conservant uniquement les descripteurs associés à une entropie inférieure au seuil $\epsilon(p, e)$, défini indépendamment pour chaque ordre p et chaque émotion e. Il peut par exemple être estimé automatiquement en fonction des performances des classifieurs basés sur les dictionnaires correspondants. Il faut néanmoins noter qu'une valeur élevée de ce seuil est préférée, afin de laisser le soin à l'algorithme d'apprentissage d'effectuer une sélection des descripteurs les plus discriminants pour les émotions considérées. Dans nos expériences ϵ est généralement plus élevé pour les unigrames (et pour les émotions rares) ce qui conforte l'intuition que le vocabulaire associé est moins spécifique pour cette représentation.

Nous notons $\mathcal{D}^p(e)$ le dictionnaire spécialisé pour l'émotion e et dont les entrées sont les p-grammes extraits du corpus :

$$\mathcal{D}^p(e) = \{f \in \mathcal{D}^p / H_e(f) < \epsilon(p, e)\}$$

l'espace de représentation associé est noté $\mathcal{X}_p(e)$. Sur le corpus d'étude M espaces de représentation sont ainsi constitués.

3.2.3 Espace de représentation final : mélange de p-grammes

Pour représenter un document, nous proposons de tenir compte à la fois de mots génériques comme les unigrammes dont les descripteurs associés présentent une bonne couverture des textes, et de descripteurs plus riches, représentés pour des ordres plus élevés. Nous mettons ici en œuvre, une méthode de fusion anticipée (voir section 1.3.2, p. 27) pour décrire un document comme un mélange de p-grammes pour des ordres $p \in [1..L]$.

Pour l'émotion e, le nombre de descripteurs associés à des grammes d'ordres différents étant très variable (voir section 1.1.1.1, p. 8), les représentations $\boldsymbol{x}^p(e) \in \mathcal{X}^p(e)$ faites d'un document sont très déséquilibrées selon l'ordre p considéré : en vue de leur mélange, un rééquilibrage est nécessaire (Ng et al., 2006). Pour ce faire, nous remarquons que pour des ordres élevés, de nombreuses combinaisons de mots apparaissent de manière accidentelle dans les corpus et ne constituent pas des descripteurs valides. Ainsi, pour homogénéiser l'ensemble des L représentations associées à un document, nous proposons d'écarter les combinaisons de mots excessivement rares dans les corpus : préalablement à la spécialisation des dictionnaires décrite précédemment, nous supposons que pour tout ordre $p \in [1..L]$, les $\mathcal{D}^p$ sont constitués de descripteurs présents dans plus de 2 documents. Cette propriété est importante, elle permet, à moindre coût, de rééquilibrer la taille des espaces de description correspondants.

Nous proposons de représenter les documents dans un espace $\mathcal{X}(e)$ construit comme la concaténation des L espaces spécialisés sur e :

$$\mathcal{X}(e) = \bigoplus_{p=1}^{L} \mathcal{X}_p(e)$$

Tout document représenté dans cet espace est décrit par un vecteur dont les composantes sont des combinaisons de $p \in [1..L]$ mots : $\boldsymbol{x}(e) = \boldsymbol{x}^1(e) \oplus \ldots \boldsymbol{x}^L(e) \in \mathcal{X}(e)$.

3.3 Apprentissage des classifieurs

Comme décrit à la section 3.1, notre approche repose sur une stratégie de classification « un contre tous » pour discriminer chacune des M étiquettes émotionnelles du problème. Nous présentons dans un premier temps les frontières de décision mise en œuvre, dans un second temps, nous remarquons que les bases d'apprentissage étiquetées selon des émotions sont souvent déséquilibrées et nous décrivons une méthode pour tenir compte de leur fréquence dans le corpus lors de l'apprentissage des frontières.

3.3.1 Frontières de décision linéaires

En notant $\boldsymbol{x}$ le vecteur de représentation final d'un document, $\boldsymbol{x}(e)$ est le vecteur d'entrée associé au sous-problème visant à discriminer l'étiquette e du reste des émotions. Pour un corpus d'étude $\mathcal{D}$ étiqueté sur $\mathcal{Y} = \{e_1, \ldots, e_M\}$ on a donc :

$$\boldsymbol{x} := \begin{cases} \boldsymbol{x}(e_1) = & \boldsymbol{x}^1(e_1) \oplus & \ldots \oplus & \boldsymbol{x}^L(e_1) \\ \ldots & & & \\ \boldsymbol{x}(e_M) = & \boldsymbol{x}^1(e_M) \oplus & \ldots \oplus & \boldsymbol{x}^L(e_M) \end{cases}$$

Dans ce cadre, pour chacun des M sous-problèmes, une frontière de décision f_e, spécifique à l'émotion e, est construite dans l'espace de représentation final correspondant $\mathcal{X}(e)$. Nous

proposons de considérer des frontières de décision linéaires : d'une part leur efficacité pour la classification de textes a été observée dans de multiples applications ; d'autre part, les frontières produites rendent explicite le rôle des descripteurs. Pour chacun des $M + 1$ problèmes bi-classes, la fonction de prédiction s'écrit sous la forme :

$$f_e(\boldsymbol{x}(e)) = \text{sign}(\boldsymbol{x}(e)^\intercal \boldsymbol{\alpha} + b) \tag{3.2}$$

où le vecteur $\boldsymbol{\alpha}$ et le réel b sont les paramètres optimisés lors de la phase d'apprentissage. Pour ce faire, nous mettons en œuvre un algorithme d'apprentissage à vaste marge (Fan et al., 2008) qui inclut dans la fonction de coût l'amplitude du vecteur de pondération $\|\boldsymbol{\alpha}\|_2$ (Cortes & Vapnik, 95) (voir section 1.2.1, p. 20).

Il faut noter qu'un sous-problème supplémentaire consiste à discriminer les textes neutres de ceux chargés émotionnellement. Les documents *neutre* (par exemple non étiquetés dans le corpus) sont ainsi écartés des bases pour l'apprentissage des M frontières de décision qui discriminent entre les émotions.

3.3.2 Déséquilibre des classes

La distribution des concepts cibles est souvent très déséquilibrée dans les bases étiquetées sur des concepts affectifs. Le tableau 3.1 l'illustre pour la base de données considérée dans les expérimentations décrites dans la section suivante. Ainsi pour la stratégie de classification que nous considérons, les erreurs d'étiquetage pour les classes les plus rares sont par défaut associées à un coût très déséquilibré par rapport au reste des autres classes.

Pour rééquilibrer ces coûts, une stratégie consiste à échantillonner la base d'apprentissage en conservant un nombre égal de documents pour représenter la classe à prédire et le reste des autres. L'apprentissage d'une frontière de décision est alors un processus répété sur plusieurs échantillons indépendants. Pour la version souple d'un algorithme à vaste marge, la méthode que nous adoptons consiste à définir une version asymétrique C_{asym} du coût de classification C, pondérée de manière inversement proportionnelle à la fréquence des classes dans le corpus.

3.4 Mise en œuvre expérimentale

Cette section décrit la mise en œuvre de l'approche détaillée précédemment sur un corpus de données réelles. Nous présentons d'abord les données étudiées ainsi que l'extraction faite des descripteurs. Nous décrivons ensuite le protocole expérimental : d'une part les classifieurs sont évalués de manière indépendante sur chacune des émotions, d'autre part les performances du système final qui agrège les décisions prises individuellement sont mesurées. Enfin, nous présentons et discutons les résultats obtenus. Nous examinons en particulier les descripteurs retenus comme les plus pertinents pour les sous-problèmes correspondants.

3.4.1 Description du corpus

Notre étude et le système que nous proposons se placent dans le cadre de la compétition I2B2 (*track 2*) (Pestian et al., 2012) dont l'objectif est la discrimination automatique d'étiquettes émotionnelles dans des textes non structurés. Le corpus fourni aux participants est constitué de 600 textes correspondant à des notes de suicide. Ces notes sont étiquetées au niveau de la phrase selon un modèle de représentation des émotions catégoriel, décrivant $M = 15$ émotions dont la liste complète est fournie dans le tableau 3.1. Il faut noter que

Emotion	Fréquence
No emotion	2 460
Instruction	800
Hopelessness	455
Love	296
Information	295
Guilt	208
Blame	107
Thankfulness	94
Anger	69
Sorrow	51
Hopefulness	47
Happiness/Peacefulness	25
Fear	25
Pride	15
Abuse	9
Forgiveness	6

TABLE 3.1 – Distribution des concepts cibles dans le corpus I2B2 (*track2*). Les étiquettes sont ordonnées de la plus fréquente à la plus rare.

nous utilisons ici le terme d'émotions de façon abusive : certaines catégories n'ont pas de valeur subjective. Le corpus d'apprentissage est composé de l'ensemble des phrases étiquetées et non étiquetées (l'absence d'émotions est représentée par l'étiquette *neutre*), vues ici comme de très courts documents. Le tableau 3.1 donne également la distribution des étiquettes correspondantes dans le corpus d'apprentissage.

De plus, pour ce corpus les phrases peuvent être associées à plusieurs émotions. Néanmoins les phrases multi-étiquetées ne représentent que 7% du corpus. Pour ces dernières un maximum de 5 émotions ont été utilisées conjointement.

La constitution d'un tel corpus constitue un effort important de la part des organisateurs. A notre connaissance il s'agit du premier corpus destiné à l'apprentissage automatique d'émotions représentées finement pour les textes, au-delà de classes binaires distinguant uniquement les émotions positives et négatives. La compétition *SemEval* 2007 (Strapparava & Mihalcea, 2007) consiste elle aussi en la discrimination d'étiquettes émotionnelles dans les textes, mais la taille du corpus ne permet pas de mettre en œuvre un apprentissage et seules les méthodes purement linguistiques donnent des résultats acceptables.

3.4.2 Extraction des descripteurs

Etant donné la petite taille des documents étudiés, nous adoptons le schéma binaire pour construire les vecteurs de représentation. Ceux-ci indiquent donc de la présence ou de l'absence des descripteurs dans les documents, sans tenir compte de leurs fréquences d'apparition. Dans la suite nous détaillons la nature de ces descripteurs : nous présentons d'une part les mots considérés, nous discutons d'autre part des ordres retenus pour les p-grammes.

Mots considérés Nous employons l'analyseur syntaxique *TreeTagger* (Schmid, 1994) et nous réduisons l'ensemble des mots extraits des documents à leur lemme (voir sec-

tion 1.1.2.1, p. 11). Nous n'effectuons pas d'autre type de filtrage sur les entrées des dictionnaires, comme observé dans bon nombre de travaux en *opinion mining* et en *emotion mining* la ponctuation, les adverbes, et les adjectifs semblent jouer un rôle important.

Extraction des p-grammes Au travers des différentes expériences que nous avons réalisées sur le corpus d'apprentissage nous avons observé une baisse de représentativité pour les descripteurs liés à des ordres $p > 3$ mots. Pour ces derniers, il tend en effet à être de moins en moins évident de calculer une frontière de décision pertinente. Ainsi, nous fixons $L = 3$ mots dans la suite et nous étudions conjointement trois espaces de représentation, respectivement composés d'unigrames $\mathcal{X}_1$, de bigrammes $\mathcal{X}_2$ et de trigrammes $\mathcal{X}_3$. Comme indiqué dans la section 3.2.3, ces espaces sont composés des seuls p-grammes apparaissant dans plus de 2 documents du corpus d'apprentissage.

Spécialisation des dictionnaires Nous appliquons la procédure proposée dans la section 3.2.2 pour spécialiser ces espaces sur chacune des étiquettes considérées. Nous estimons empiriquement les seuils $\epsilon(p, e)$ en observant leurs effets sur les performances moyennes des classifieurs associés à chacune des sous-problèmes bi-classes. Comme illustré sur la figure 3.2, p. 58, les seuils résultants varient entre 0.8 et 1 pour tout e et pour tout p. Nous avons de plus constaté de meilleures performances sur les unigrammes et sur les émotions rares pour un seuillage presque inexistant, c'est-à-dire pour des valeurs de $\epsilon(p, e)$ très proches de 1.

3.4.3 Protocole expérimental

Nous décrivons le protocole expérimental pour évaluer la performance du système proposé : dans un premier temps, les performances des classifieurs spécialisés sur chacune des émotions sont mesurées, dans un second temps, le système final est évalué.

Apprentissage individuel des étiquettes L'apprentissage des $M + 1$ frontières de décision (dont l'étiquette *neutre*) nécessite chacun l'ajustement d'un coût de classification C qui contrôle la tolérance aux erreurs de classification en phase d'apprentissage. Nous définissons une grille de recherche composée des puissances de 2 successives entre 0 et 10 ; pour chacune de ces valeurs nous réalisons une validation croisée sur le corpus d'apprentissage. Pour ce faire, le corpus est divisé en 10 partitions (3 partitions pour les étiquettes rares) et la valeur de C retenue est celle dont le score $F1$ est en moyenne le plus élevé.

De plus, il n'est pas tenu compte de l'étiquetage multiple : chaque document est associé à une étiquette unique. Le corpus d'étude ne contient que 7% d'étiquettes multiples et nous estimons que le biais introduit par notre approche est acceptable. Bien que d'autres approches tiennent compte de la nature multi-étiquetée de la base, l'apprentissage en présence d'exemples multi-étiquetés est une tâche difficile, nous nous concentrons ici sur le problème de discrimination de concepts affectifs.

Pour chacune des émotions et chacun des ordres considérés, nous reportons, de manière classique, les valeurs moyennes de scores $F1$, rappel et précision.

Evaluation du système final Le système final, qui est composé des $M + 1$ meilleurs classifieurs, est évalué sur un corpus d'évaluation indépendant du corpus d'apprentissage, fourni par les organisateurs de la compétition. La taille du corpus d'évaluation représente

environ 1/3 de celle du corpus d'apprentissage et la distribution des étiquettes y est similaire.

Contrairement à l'évaluation individuelle des classifieurs, l'évaluation du système final tient compte de la nature multi-étiquetée des documents ainsi que de la nature multi-classes du problème. Les mesures d'évaluation employées sont ainsi les micro-moyennes des mesures de $F1$, de rappel et de précision (Yang & Liu, 1999). Nous remarquons de plus que dans le protocole défini par les organisateurs, l'étiquette *neutre* ne produit pas de *vrais positifs*.

3.4.4 Résultats et discussions

Nous présentons dans un premier temps les performances individuelles des $M + 1$ classifieurs associés à chacune des étiquettes. Nous discutons d'abord des résultats obtenus pour chacun des ordres considérés, que nous comparons ensuite aux performances réalisées dans le cas de combinaisons. Nous discutons alors des descripteurs les plus discriminants pour les émotions.

Dans un second temps, nous présentons les résultats obtenus par le système final sur le corpus indépendant du corpus d'apprentissage.

3.4.4.1 Evaluations individuelles selon les émotions

Dans cette section sont décrits les résultats individuels obtenus sur chacune des étiquettes : il s'agit de performances moyennes réalisées par validation croisée sur le corpus d'apprentissage.

Performances selon les p-grammes Les tableaux 3.2, 3.3, et 3.4 présentent les scores moyens de $F1$, de rappel et de précision associés aux meilleures performances individuelles : c'est-à-dire aux valeurs de C identifiées par validation croisée indépendamment pour chaque émotion et chaque ordre. Ces tableaux décrivent respectivement les résultats obtenus d'après les unigrammes, les bigrammes, et les trigrammes extraits du corpus. Il est important de noter que pour le système final qui implémente un processus de décision en deux étapes, les performances sur les étiquettes émotionnelles sont bornées par les performances du classifieur individuel sur l'étiquette *neutre*.

Nous observons que les classifieurs basés sur des grammes d'ordres inférieurs obtiennent globalement de meilleures performances que les classifieurs basés sur des grammes d'ordres supérieurs. Cela est conforme à l'intuition que les grammes d'ordres supérieurs sont plus spécifiques que les grammes d'ordres inférieurs et ainsi que les descripteurs qui reposent uniquement sur ces derniers ne couvrent pas suffisamment les documents. En particulier, tandis que les bigrammes tendent à augmenter la précision des classifieurs, ils tendent aussi à baisser leur rappel : le gain en précision ne compense pas à lui seul la rareté des marqueurs d'émotions correspondants. Les classifieurs qui reposent uniquement sur ces derniers ne peuvent alors correctement discriminer entre les étiquettes émotionnelles. Ce phénomène est particulièrement remarquable pour les trois émotions les plus rares : *pride* (15 documents), *abuse* (9 documents) et *forgiveness* (6 documents). Dû à l'extrême rareté de ces étiquettes dans le corpus et en dépit du schéma de classification asymétrique que nous avons utilisé, les bigrammes et les trigrammes ne suffisent pas à eux seuls à induire une frontière de décision acceptable : dans les tableaux, les *N/A* indiquent que l'apprentissage ne peut s'effectuer correctement, dans ce cas le classifieur produit des décisions par vote majoritaire.

Etiquette	F1	Précision	Rappel
Neutre	0.68 ± 0.02	0.71 ± 0.02	0.66 ± 0.03
Instruction	0.85 ± 0.02	0.86 ± 0.03	0.84 ± 0.03
Hopelessness	0.69 ± 0.04	0.63 ± 0.06	0.76 ± 0.05
Love	0.76 ± 0.04	0.73 ± 0.05	0.8 ± 0.07
Information	0.54 ± 0.04	0.45 ± 0.04	0.68 ± 0.07
Guilt	0.52 ± 0.09	0.42 ± 0.08	0.66 ± 0.1
Blame	0.23 ± 0.09	0.19 ± 0.07	0.31 ± 0.15
Thankfulness	0.98 ± 0	0.99 ± 0.01	0.98 ± 0.01
Anger	0.17 ± 0.06	0.12 ± 0.04	0.29 ± 0.11
Sorrow	0.17 ± 0	0.14 ± 0.01	0.22 ± 0.03
Hopefulness	0.24 ± 0.11	0.18 ± 0.08	0.38 ± 0.16
Happiness/Peacefulness	0.19 ± 0.13	0.19 ± 0.11	0.2 ± 0.15
Fear	0.19 ± 0.08	0.19 ± 0.04	0.19 ± 0.12
Pride	0.11 ± 0.04	0.06 ± 0.02	0.4 ± 0.2
Abuse	0.02 ± 0	0.01 ± 0	0.44 ± 0.19
Forgiveness	0.26 ± 0.1	0.16 ± 0.06	0.83 ± 0.29

TABLE 3.2 – Unigrammes : scores moyens de F1, de précision, et de rappel ainsi qu'écarts-types associés (ordonnés par la fréquence décroissante des étiquettes dans le corpus).

Néanmoins, nous observons que dans certains cas les grammes d'ordres supérieurs fournissent tout de même une meilleure représentation : pour décrire l'émotion *sorrow*, les trigrammes (0.98 ± 0.01) conduisent par exemple à une représentation des documents bien plus pertinente que les unigrammes (0.17 ± 0) ou que les bigrammes (0.05 ± 0.02). Dans une moins grande mesure, nous observons le même effet pour l'émotion *hopelessness*.

Malgré la baisse évidente de performance associée aux étiquettes peu fréquentes du corpus, nous constatons que certaines émotions se distinguent naturellement des autres : aussi, bien que les émotions *love* (296 documents) et *thankfulness* (94 documents) ne soient pas les plus fréquentes du corpus, les classifieurs associés obtiennent de très bonnes performances (le score $F1$ associé est toujours supérieur à 0.7, dans l'ensemble il est à 0.9 en moyenne) quand ils reposent sur des unigrammes ou sur des bigrammes. Cela suggère que pour certaines émotions il existerait un vocabulaire spécifique qu'il est plus facile d'identifier.

Performances pour des mélanges de p-grammes Nos observations semblent concorder sur le fait que pour ce corpus, les bigrammes capturent des constructions plus riches au détriment de la généricité des descripteurs correspondants (précision plus élevée mais rappel plus faible). Au contraire, les unigrammes semblent capturer un vocabulaire simple et générique (précision plus faible mais rappel plus élevé). Il est alors naturel de supposer qu'un classifieur qui repose sur la fusion des deux représentations puisse tirer parti d'un meilleur compromis entre sa précision et son rappel. Tandis que les trigrammes semblent pertinents pour les émotions caractérisées par des constructions plus complexes, au travers d'expériences supplémentaires que nous avons réalisées, nous n'avons pas observé de gain significatif lorsque ces derniers sont mélangés aux deux autres représentations. Dans la suite nous ne considérons donc que le cas de la fusion d'unigrammes et de bigrammes. Le

Etiquette	F1	Précision	Rappel
Neutre	0.72 ± 0.03	0.84 ± 0.03	0.63 ± 0.04
Instruction	0.82 ± 0.01	0.8 ± 0.02	0.84 ± 0.03
Hopelessness	0.64 ± 0.05	0.66 ± 0.04	0.62 ± 0.08
Love	0.74 ± 0.07	0.76 ± 0.08	0.72 ± 0.08
Information	0.47 ± 0.1	0.43 ± 0.09	0.53 ± 0.14
Guilt	0.5 ± 0.08	0.5 ± 0.08	0.5 ± 0.09
Blame	0.28 ± 0.1	0.27 ± 0.08	0.32 ± 0.14
Thankfulness	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.01
Anger	0.14 ± 0.01	0.11 ± 0.01	0.2 ± 0.02
Sorrow	0.05 ± 0.02	0.03 ± 0.01	0.16 ± 0.09
Hopefulness	0.2 ± 0.1	0.2 ± 0.06	0.21 ± 0.13
Happiness/Peacefulness	0.15 ± 0.04	0.26 ± 0.21	0.12 ± 0.01
Fear	0.13 ± 0.06	0.11 ± 0.04	0.16 ± 0.08
Pride	N/A	0 ± 0	0 ± 0
Abuse	N/A	0 ± 0	0 ± 0
Forgiveness	N/A	0 ± 0	0 ± 0

TABLE 3.3 – Bigrammes : scores moyens de F1, de précision, et de rappel ainsi qu'écarts-types associés (ordonnés par la fréquence décroissante des étiquettes dans le corpus).

vecteur de représentation $\boldsymbol{x}$ s'exprime ainsi comme :

$$\boldsymbol{x} := \begin{cases} \boldsymbol{x}_{\text{neutre}} = \boldsymbol{x}^1_{\text{neutre}} \oplus \quad \boldsymbol{x}^2_{\text{neutre}} \\ \ldots \\ \boldsymbol{x}_{\text{colère}} = \boldsymbol{x}^1_{\text{colère}} \oplus \quad \boldsymbol{x}^2_{\text{colère}} \end{cases}$$

Comme pour les résultats obtenus individuellement sur chacune des représentations, le tableau 3.5 présente les scores moyens de $F1$, de précision et de rappel associés aux meilleures performances des classifieurs reposant sur une fusion d'unigrammes et de bigrammes. Encore une fois, pour le système final les résultats présentés sont bornés par les performances obtenues sur l'étiquette *neutre*.

On observe que la combinaison augmente légèrement les performances par rapport aux p-grammes considérés isolément, ou obtient des résultats similaires. Ainsi pour *love*, la fusion améliore à la fois la précision et le rappel et donc le score $F1$. Pour *instruction*, les bigrammes obtiennent individuellement une précision moins bonne que les unigrammes, et la fusion n'améliore pas les performances obtenus par les unigrammes seuls.

3.4.4.2 Analyse des dictionnaires pour des mélanges de p-grammes

Nous considérons ici l'effet de l'étape de spécialisation des dictionnaires, à la fois en terme de taille et de contenu, lorsque les entrées décrivent des mélanges de p-grammes.

Taille des dictionnaires On constate une grande disparité des tailles de dictionnaires après spécialisation : le plus petit, associé à la classe *neutre*, contient 1 904 entrées, le plus grand associé à la classe *abuse*, en contient 3 511. Comme présenté à la section 3.4.2, ceci est une conséquence des valeurs de $\epsilon(p, e)$ estimées d'après les performances des classifieurs

Etiquette	F1	Précision	Rappel
Neutre	0.6 ± 0.02	0.85 ± 0.02	0.47 ± 0.02
Instruction	0.53 ± 0.07	0.61 ± 0.09	0.47 ± 0.08
Hopelessness	0.8 ± 0.02	0.73 ± 0.03	0.88 ± 0.02
Love	0.22 ± 0.06	0.34 ± 0.15	0.17 ± 0.03
Information	0.53 ± 0.04	0.63 ± 0.09	0.47 ± 0.04
Guilt	0.37 ± 0.06	0.4 ± 0.04	0.35 ± 0.08
Blame	0.1 ± 0.01	0.05 ± 0	0.77 ± 0.02
Thankfulness	0.03 ± 0.01	0.02 ± 0.01	0.51 ± 0.15
Anger	0.4 ± 0.08	0.51 ± 0.07	0.33 ± 0.08
Sorrow	0.98 ± 0.01	0.97 ± 0	0.99 ± 0.01
Hopefulness	N/A	0 ± 0	0 ± 0
Happiness/Peacefulness	0.04 ± 0.01	0.02 ± 0	0.49 ± 0.08
Fear	N/A	0 ± 0	0 ± 0
Pride	N/A	0 ± 0	0 ± 0
Abuse	N/A	0 ± 0	0 ± 0
Forgiveness	N/A	0 ± 0	0 ± 0

TABLE 3.4 – Trigrammes : scores moyens de F1, de précision, et de rappel ainsi qu'écarts-types associés (ordonnés par la fréquence décroissante des étiquettes dans le corpus).

Etiquette	F1	Précision	Rappel
Neutre	0.73 ± 0.04	0.8 ± 0.03	0.68 ± 0.05
Instruction	0.85 ± 0.02	0.85 ± 0.03	0.86 ± 0.03
Hopelessness	0.68 ± 0.04	0.68 ± 0.04	0.69 ± 0.06
Love	0.78 ± 0.06	0.78 ± 0.07	0.78 ± 0.08
Information	0.54 ± 0.08	0.52 ± 0.06	0.58 ± 0.12
Guilt	0.53 ± 0.07	0.51 ± 0.08	0.55 ± 0.08
Blame	0.32 ± 0.09	0.34 ± 0.11	0.31 ± 0.08
Thankfulness	0.99 ± 0	0.98 ± 0.01	0.99 ± 0.01
Anger	0.2 ± 0.1	0.18 ± 0.08	0.23 ± 0.14
Sorrow	0.16 ± 0.05	0.1 ± 0.03	0.39 ± 0.12
Hopefulness	0.23 ± 0.07	0.16 ± 0.05	0.38 ± 0.1
Happiness/Peacefulness	0.16 ± 0.12	0.12 ± 0.07	0.29 ± 0.29
Fear	0.21 ± 0.05	0.23 ± 0.06	0.2 ± 0.07
Pride	0.08 ± 0.02	0.05 ± 0.01	0.27 ± 0.12
Abuse	0.02 ± 0	0.01 ± 0	0.44 ± 0.19
Forgiveness	0.24 ± 0.1	0.14 ± 0.07	0.83 ± 0.29

TABLE 3.5 – Fusion des unigrammes et des bigrammes : scores moyens de F1, de précision et de rappel ainsi qu'écarts-types associés (ordonnés par la fréquence décroissante des étiquettes dans le corpus).

appris sur les représentations induites : nous avons constaté que pour les unigrammes ainsi que pour les émotions rares, des valeurs de seuil très proche de 1 donnent de meilleurs résultats.

Ces observations sont compatibles avec l'intuition que lorsqu'il n'est pas tenu compte

Etiquettes	Descripteurs
Thankfulness	thank appreciate than nice effort kindness under be swell than you you dear appreciate it too . have be for your
Instruction	cremate call please sell funeral teach notify to be forget me be good to have bury me dispose of care of
Love	love wonderful bless watch beloved most loving you . do . be wonderful love you god bless your john me on
Hopelessness	cancer am suffer die struggle everybody tired without you go on dear jane can not . my be . of all
Information	bldg insurance key paper owe ticket in of cincinnati be pay ohio . in this no . and my the key
Guilt	sorry forgive excuse fail hurt could burden have be forgive me please forgive have do understand . not to to help
Blame	sorry thank love please give wish go to be cause you of it you . you to in the to go

TABLE 3.6 – Descripteurs les plus discriminants pour les SVM (7 meilleurs unigrammes et 7 meilleurs bigrammes) pour les classifieurs dont le score $F1$ est supérieur à 0.3. Le tableau est ordonné par ordre décroissant des performances des classifieurs sur les étiquettes correspondantes.

du contexte d'apparition des mots, les descripteurs résultants sont moins discriminants pour des concepts affectifs. En effet, en éliminant les entrées les moins discriminantes des dictionnaires, l'information restante ne suffit plus à décrire les étiquettes émotionnelles correspondantes. Pour les émotions rares, cet effet est dû au fait que de nombreuses entrées du dictionnaire apparaissent uniquement chez les contre-exemples de l'émotion considérée, et sont donc considérées comme discriminantes.

Il faut aussi noter que la spécialisation permet dans tous les cas de réduire significativement les dictionnaires obtenus sur les unigrammes et sur les bigrammes, qui ont pour tailles respectives avant spécialisation 1 206 entrées et 2 968 entrées, autorisant une taille maximale de dictionnaire de 4 174 entrées.

Descripteurs discriminants pour les émotions Afin d'étudier les vocabulaires identifiés comme spécifiques des émotions sans considérer l'intégralité de leurs contenus, le tableau 3.6 donne les descripteurs les plus discriminants : pour les émotions associées à des classifieurs de score $F1$ supérieur à 0.3, le tableau indique les 7 unigrammes et les 7 bigrammes correspondant aux valeurs maximales du vecteur de pondération $\boldsymbol{\alpha}$ qui définit la frontière de décision (voir éq. (3.2)). La présence de ces termes influence en effet la détection de l'émotion correspondante.

On observe alors par exemple naturellement que les descripteurs *love* et *thank* sont discriminants pour les émotions correspondantes. On peut de plus noter que les unigrammes discriminants peuvent être communs à plusieurs émotions, comme *please* associé à *ins-*

Système	F1 micro	Enrichissements sémantiques/ externes au corpus
Open university	0.61	oui
Msra	0.59	oui
Mayo	0.56	oui
Nrciit	0.55	oui
Oslo	0.54	oui
Limsi	0.54	oui
Swatmrc	0.53	oui
Uman	0.53	oui
Cardiff	0.53	oui
Lt3	0.53	oui
Utd	0.52	oui
Wolverine	0.50	oui
Clips	0.50	oui
Sri et UcDavis	0.48	oui
Diego-Acu	0.48	oui
Senti6	0.47	**non**
Ebi	0.46	**non**
Duluth	0.45	**non**
Tpavacoe	0.38	**non**
Lassa	0.38	**non**

TABLE 3.7 – Résultats de la compétition I2B2 (*track2*), pour chacun des systèmes est donnée la micro-moyenne du score de $F1$ obtenu sur un corpus d'évaluation indépendant du corpus d'apprentissage (Pestian et al., 2012). Les systèmes sont ordonnés par ordre décroissant du score de $F1$.

truction et *blame*, ou *love* associé à *love* mais aussi à *blame*. Au contraire, les bigrammes discriminants apparaissent comme spécifiques à une émotion donnée, et non partagés.

Il est aussi intéressant d'observer que les unigrammes les plus influents n'apparaissent pas nécessairement dans les bigrammes les plus pertinents, et que, réciproquement, les bigrammes apparaissent comme plus discriminants que les unigrammes qui les constituent pris isolément.

Finalement, on peut constater que dans une certaine mesure la position des mots dans la phrase importe : des bigrammes de la forme « *mot* . », indiquant que le mot est le dernier de la phrase, apparaissent fréquemment comme bigrammes discriminants.

3.4.4.3 Evaluation du système final sur un corpus indépendant

Le système final repose sur une fusion des unigrammes et des bigrammes. Comme décrit à la section 3.1, les documents sont d'abord pré-traités puis étiquetés au travers d'un processus de décision en deux étapes.

Il faut noter que le système proposé n'est pas capable de fournir plusieurs étiquettes émotionnelles pour un même document : pour les documents identifiés comme porteurs d'émotions, l'étiquette émotionnelle pour laquelle la certitude est maximale l'emporte sur les autres. Sur le corpus d'évaluation, le système obtient un score de $F1$ de 0.47, une

précision de 0.49 et un rappel de 0.46 (Pestian et al., 2012) [1]. Ces scores correspondent aux moyennes micro calculées sur l'ensemble des $M = 15$ étiquettes émotionnelles.

Parmi l'ensemble des systèmes évalués par les organisateurs de la compétition, la plus faible performance en $F1$ est de 0.3, la meilleure de 0.61 et en moyenne les systèmes ont obtenu un score de 0.49 ± 0.007, l'ensemble des résultats est donné dans le tableau 3.7. Il est observé que les systèmes reposant uniquement sur l'exploitation de descripteurs bas niveau semblent bornés par une performance maximale (Pestian et al., 2012). Au contraire, les 15 premiers systèmes, qui exhibent les meilleures performances, exploitent ou reposent sur des enrichissements sémantiques qui apportent une information supplémentaire sur les concepts étudiés.

3.5 Pistes d'enrichissements

Nous relevons les limitations de l'approche proposée : notre discussion est alimentée par les résultats obtenus expérimentalement, le retour de la compétition I2B2, mais aussi par les méthodes existantes pour l'analyse automatique des émotions.

Fossé sémantique Bien que l'espace de représentation que nous proposons exploite des descripteurs de contextes de tailles différentes et tienne simultanément compte de mots génériques ainsi que de constructions plus riches, l'information décrite ne semble pas suffisamment riche pour décrire à bien les concepts affectifs. Cela est notamment remarquable d'après les résultats de la compétition (voir tableau 3.7) : il semblerait en effet que l'utilisation de descripteurs sémantiques permette de franchir un seuil de performance bornant les méthodes reposant uniquement sur l'exploitation de descripteurs bas niveau. La discrimination automatique de concepts affectifs est une tâche réputée difficile, et bien que pour certaines émotions nous montrions qu'il existe un vocabulaire naturellement discriminant (par exemple *love*), l'expression écrite des états affectifs semble emprunter des schémas complexes, nécessitant des enrichissements sémantiques supplémentaires.

Emotions rares La pertinence des frontières de décisions induites dans un espace de représentation bas niveau est soumise au nombre d'exemples observés en phase d'apprentissage. Dans le cas des émotions, on constate que l'étiquetage de corpus est difficile, même pour des annotateurs humains, et de nombreux documents, présentant un taux d'accord insuffisant, sont écartés des bases d'apprentissage. Ainsi, au sein des corpus certains concepts ne sont pas suffisamment représentés : les documents correspondants n'emploient pas un vocabulaire discriminant et la frontière de décision ne discrimine pas les classes les plus rares des classes les plus fréquentes. Ici, l'utilisation de descripteurs sémantiques permet d'introduire une forme de supervision nouvelle, implémentée à l'échelle des descripteurs.

Spécificité des mélanges selon les émotions Notre approche consiste à adopter une stratégie de fusion unique pour l'ensemble des concepts cibles : les mêmes mélanges sont considérés pour chacune des étiquettes étudiées. Or nous avons observé dans nos expériences que selon les émotions, les espaces de représentation les plus pertinents peuvent être composés de mélanges de p-grammes pour des ordres variables. Par exemple les trigrammes seuls offrent une représentation bien plus pertinente pour décrire l'émotion *sorrow*

1. Notre système correspond à la méthode *senti6* : nos résultats publiés sont soumis à une erreur commise sur l'orthographe d'une étiquette émotionnelle. Cette coquille a été corrigée et discutée avec les organisateurs de la compétition : dans ce document nous reportons les résultats correspondants.

(0.98 ± 0.01 en score de F1) que les autres formes de représentation. Il semblerait alors que pour une tâche de discrimination de concepts affectifs, il soit important de tenir compte de la nature individuelle de chacun des concepts et d'adapter ainsi la représentation faite des documents selon les concepts cibles.

3.6 Conclusions et perspectives

Dans ce chapitre nous avons présenté une approche pour la discrimination de concepts affectifs qui repose sur trois caractéristiques : la fusion anticipée de grammes d'ordres croissants, une méthode pour l'élimination des p-grammes les moins pertinents et un système de décision en deux étapes identifiant d'abord les documents neutres.

On observe que les unigrammes, pris isolément, ne suffisent pas à discriminer les concepts émotionnels naturellement complexes et subtils. En tenant compte de mélanges, composés également des bigrammes, on constate des performances légèrement supérieures en moyenne à celles réalisées sur chacune des représentations prise individuellement. De manière générale, il semble nécessaire que les espaces de représentation fusionnés décrivent chacun une information pertinente et différente pour que leur fusion produise une information de plus grande qualité. Dans ce cadre, les unigrammes semblent améliorer le rappel des classifieurs tandis que les bigrammes semblent améliorer leur précision.

On constate également que les trigrammes permettent de construire des descripteurs plus discriminants en représentant des constructions plus complexes. Néanmoins, la qualité des descripteurs obtenus dépend des concepts étudiés et pour le système que nous proposons, la prise en compte des trigrammes dans une stratégie de fusion anticipée ne donne, en moyenne, pas de meilleurs résultats.

Il semblerait donc intéressant de considérer d'autres types de fusion, tenant compte à la fois de l'hétérogénéité des représentations et de l'importance relative des espaces d'origine pour décrire les concepts cibles. Enfin pour la discrimination automatique de concepts affectifs, les ressources sémantiques permettent de représenter des concepts pour lesquels il n'existe pas un vocabulaire naturellement discriminant ; de même elles autorisent le calcul de frontières de décision acceptables pour les concepts sous-représentés dans les corpus. Au chapitre suivant, nous étudions l'exploitation de ressources sémantiques exploitant une représentation fine des émotions.

Chapitre 4

Un espace sémantique pour une caractérisation affective

Les enrichissements sémantiques permettent de pallier un certain nombre de limitations spécifiques aux espaces de représentation bas niveau : elles offrent en particulier une alternative au problème du fossé sémantique inhérent aux méthodes automatiques pour l'analyse des émotions. Dans ce cadre, les ressources associées à une représentation fine des émotions offrent de plus la possibilité d'extraire, des corpus d'étude, une information proche des mécanismes, souvent complexes, en jeu lors de leur expression écrite. Contrairement à l'approche proposée au chapitre précédent, celle que nous décrivons ici repose sur un enrichissement sémantique des documents étudiés, et s'inscrit dans les méthodes non supervisées pour l'analyse des émotions. Notre proposition ne met pas en œuvre un apprentissage mais consiste plutôt en une caractérisation fine de la charge émotionnelle associée aux documents.

Dans le cadre d'une approche dimensionnelle des émotions nous exploitons un lexique qui caractérise la charge émotionnelle de 3 000 mots de la langue française dans dans un espace formé de trois axes continus. Nous proposons de relever les termes du lexique dans les documents étudiés afin de représenter ces derniers comme des nuages de points dans un espace sémantique. L'étude des propriétés de ces ensembles fournit alors des informations fines et graduelles sur la charge émotionnelle portée par les documents. Après avoir précisé le contexte et les motivations dans la section 4.1, nous présentons dans la section 4.2 le lexique employé : nous détaillons à cet effet le choix des mots qui le structurent, l'espace dimensionnel qu'il utilise pour décrire les émotions et l'expérience psychologique à son origine. Nous y décrivons également la méthode proposée qui consiste à calculer l'intersection entre les mots d'un document et les termes du lexique. Comme présenté à la section 4.3, pour cette méthode, nous envisageons deux cadres d'étude : pour le premier nous étudions et comparons les propriétés statistiques des nuages de points associés à deux textes chargés émotionnellement, pour le second nous analysons les évolutions temporelles des ensembles de points qui décrivent à tout moment la charge émotionnelle des dialogues d'un film. Enfin, à la section 4.4 nous discutons de pistes d'enrichissement pour la méthode proposée, et les conclusions de ce chapitre sont fournies à la section 4.5.

Ces travaux on fait l'objet d'une publication dans une conférence (Dzogang et al., 2010a).

4.1 Contexte et motivations

L'intérêt d'un enrichissement sémantique réside dans l'apport fourni par l'association d'un vocabulaire à une sémantique émotionnelle. Les ressources qui exploitent à cet effet une catégorisation des émotions permettent par exemple d'extraire des documents des marqueurs d'émotions difficiles à identifier automatiquement (voir chapitre 3). De plus, les ressources qui mettent en œuvre une représentation dimensionnelle des émotions caractérisent finement les marqueurs identifiés dans un espace sémantique multi-dimensionnel. Ainsi, les états affectifs identifiés dans un corpus demeurent plus robustes que pour une catégorisation parfois subjective et ambiguë des émotions.

Nous considérons ici cette approche, et nous l'exploitons afin d'associer aux documents d'un corpus, un état affectif décrit par un ensemble de mesures dans un espace sémantique. Dans ce contexte, nous proposons de projeter le vocabulaire d'un document dans un espace sémantique multi-dimensionnel et nous interprétons le nuage de points obtenu comme un signal statique ou temporel selon le cadre d'étude.

Comme rappelé à la section 2.2, p. 41, dans cet espace il est classique de réserver des régions spécifiques à certaines émotions primaires. Nous proposons plutôt de définir la sémantique portée par les signaux selon l'application considérée : notre proposition peut ainsi être vue comme une méthode d'extraction de la charge émotionnelle d'un texte dont une interprétation catégorielle dépend du cas d'utilisation.

4.2 Méthode proposée

La méthode que nous proposons repose sur trois étapes : la première consiste à choisir un lexique qui associe aux mots des coordonnées dans un espace multi-dimensionnel pour décrire les émotions. A la section 4.2.1 nous présentons un lexique composé de 3 000 mots pour le français. Comme décrit à la section 4.2.2, les termes définis dans le lexique sont dans un second temps extraits des documents par une simple identification de mots clefs. Finalement, comme présenté à la section 4.3.1, les propriétés de l'ensemble de points résultant peuvent être analysées à plusieurs niveaux.

4.2.1 Lexique considéré : représentation dimensionnelle des émotions

De nombreux lexiques ont été proposés pour enrichir les documents d'une sémantique émotionnelle. A la section 2.3.2.1, p. 47, nous en avons présenté de différentes natures et nous les avons organisés selon le type de vocabulaire employé, qui peut être générique ou spécialisé sur un corpus d'étude. Le tableau 2.1, p. 53 décrit des ressources qui regroupent un vocabulaire générique. Comme on peut le constater, une majorité repose sur une catégorisation des émotions et en particulier sur un modèle de représentation bi-classe, positif/négatif.

Au contraire, l'approche que nous proposons s'appuie sur une représentation fine des émotions : nous proposons d'exploiter le lexique de Leleu (1987) qui regroupe 3 000 mots de la langue française et qui utilise une représentation tridimensionnelle des émotions. Dans la suite nous présentons et analysons les propriétés de ce lexique.

4.2.1.1 Description du lexique

Nous présentons ici le lexique de Leleu (1987), nous détaillons d'abord le choix du vocabulaire qu'il définit, puis nous considérons l'espace de représentation qu'il emploie

Termes	Activation	Emotionalité	Valence
amour	61	68	68
haine	58	59	20
confiance	44	53	59
mépris	35	46	16
mort	13	56	16
arme	45	45	25
fromage	11	13	45

TABLE 4.1 – Echantillon du lexique de Leleu

pour décrire les émotions. Enfin, nous détaillons l'expérience psychologique qui a mené à sa constitution.

Choix du vocabulaire Conformément à l'approche dimensionnelle pour décrire les émotions (voir section 2.2, p. 41), ce lexique constitue un vocabulaire générique et contient aussi bien des termes qui dénotent des émotions précises comme *amour* ou *tristesse* que des termes dont la charge émotionnelle est plus subtile, voire plus faible. Parmi ceux-là, certains comme *arme* ou *mort* connotent une émotion et sont très chargés émotionnellement, d'autres comme *table* ou *chaise* sont plus communs. Le tableau 4.1 fournit, à titre d'exemple, 7 mots du lexique.

Tandis que les termes dénotant un état affectif précis renvoient directement à une émotion dite primaire (voir section 2.2, p. 41) et peuvent être décrits à partir d'une catégorisation des émotions, ceux qui connotent cet état affectif de manière plus subtile nécessitent une représentation fine et graduelle des émotions. Selon l'hypothèse dimensionnelle l'ensemble des mots d'une langue connotent un état affectif (entre autres au travers des expériences personnelles ou de la culture). Le lexique regroupe ainsi 3 000 mots jugés les plus courants de la langue française. Parmi ces mots figurent des formes déclinées telles que les féminins, les pluriels ou certaines formes conjuguées comme les participes passés.

Représentation des émotions Le lexique exploite une représentation tridimensionnelle des émotions, composée des axes de *valence*, d'*activation* et d'*émotionalité*. Comme rappelé à la section 2.2, p. 41, les deux premiers sont utilisés de manière classique dans les travaux psychologiques : la valence décrit la polarité associée à un état affectif, et l'activation en donne une mesure d'intensité. L'émotionalité peut ici être interprétée comme une évaluation des termes sur une échelle subjectif/objectif.

De plus, chacun de ces axes décrit un intervalle de valeurs entre 10 et 70, un état dont la valence est inférieure à 40 (qui constitue la médiane de l'échelle de mesure) est de polarité négative, et il est de polarité positive pour une valence supérieure. Par exemple, dans le lexique, le mot *confiance* est associé à un état positif, et le mot *haine* à un état très intense.

L'axe d'émotionalité est spécifique au modèle utilisé par ce lexique, il décrit la charge émotionnelle associée aux états : comme décrit dans le tableau 4.1, un mot très chargé émotionnellement comme *amour* est associé à une émotionalité élevée, tandis qu'un mot qui l'est moins comme *fromage* est décrit par une émotionalité faible.

Constitution du lexique Comme présenté à la section 2.3.2.1, p. 47, ce lexique est le résultat d'une expérience psychologique, au cours de laquelle il a été demandé à des sujet

humains d'évaluer différents termes sur les axes de valence, d'activation et d'émotionalité. Pour chacun des termes du lexique ont ainsi été obtenus différents états dans l'espace sémantique, ces résultats ont alors été agrégés sur chacun des axes, puis ramenés à des valeurs entières entre 10 et 70 (Leleu, 1987).

4.2.1.2 Analyse des propriétés du lexique

Pour évaluer la pertinence du lexique nous analysons les propriétés géométriques et statistiques de l'ensemble de points que constitue son vocabulaire. En particulier, nous examinons sa dispersion dans l'espace en étudiant la corrélation des axes, mesurée sur l'ensemble des états affectifs décrits dans le lexique. Le choix du vocabulaire ayant pour motivation sa généricité, il est en effet nécessaire que l'information portée soit suffisamment homogène. La figure 4.1 représente l'ensemble des mots du lexique pour chacune des coupes bi-dimensionnelles de l'espace.

Sur les figures 4.1(a) et 4.1(b), on observe un léger éparpillement du nuage sur l'axe de la valence, autour de sa médiane : dans les travaux psychologiques, cet axe est considéré de manière consensuelle comme le plus discriminant pour organiser les états affectifs.

De plus, sur la figure 4.1(a) le nuage de points couvre la totalité du plan défini par la valence et l'activation ce qui semble indiquer une certaine indépendance entre ces deux axes. Nous constatons néanmoins que pour des valeurs d'activation faibles (inférieures à 20) comme pour des valeurs de valence élevées (supérieures à 50) une distribution des points plus clairsemée qui indique que peu d'états sont très positifs et que peu d'états sont décrits par une activation très faible, et ainsi que les mots correspondants sont rares dans le lexique.

Sur la figure 4.1(b), un motif en V est identifiable et semble indiquer une association entre les axes de valence et d'émotionalité. Dans le tableau 4.2 nous avons reporté les coefficients de corrélation linéaire de Pearson mesurés sur le nuage de points pour la coupe valence/activation d'une part, et pour la coupe valence/émotionalité d'autre part. Pour chacune des coupes, nous segmentons le plan en deux en séparant les valeurs inférieures à 40 des valeurs supérieures. Les mesures reportées sont spécifiques à chacune des régions.

On constate ainsi une forte corrélation entre les axes d'émotionalité et de valence dans la région supérieure et dans une moindre mesure une association certaine dans la région inférieure. Pour la coupe valence/activation, tandis que l'indice de corrélation indique une indépendance des deux axes dans la région inférieure, il montre une légère dépendance dans la région supérieure. Comme nous l'avons suggéré précédemment, cela peut être un effet secondaire lié à la parcimonie du nuage dans cette région.

Nous pouvons ainsi conclure que parmi les axes qui composent cet espace, l'information portée par l'émotionalité reproduit dans une certaine mesure celle portée par la valence. Une explication résiderait dans l'interprétation qui est faite de l'émotionalité par les sujets de l'expérience : il est en effet possible que ces derniers aient interprété ce descripteur comme une mesure d'intensité liée à la polarité. Il faut néanmoins noter que l'émotionalité n'est pas totalement corrélée à la valence et que cet axe peut par exemple décrire une information précieuse afin de distinguer des états platoniques d'états passionnés.

4.2.2 Projection des textes dans un espace sémantique

Nous présentons dans cette section la méthode proposée : pour chacun des textes étudiés, nous normalisons la casse des caractères de sorte que les textes soient tous représentés par des caractères minuscules. Comme présenté précédemment, le lexique contient de nombreuses formes déclinées, aussi nous n'effectuons pas d'autre forme de pré-traitements.

(a) Valence/Activation

(b) Valence/Emotionalité

(c) Activation/Emotionalité

FIGURE 4.1 – Projections $2D$ du lexique de Leleu.

	Activation/Valence	**Emotionalité/Valence**
Valence $<$ 40	-0.03	-0.5
Valence $\geq$ 40	0.36	0.67

TABLE 4.2 – Indice de corrélation linéaire de Pearson pour les coupes valence/activation et émotionalité/valence des états du lexique de Leleu.

La projection d'un texte dans l'espace sémantique consiste alors à identifier l'intersection entre les mots qui le composent et le vocabulaire défini dans le lexique. Pour ce faire, nous proposons de réaliser une simple identification de mots-clefs (*keyword spotting*) telle que présentée à la section 2.4, p. 49.

A la section 2.3.1, p. 45 nous avons relevé l'importance de la ponctuation, nous proposons d'ajouter au lexique une entrée configurable correspondant au point d'exclamation. Nous ajustons empiriquement ses coordonnées aux valeurs médianes des échelles de mesure pour la valence et l'émotionalité (i.e 40) et à 75% de l'échelle de mesure pour l'activation (i.e. 60).

4.3 Mises en œuvre expérimentales

Nous évaluons la méthode proposée au travers de deux études réalisées sur des données réelles : pour deux textes courts, nous exploitons les propriétés statistiques des nuages de

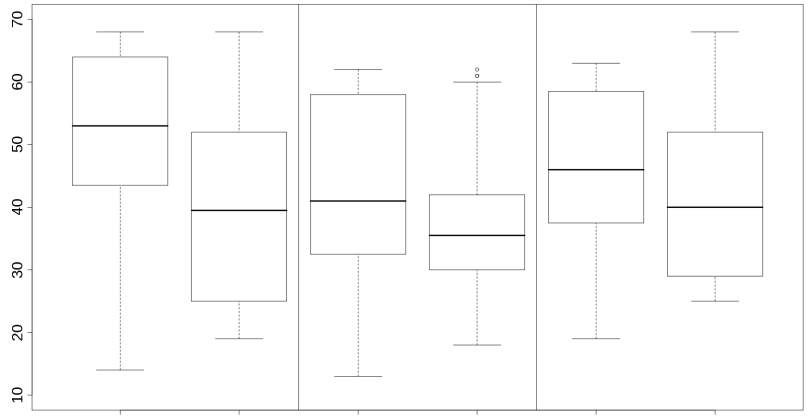


FIGURE 4.2 – Distribution de textes issus des paroles de *Hymne à la joie* (gauche) et *You are not alone* (droite) dans l'espace affectif associé au lexique de Leleu : valence (gauche), activation (milieu) et émotionalité (droite).

points afin de discriminer leur charge émotionnelle. Dans la seconde étude nous étudions la dynamique temporelle du signal affectif constitué dans l'espace de représentation sémantique en examinant son évolution temporelle.

4.3.1 Etude statique : discrimination d'états affectifs

Afin de discriminer les états affectifs liés à deux textes chargés émotionnellement, nous analysons les propriétés des nuages de points qui leurs sont associés. Ces deux textes sont respectivement issus des paroles de « Hymne à la joie »[1] et de la traduction française des paroles de « *You are not alone* » de Michael Jackson[2], les documents utilisés dans cette étude sont donnés en annexe B, p. 201. Nous appliquons à ces deux documents la méthode proposée. Nous souhaitons évaluer si elle permet d'identifier, d'une part l'enthousiasme global et le caractère positif de *Hymne à la joie*, d'autre part, la tristesse et le caractère globalement négatif de *You are not alone*.

Sur les trois axes de l'espace sémantique, la figure 4.2 représente, sous forme de diagramme en boîte, la distribution des nuages de points extraits des deux textes. Un diagramme en boîte représente la distribution d'une série de valeurs en représentant sa médiane dans une boîte définie par son 1^{er} et son $3^{ème}$ quartile. La position de la médiane dans cette boîte est interprétée comme un indicateur de déséquilibre, tandis que la hauteur de cette boîte comme un indicateur de dispersion. De plus, les exceptions, définies pour les valeurs supérieurs à une fois et demi l'écart inter-quartile, sont représentées par des points en dehors de la boîte. Enfin, les valeurs extrêmes (minimum et maximum) qui ne sont pas des exceptions sont représentées en dehors de la boîte par des barres horizontales.

On observe que 75% des mots détectés dans *Hymne à la joie* sont liés à une valence supérieures à 40, la moitié de ces mots est de plus associée à une valence très élevée (entre 53 et 65). On constate également que le vocabulaire identifié se répartit de manière uniforme sur l'axe d'activation avec une légère tendance pour des valeurs élevées : la moitié est associée à des valeurs entre 43 et 58 et seulement le quart est associé à des valeurs inférieures à 35.

1. Le texte est disponible à l'adresse `http://fr.wikipedia.org/wiki/Ode_a_la_joie`.
2. Le texte est disponible à l'adresse `http://www.lacoccinelle.net/242720.html`.

Pour *You are not alone*, on observe que le vocabulaire détecté est étonnamment positif ; cependant les mots liés à une valence très élevée (> 51) ne représentent que le quart du vocabulaire identifié. Sur cet axe le nuage de points est bien plus dispersé que précédemment, cela traduit l'emploi conjoint de mots très positifs et de mots très négatifs : le thème amoureux abordé dans les paroles semble expliquer ce phénomène. Sur l'axe d'activation, nous constatons que le nuage de points est concentré autour de valeurs faibles, en dépit des deux exceptions observées pour des valeurs élevées.

Pour *Hymne à la joie* comme pour *You are not alone* nous n'observons pas de différence marquée entre la valence et l'émotionalité : comme rappelé à la section précédente, ces deux axes exhibent une forte corrélation dans le lexique. Néanmoins sur l'axe d'émotionalité, nous constatons, pour le premier, une dispersion du vocabulaire légèrement plus élevée : cela semble dû à une utilisation de mots communs plus répétée dans *Hymne à la joie* que dans *You are not alone*.

Bien que les différences entre les états affectifs liés à ces deux documents puissent être plus marquées, l'approche proposée permet d'identifier un état plus positif et intense pour *Hymne à la joie* que pour *You are not alone*. Elle permet également de caractériser finement chacun de ces états.

4.3.2 Etude temporelle : courbes émotionnelles

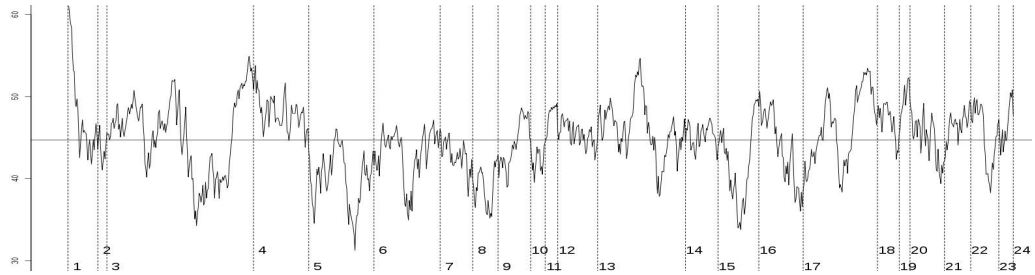
Nous réalisons une seconde étude dans laquelle nous exploitons l'espace sémantique pour identifier et suivre dans le temps l'état affectif correspondant aux dialogues du film *Little miss sunshine*[3]. Pour ce faire, nous avons segmenté le texte correspondant selon les 24 scènes du film[4]. La figure 4.3 tient compte de ce découpage et représente l'évolution du contenu émotionnel du film sur les axes de valence, d'activation et d'émotionalité. La charge émotionnelle du film est caractérisée au travers de la méthode proposée, à tout nouveau point de la figure correspond un terme identifié dans le lexique. Son évolution dans l'espace sémantique décrit une courbe émotionnelle qui est de plus lissée par moyennes mobiles exponentielles imposant un fenêtrage sur 20 mots. Pour chacun des axes, une ligne horizontale représente de plus l'état moyen obtenu en considérant l'ensemble des termes employés dans les dialogues.

Sur l'axe de la valence (figure 4.3(a)), avec une moyenne de 45 pour l'ensemble du film, on observe que la courbe se situe relativement haut par rapport à la valeur médiane de l'échelle de mesure. Ainsi, les dialogues de ce film semblent exhiber une tendance relativement positive. De plus, nous constatons que les scènes associés à une valence inférieure à la médiane, et donc à un vocabulaire négatif, correspondent systématiquement à des passages mettant en jeu des confrontations ou des disputes entre les personnages (scènes 3, 5, 8, 15, et 16). De même, les pics observés sur la valence capturent des passages marqués positivement ; en particulier les pics correspondants aux scènes 3 et 13 sont dûs à l'emploi d'un vocabulaire lié au thème amoureux.

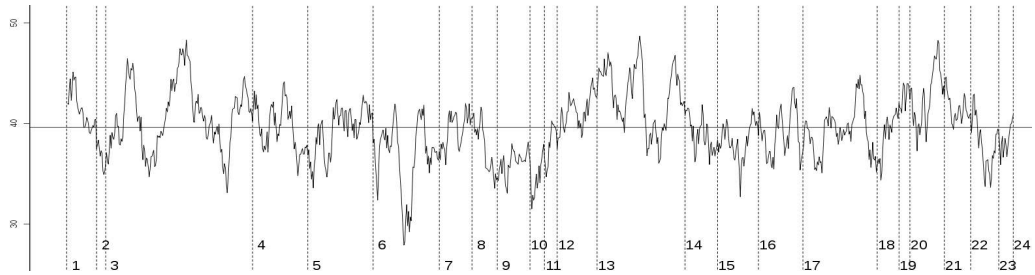
Le point culminant du film est atteint à la scène 20, l'action s'accélère et sur les figures 4.3(b) et 4.3(c) sont observés deux pics correspondants en activation et en émotionalité. Les pics formés sur ces deux axes sont indicateurs d'un enthousiasme général ou de passages fortement chargés émotionnellement. Ainsi au milieu de la scène 3 l'un des personnage raconte sa rencontre avec son mari, dans la scène 13 une mère explique à son enfant la mort de son grand-père. Au contraire, nous observons de faibles valeurs d'activation

3. Les dialogues sont extraits du sous-titrage français du film, disponible à l'adresse `http://www.opensubtitles.org/en/subtitles/3091816/little-miss-sunshine-en`

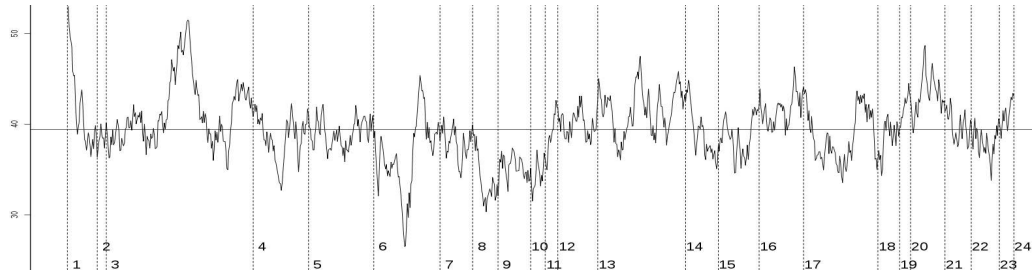
4. Les scènes ont été extraites du DVD et correspondent à des passages clefs de l'action.



(a) Valence.



(b) Activation



(c) Emotionalité

FIGURE 4.3 – Courbes émotionnelles obtenues sur la traduction française des dialogues du film *Little miss sunshine* : les états affectifs associés aux différents instants du film nécessitent une lecture verticale tandis qu'une lecture horizontale représente l'évolution de ces états, sur chacun des axes. Les lignes verticales correspondent à un découpage du film selon ses 24 scènes, les lignes horizontales représentent l'état moyen pour l'ensemble du film.

pour des passages plus sombres, représentés notamment par la scène 9 qui constitue un ralentissement dans la narration et la chute émotionnelle de l'un des personnages.

Bien que les courbes semblent refléter les passages clefs de l'histoire, nous observons quelques difficultés comme c'est le cas pour la scène 6 où une répétition des termes *grosse* et *gras*, jugés comme très peu actifs, peu émotionnels et très négatifs dans le lexique causent une chute marquée des courbes. De plus, dans la scène 12, une succession de termes positifs (entre autres *famille*, *confiance*, *amour*) ralentit la descente de la courbe sur l'axe de valence alors que le passage correspond à une scène de dispute.

4.4 Pistes d'enrichissement

Les résultats obtenus par la méthode proposée sont encourageants, deux points d'amélioration particuliers sont décrits dans cette section.

Couverture des textes Bien que le lexique que nous employons contienne de nombreux mots fréquents de la langue française, beaucoup de marqueurs d'émotions ne sont pas identifiés et manquent à la représentation des états affectifs. D'après les expériences que nous avons réalisées, nous constatons que pour de longs textes, ce problème de couverture est atténué par la densité du vocabulaire utilisé. Pour des textes plus courts cependant, il est possible que les marqueurs importants échappent à notre système. Il s'agit là d'un problème inhérent aux représentations sémantiques, auquel l'emploi de descripteurs bas niveau pourrait apporter une réponse.

Extraction avancée de descripteurs sémantiques En dépit de la simplicité du processus d'extraction que nous mettons en œuvre, la projection sémantique que nous proposons de construire permet de caractériser finement les états affectifs des documents. Néanmoins nous relevons certaines difficultés liées en partie au manque d'analyses avancées lors de l'extraction du vocabulaire : comme rappelé à la section 2.4.2, p. 50, la négation ainsi que les modificateurs d'intensité linguistique introduisent des imprécisions et des ambiguïtés. Aussi, l'approche présentée bénéficierait-elle de méthodes plus avancées pour l'extraction des descripteurs sémantiques.

4.5 Conclusions

Nous avons proposé une approche qui repose sur un enrichissement sémantique des documents pour l'étude de leur contenu émotionnel. La méthode proposée consiste à décrire un document comme un nuage de points dans un espace sémantique dont les axes mesurent les caractéristiques émotionnelles des états représentés. A cet effet, nous avons exploité un lexique qui associe aux mots des coordonnées dans un espace multi-dimensionnel pour décrire les émotions.

Nous avons réalisé deux expériences pour évaluer la pertinence de l'approche proposée, pour chacune nous avons utilisé le lexique de Leleu (1987) qui associe à un vocabulaire jugé le plus courant de la langue française, 3 000 états affectifs dans un espace composé des axes de valence, d'activation et d'émotionalité. Dans la première, nous avons discriminé les états affectifs associés à deux textes courts chargés émotionnellement. Dans la seconde nous avons exploité les dialogues d'un film afin de décrire l'évolution temporelle de leur contenu émotionnel. En dépit de la simplicité de l'approche proposée, nous avons montré

l'intérêt de l'enrichissement sémantique considéré, à savoir l'identification de marqueurs d'émotions et une caractérisation fine et graduelle de ces derniers.

Le manque de couverture des textes ainsi que les imprécisions introduites par le caractère très local des analyses s'imposent comme des limitations à l'approche proposée. Nous souhaitons exploiter cette méthode de manière conjointe à un apprentissage réalisé dans un espace de représentation bas niveau.

Au chapitre suivant nous proposons un modèle de représentation des émotions pour les textes dont l'enjeu est la prise en compte des imprécisions et des ambiguïtés inhérentes aux états affectifs, mais aussi la constitution de ressources sémantiques destinées à enrichir la description faite des documents dans un espace de représentation bas niveau.

Chapitre 5

Un modèle des états affectifs pour le texte

Dans la littérature, peu de travaux abordent le problème de l'analyse automatique de textes en tenant compte d'une modélisation fine des émotions : comme présenté au chapitre 2, les modèles de représentation employés pour une discrimination de concepts affectifs sont réduits à une catégorisation souvent bi-classe positif/négatif des émotions ; les représentations plus riches sont, elles, réservées à une caractérisation fine de la charge émotionnelle d'un document, cependant les méthodes utilisées ne mettent généralement pas en œuvre un apprentissage mais reposent plutôt sur un enrichissement sémantique des corpus. Au chapitre 3 nous avons observé les difficultés liées à un apprentissage de concepts affectifs dans un espace de représentation bas niveau, même lorsque le vocabulaire considéré modélise des descripteurs plus riches que les mots uniques. Au chapitre 4 nous avons réalisé deux études qui montrent l'intérêt d'un enrichissement sémantique des corpus reposant sur une représentation fine et graduelle des émotions.

Motivés par la prise en compte combinée de descripteurs bas niveau et de descripteurs sémantique, dans ce chapitre nous proposons un modèle pour décrire finement et graduellement les émotions en vue de leur analyse automatique dans les textes. Le modèle que nous proposons se place dans le cadre de la théorie des sous-ensembles flous (Zadeh, 1965) et repose à la frontière entre les représentations catégorielles et les représentations dimensionnelles des émotions.

Le modèle est détaillé à la section 5.1 ; en vue de la constitution d'un espace de représentation sémantique, un ensemble de recommandations est donné à la section 5.2 pour la constitution de ressources linguistiques à partir des caractéristiques proposées par le modèle. En supposant que soit disponible un tel lexique qui associe à un vocabulaire générique l'un des états affectifs décrits par le modèle, à la section 5.3 nous proposons et étudions la constitution d'un espace de représentation sémantique qui décrit finement et graduellement les émotions d'un document. Enfin, à la section 5.4 est présentée une mise en œuvre expérimentale du modèle sur des données réelles provenant du projet DoXa visant à l'identification et à l'analyse des émotions dans les textes.

5.1 Représentation graduelle des états affectifs pour le texte

Au delà de la dichotomie psychologique catégorielle/dimensionnelle exploitées respectivement dans les deux chapitres précédents, nous proposons un modèle de représentation des émotions adapté à leur analyse automatique dans les textes. Notre proposition repose

sur une catégorisation des émotions ainsi que sur la théorie des sous-ensembles flous afin de décrire des états complexes et imprécis : ces derniers sont décrits par un vecteur d'appartenance décrivant une association graduelle à des états basiques. Le modèle peut ainsi décrire des états qui dénotent totalement une émotion primaire, les combiner pour décrire une émotion transitoire, ou connoter des mélanges et représenter une charge émotionnelle imprécise. De plus, afin de distinguer les états platoniques des états passionnés, le modèle intègre une notion d'intensité lors de la description faite d'un état sur les émotions primaires.

5.1.1 Catégories sémantiques

Le modèle met en œuvre une représentation catégorielle des états affectifs : nous proposons l'emploi d'un ensemble fini de M émotions primaires que nous notons $\mathcal{C}$. Comme présenté à la section 2.2, p. 41, l'hypothèse catégorielle est motivée par l'existence d'un vocabulaire affectif dénotant des étiquettes émotionnelles. Dans le cadre d'un apprentissage supervisé, la construction de frontières de décision linéaires dans un espace de représentation bas niveau équivaut ainsi à l'identification du vocabulaire discriminant pour chacune des M émotions du modèle (voir section 2.4.1, p. 49). Bien que nous ne discutions pas du choix des émotions primaires composants $\mathcal{C}$, il est possible de considérer l'ensemble classique du *big six set*. Dans la suite, nous notons $M = |\mathcal{C}|$ la taille de cet ensemble.

De manière classique, nous associons de plus aux éléments de $\mathcal{C}$ une polarité positive ou négative :

$$\begin{aligned} pol : \mathcal{C} &\rightarrow \{-, +\} \\ c &\mapsto pol(c) \end{aligned}$$

En outre, comme c'est par exemple le cas pour la compétition I2B2 (*track2*) présentée au chapitre 3, dans les textes, il est souhaitable que les étiquettes considérées ne soient pas uniquement des émotions au sens psychologique, mais représentent également tout concept nécessaire au problème étudié, par exemple pour discriminer entre des états subjectifs et objectifs (par exemple *information* dans I2B2), ou affectifs et intellectifs (par exemple *valorisation* ou *désaccord*). Nous autorisons ainsi $\mathcal{C}$ à contenir des étiquettes plus générales, auxquelles nous faisons référence par *catégories sémantiques*, et nous décrivons ces dernières par une polarité *neutre*.

5.1.2 Gradualité par modélisation floue

Afin de prendre en compte la nature imprécise des émotions, nous proposons de plus de définir $\mathcal{C}$ comme une partition floue sur l'univers des documents (Zadeh, 1965). L'appartenance d'un document $d \in \mathcal{D}$ à un élément c de $\mathcal{C}$ est ainsi représentée par un degré d'appartenance :

$$\begin{aligned} \mu_c : \mathcal{D} &\rightarrow [0, 1] \\ d &\mapsto \mu_c(d) \end{aligned}$$

où μ_c vaut 0 lorsque d n'est pas associé à c, et vaut 1 quand l'appartenance de d à c est totale.

Tout comme pour l'hypothèse dimensionnelle, il est alors possible de décrire finement et graduellement un état affectif. Bien que notre proposition ne consiste pas en

une représentation continue de ces concepts, elle tolère une caractérisation pour des états transitoires comme les émotions complexes présentées à la section 2.2, p. 41. En effet un document associé à une étiquette *mépris*, défini par Plutchik (1990) comme un mélange des deux émotions primaires *dégoût* et *colère* peut par exemple être décrit par les degrés d'appartenance : $\mu_{\text{dégoût}} = 1$ et $\mu_{\text{colère}} = 1$.

De manière générale, un document est décrit par un vecteur de degrés d'appartenance :

$$\boldsymbol{\mu}(d) = (\mu_1, ..., \mu_M) \in [0, 1]^M$$

dont chacune des composantes représente une mesure de l'association de d à un élément c de $\mathcal{C}$. Cette représentation a été exploitée dans la littérature, notamment pour identifier des états transitoires comme des conjonctions d'états basiques (Subasic & Huettner, 2000; Moriyama & Ozawa, 2001; Martin et al., 2006; Fitrianie & Rothkrantz, 2008).

Il faut noter que lorsque pour tout c, les μ_c sont des degrés binaires (à valeurs dans $\{0, 1\}$), notre proposition se réduit à une représentation catégorielle classique des émotions, pour laquelle $\mathcal{C}$ définit un ensemble de concepts de la même manière que les modèles employés traditionnellement en apprentissage pour des concepts affectifs.

5.1.3 Intensité des états affectifs

Afin de représenter la nature positive ou négative d'un état affectif décrit par un vecteur d'appartenance sur $\mathcal{C}$, nous proposons d'associer les documents à une polarité :

$$pol : d \mapsto pol(d) \in \{-, mixte, +\}$$

où le rôle du niveau *mixte* est de caractériser les mélanges de concepts positifs et négatifs pour lesquels il est impossible de décider. Il faut noter que cette fonction de polarité est différente de celle définie pour les catégories sémantiques : la valeur *mixte* est notamment à différencier de la valeur *neutre* pour décrire les documents dépourvus de charge émotionnelle.

De plus, afin de distinguer les états platoniques des états passionnés, nous proposons d'associer aux documents, une intensité en remarquant que dans la littérature, polarité et intensité ont des rôles très dépendants. Nous proposons de définir définir une échelle de polarité augmentée, comprenant par exemple les 5 niveaux suivants $pol : \mathcal{D} \rightarrow \{--, -, mixte, +, ++\}$. Cette échelle décrit donc 2 niveaux d'intensité pour chacune des polarités, elle permet notamment de distinguer les états très positifs (resp. négatifs) de ceux qui le sont moins.

5.1.4 Relations avec les approches classiques

Nous mettons en évidence différentes instanciations du modèle, pour des tâches spécifiques d'analyse des émotions. Tandis que nous avons exploité certaines de ces instanciations dans nos travaux (voir section 5.4 notamment), nous donnons également des exemples d'utilisation intéressants mais non explorés.

Apprentissage bipolaire positif/négatif Lorsque $M = 0$ et que $\mathcal{C}$ est l'ensemble vide, alors les documents de $\mathcal{D}$ ne sont décrits que par leur polarité. Pour une tâche d'apprentissage automatique, l'ensemble des étiquettes est alors $\mathcal{Y} = \{--, -, mixte, +, ++\}$ si différents degrés d'intensité sont pris en compte et les méthodes d'apprentissage de frontières de décision dans un tel cadre s'inscrivent dans le domaine de l'*opinion mining* (Pang & Lee, 2008). Les corpus d'apprentissage étudiés dans ce domaine consistent

souvent en des corpus de commentaires utilisateurs sur Internet (*Internet user ratings*) où les documents sont étiquetés sur l'échelle de polarité. Plusieurs tâches peuvent alors être définies comme par exemple la discrimination entre les concepts *positif* et *négatif* ou la régression d'une fonction à valeurs sur l'échelle de polarité définie.

Apprentissage de concepts affectifs fins Lorsque $M > 1$ et que les degrés d'appartenance sont binaires alors le modèle que nous proposons permet de plus d'associer aux documents un ensemble d'étiquettes émotionnelles. Il est possible d'étiqueter des corpus d'apprentissage pour une tâche de classification multi-classes composée de M concepts. Cette tâche est celle de l'*emotion mining*, plus générale que la discrimination positif/négatif classique, elle est notamment étudiée dans le cadre de la compétition I2B2 (*track2*) (Pestian et al., 2012).

Caractérisation fine des émotions et états transitoires Quand $M > 1$ et que les degrés d'appartenance sont continus dans l'intervalle unité, le modèle que nous proposons autorise de plus une représentation fine des états affectifs et permet par exemple l'identification de clusters définissant des états transitoires. Ainsi, un cluster d'états dont le représentant est décrit par le vecteur $\boldsymbol{\mu}(d) = (\mu_{joie}, \mu_{peur}, 0, ..., 0)$ avec $\mu_{joie} > 0$ et $\mu_{peur} > 0$ caractérise un état exprimant à la fois de la *joie* et de la *peur* qui peut être interprété comme l'émotion complexe *dégoût* dans le modèle de Plutchik (1990).

Les degrés d'appartenance peuvent être définis et encoder l'imprécision inhérente aux états. Dans une tâche de discrimination des émotions, ils peuvent représenter des degrés de certitude associés aux décisions prises sur les M concepts étudiés (Hüllermeier, 2011). Dans ce cadre, il est de plus possible, comme proposé par Subasic et Huettner (2000), de comparer la charge émotionnelle des documents d'un corpus de manière visuelle.

Passage à une représentation dimensionnelle Il peut être souhaitable de décrire les états affectifs dans un espace multi-dimensionel. Etant donné un espace multi-dimensionnel dans lequel une région est associée à chacun des éléments de $\mathcal{C}$, il est possible de définir la projection d'un document dans cet espace comme le barycentre des catégories qui lui sont associées, pondéré par ses degrés d'appartenance.

5.2 Représentation des états affectifs pour les ressources linguistiques

Nous décrivons ici un ensemble de recommandations pour la constitution de lexiques sémantiques qui emploieraient le modèle de représentation des émotions que nous proposons pour décrire un vocabulaire générique.

5.2.1 Motivations et principe

Dans (Dzogang et al., 2010b), nous relevons plusieurs difficultés spécifiques à la modalité du texte : la négation joue un rôle important pour l'analyse des émotions et introduit de nombreuses ambiguïtés lors de la constitution des lexiques, de manière plus générale les ambiguïtés inhérentes au langage comme la polysémie font obstacle à la constitution de ressources regroupant un vocabulaire générique.

Ainsi, nos recommandations portent essentiellement sur les ambiguïtés liées au langage naturel et consistent par exemple en l'exploitation de marqueurs supplémentaires. Pour une

tâche d'apprentissage, lorsque ces marqueurs sont intégrés à l'espace de représentation des textes, ils constituent des descripteurs supplémentaires pour la construction d'une frontière de décision. Pour une tâche de caractérisation fine, des règles de traitement tirant parti du contexte d'énonciation peuvent mettre en œuvre ces marqueurs.

Les spécifications que nous proposons visent à enrichir les formes d'un document d'une information sémantique proche des concepts étudiés. Pour ce faire nous considérons qu'une fonction π (implémentée comme une grammaire d'extraction linguistique par exemple) annote les mots ou les groupes de mots d'un document selon quatre caractéristiques détaillées dans la suite : une catégorisation sémantique $\mathcal{C}$, une échelle d'intensité I, et deux marqueurs binaires indiquant respectivement une ambiguïté A ou une négation N. Formellement, la fonction π est la suivante :

$$\pi : f \mapsto ((\mu_1, t_1), ..., (\mu_M, t_M))$$

où pour tout i dans $[1..M]$, $t_i \in (I \times A \times N)$. Par exemple, la forme $f =$ *ni très en colère ni très ravi* peut être décrite comme :

$$\begin{aligned}\pi(f) = (&\\ &(\mu_{colere}, (9, 0, 1)),\\ &(\mu_{joie}, (9, 0, 1)),\\ &(0, (0, 0, 0)), ..., (0, (0, 0, 0))\\ )&\end{aligned}$$

5.2.2 Représentation des catégories

Tout comme pour les documents, nous recommandons de décrire le vocabulaire d'un lexique sur $\mathcal{C}$ au travers d'un vecteur de M degrés d'appartenance. Un forme identifiée dans un document est un état affectif, potentiellement complexe dans quel cas le vecteur décrit plusieurs composantes non nulles et exprime une conjonction des éléments de $\mathcal{C}$.

5.2.3 Marqueurs d'ambiguïté

Comme présenté à la section 2.3.2.1, p. 47, les entrées d'un lexique sont soumises aux ambiguïtés du langage : la polysémie en particulier est un problème récurrent pour les ressources qui regroupent un vocabulaire générique. Selon son contexte d'apparition le mot *explosion* peut par exemple faire référence à de la *joie* ou à de la *peur*, dans l'expression *un film terrible* le mot *terrible* n'a pas la même sémantique que dans l'expression *une terrible journée*. Cette ambiguïté peut donc porter sur la description sémantique faite des mots, mais aussi sur l'ensemble des autres caractéristiques (en particulier l'intensité et la négation présentée ci-dessous) qui leur sont associées.

Ainsi, nous recommandons l'usage d'un marqueur d'ambiguïté à valeurs dans $A = \{0, 1\}$ qui implémente une disjonction entre les différentes caractéristiques qu'il est possible d'associer à une forme. Dans un lexique, le mot *explosion* sera par exemple décrit par deux degrés d'appartenance $\mu_{\text{joie}} > 0$ et $\mu_{\text{peur}} > 0$ et un marqueur d'ambiguïté à 1. Selon son contexte d'apparition, l'un ou l'autre de ces deux états sera préféré.

5.2.4 Marqueurs de négation

Enfin, à un niveau différent de celui du lexique, la prise en compte des négations est importante dans les annotations fournies par π : c'est particulièrement le cas lorsque le

corpus d'étude emploie peu de vocabulaire. Bien qu'il soit possible de mettre à profit des catégories sémantiques antonymes dans $\mathcal{C}$ (par exemple *joie* comme opposé de *tristesse*), pour certaines catégories il n'existe simplement pas d'opposé naturel. Par exemple dans la catégorisation proposée par Plutchik (1990), aucune étiquette ne constitue de candidat au concept de *non dégoût*.

De plus, la négation peut adopter un comportement de modificateur d'intensité linguistique comme c'est par exemple le cas pour l'expression *pas très en colère*.

Ainsi, nous recommandons l'usage d'un marqueur de négation à valeurs dans $N = \{0, 1\}$ et dont le rôle est d'indiquer, aux chaînes de traitement ultérieures, les négations non résolues dans les descripteurs sémantiques extraits par π.

5.3 Construction d'un espace de représentation sémantique

Dans cette section nous considérons plus particulièrement une tâche d'apprentissage pour des concepts affectifs par opposition à leur modélisation présentée dans les section précédentes. Nous considérons ici que le modèle est implémenté au complet, c'est-à-dire que les documents d'un corpus d'apprentissage $\mathcal{D}$ sont étiquetés sur un ensemble de concepts cibles $\mathcal{Y}$ dont la définition dépend du problème considéré : par exemple $\mathcal{Y} = \mathcal{C}$ pour un problème d'*emotion mining* ou $\mathcal{Y} = \{--, -, \textit{mixte}, +, ++\}$ pour un problème d'*opinion mining*. Nous supposons de plus qu'il est fourni un lexique affectif ainsi qu'une fonction π qui suivent les spécifications proposées à la section précédente pour annoter les documents de $\mathcal{D}$.

Dans ce cadre nous étudions la constitution d'un espace de représentation $\mathcal{X}_{\text{sém}}$ formé de descripteurs sémantiques liés au modèle proposé. Nous motivons de plus le choix et la constitution de ces descripteurs et les avantages qu'ils apportent pour une tâche de discrimination fine des concepts. Comme au chapitre 4, l'espace que nous décrivons représente des états affectifs de manière graduelle, il est de plus motivé par son utilisation conjointe avec des espaces de représentation bas niveau (voir chapitre 3), à la section 5.4 nous présentons une telle mise en œuvre.

5.3.1 Principe général d'agrégation

A partir des caractéristiques associées aux annotations réalisées par π sur $\mathcal{D}$, de nombreux descripteurs peuvent être définis afin de raffiner l'information extraite du corpus pour les concepts cibles. De plus, contrairement aux méthodes de comptage classiques utilisées sur des descripteurs bas niveau, nous proposons une extraction plus appropriée pour des caractéristiques sémantiques en réalisant une agrégation pertinente de ces dernières. Dans ce cadre, un document est décrit par le vecteur $\boldsymbol{x} \in \mathcal{X}_{\text{sém}}$ qui contient les agrégats formés pour chacune des caractéristiques associées aux annotations.

Les agrégats considérés dépendent de la tâche d'apprentissage et sont spécifiques à chacun des concepts $\mathcal{Y}$ étudiés. Aussi lorsque les concepts décrivent une catégorisation des émotions, les agrégats sont définis pour chacune des catégories, lorsqu'ils décrivent une échelle de polarité, ils sont définis pour chacun de ses niveaux. Il faut noter que les concepts définis dans le lexique et ceux utilisés pour l'étiquetage de $\mathcal{D}$ peuvent différer, pour un étiquetage en polarité par exemple, les agrégats peuvent être formés pour chacune des catégories sémantiques, en considérant l'ensemble des catégories sémantiques de même polarité, ou pour cet ensemble décliné de plus selon des niveaux d'intensité.

Dans la suite nous précisons la nature de ces agrégats selon les caractéristiques considérées. De manière générale, en notant f une forme annotée par π dans le document d et $e_j(f)$ la

Opérateur	Sémantique	μ continus	μ binaires	Intensité	Positions
min	pessimiste	x	-	-	-
max	optimiste	x	-	-	-
moyenne	compensation	-	x$^{(*)}$	x	x
somme	renforcement faible	-	x	x	-
produit	renforcement fort	-	-	x	-
variance	critère de dispersion	-	-	x	x

TABLE 5.1 – Résumé des différents agrégats considérés et de leurs propriétés pour la construction d'un espace sémantique. (*) la moyenne est rapportée au nombre total d'annotations identifiées dans un document.

valeur associée à cette forme pour la caractéristique e et le concept j, nous construisons l'agrégat correspondant de la manière suivante :

$$\varphi\left[e_j(f)\right]_{\{f \in d/\mu_j(f)>0\}} \in \mathbb{R}$$

où selon la tâche considérée, j parcourt l'intervalle $[1..M]$ ou l'échelle de polarité, et $e_j(f)$ représente le degré d'appartenance ou l'intensité associé à la forme f pour le $j^{\text{ème}}$ concept. Dans la suite nous précisons la nature de φ pour chacun de ces deux cas, nous considérons de plus deux agrégats supplémentaires pour tenir compte d'indications spatiales et fréquentielles sur les annotations. L'ensemble des agrégats considérés est résumé dans le tableau 5.1.

5.3.2 Agrégations sur les degrés d'appartenance

Dans le cadre de la théorie des sous-ensembles flous, une fonction d'agrégation de degrés d'appartenance peut être vue comme une opération ensembliste et peut entre autres prendre la forme d'une *t-norme* ou d'une *t-conorme*. Deux cas particuliers de cette famille ont été présentés dans ce document (voir section 1.3.3.2, p. 29) : il s'agit respectivement des fonctions *minimum* et *maximum*. Nous proposons de construire deux agrégats à partir de ces deux fonctions : une sémantique pessimiste, respectivement optimiste, est associée aux agrégats correspondants.

5.3.3 Agrégations sur les intensités

Nous proposons d'exploiter des opérateurs qui présentent des propriétés de renforcement et de compensation pour l'agrégation des intensités. Comme rappelé à la section 1.3.3.2, la propriété de compensation impose un résultat intermédiaire entre la valeur minimale agrégée et la valeur maximale agrégée, celle de renforcement permet de modéliser les effets d'une accumulation émotionnelle de valeurs élevées ou basses dans les documents. Ici, les opérateurs correspondants permettent de compenser des intensités faibles par des intensités élevées ou bien d'accentuer une accumulation d'intensités élevées.

Ainsi, nous proposons de considérer trois agrégats respectivement obtenus à partir des fonctions *moyenne*, *somme* et *produit* : tandis que la moyenne est un opérateur de compromis qui exhibe des propriétés de compensation, la somme et le produit pour des valeurs toutes supérieures à 1, présentent des propriétés de renforcement positif, d'influence croissante. Pour le produit, ce renforcement est négatif pour des valeurs de l'intervalle unité.

De plus, nous souhaitons tenir compte de la dispersion des intensités autour de leur moyenne et nous considérons un agrégat supplémentaire consistant à mesurer leur variance.

5.3.4 Agrégations sur les positions et sur les fréquences

Nous souhaitons tenir compte de la position des annotations dans un document : les annotations peuvent en effet jouer un rôle différent, selon qu'elles sont identifiées en début ou en fin de document. De même leur éparpillement dans un document peut constituer une information pertinente pour décrire les concepts cibles. Ainsi, nous proposons de considérer deux agrégats pour la moyenne des positions des annotations et la variance de ces positions.

Nous considérons de plus la représentativité d'un concept dans un document au travers de deux agrégats supplémentaires, calculés respectivement comme le nombre d'annotations correspondantes, et comme leur fréquence relative à l'ensemble des annotations du document. Pour ce faire, nous proposons de former la somme des degrés d'appartenance binaires et cette même quantité rapportée au nombre total d'annotation produit par π sur d.

5.3.5 Traitement des négations et des ambiguïtés

Une méthode simple pour traiter des négations non résolues par π consiste à définir un nouveau concept antonyme pour tout concept associé à une négation non résolue. Ainsi, l'ensemble des agrégats présentés précédemment (voir tableau 5.1) sont également formés sur les concepts antonymes $\neg c$ définis pour les concepts $c \in \mathcal{C}$ pour lesquels il existe, dans d, une annotation associée à une négation non résolue.

Pour l'ambiguïté, nous proposons de former dans un premier temps les agrégats correspondants aux annotations non ambiguës puis, lorsque le contexte le permet, d'exploiter la fréquence des catégories sémantiques (resp. des niveaux de polarité) pour décider du sens dominant. Cette méthode simple permet de tenir compte du vocabulaire ambigu dans les lexiques, des méthodes plus avancées qui reposent sur des règles de désambiguisation ou sur un corpus d'apprentissage sont plus coûteuses mais permettent d'obtenir des résultats plus précis.

5.4 Mise en œuvre expérimentale

Le projet DGCIS [1] DoXa (Paroubek et al., 2010) vise au traitement automatique des états affectifs dans des ensembles de documents rédigés en français et en anglais. Les ensembles traités intègrent de grands volumes de données à la fois non structurées et issues de l'Internet comme par exemple de blogs, de forums, de médias d'information, ou de réseaux sociaux ; et de données structurées issues de bases de données clients. Le projet, coordonné par Thales, a réuni 11 partenaires, industriels avec entre autres EDF, Arisem, ILObjets, et Pertimm ; et académiques avec entre autres le LIP6/CNRS, le Limsi/CNRS, et l'IGM.

L'un des enjeux du projet est l'étude et l'exploitation d'une représentation plus fine qu'une catégorisation bi-classe, positif/négatif, des émotions, pour leur analyse automatique dans les documents. Le modèle que nous proposons s'inscrit dans ce cadre. Notre participation au projet a notamment mené à la mise en œuvre expérimentale que nous présentons dans cette section. Celle-ci a nécessité une collaboration étroite avec les partenaires linguistes du projet, qui ont constitué un lexique affectif, et étiqueté un corpus de documents selon une catégorisation fine des émotions. La mise en œuvre que nous présentons consiste en une discrimination automatique des émotions composant un corpus de documents non structurés, rédigés en français.

1. Direction Générale de la Compétitivité, de l'Industrie et des Services.

A la section 5.4.1 nous présentons le corpus qui nous est fourni ; l'enrichissement sémantique de ce corpus, produit par les partenaires linguistes du projet, et le lexique affectif constitué à cet effet sont détaillés à la section 5.4.2. Nous présentons, à la section 5.4.3, la construction d'un espace de description enrichi pour décrire les documents, ce dernier est composé aussi bien de descripteurs sémantiques tels que présentés à la section précédente, que de descripteurs bas niveau. A la section 5.4.4 nous présentons la mise en œuvre, réduite à la catégorie la plus fréquente du corpus. A ce titre sont comparées différentes représentations des documents. Les résultats obtenus pour la catégorie la plus fréquente du corpus sont détaillés à la section 5.4.5. Enfin, à la section 5.4.6 nous discutons des difficultés rencontrées et de leurs causes potentielles

5.4.1 Corpus d'apprentissage mis à disposition

Nous décrivons les documents considérés, l'instanciation faite du modèle proposé pour décrire les émotions, et nous proposons une analyse du corpus fourni en entrée.

Documents considérés Le corpus $\mathcal{D}$, destiné à l'apprentissage des concepts affectifs, est constitué de 82 textes *web* portant sur le thème du jeu vidéo, qui correspondent à des articles de blogs ou de média d'information, ainsi qu'à des messages provenant de forums de discussion. Ces textes sont segmentés en 650 paragraphes au total (dans le cadre du projet, un paragraphe est défini comme une suite consécutive de 100 mots).

Comme détaillé dans la suite, chacun des textes et chacun des paragraphes est associé à une ou plusieurs étiquettes émotionnelles ainsi qu'à un indice sur une échelle de polarité augmentée d'une intensité. Cet étiquetage est le résultat d'évaluations humaines organisées par les partenaires linguistes du projet.

Ainsi, le corpus est divisé en deux sous-corpus selon que les textes entiers, ou les paragraphes, vus comme de courts documents, sont considérés.

Instanciation du modèle La catégorisation des émotions employée dans le cadre du projet repose sur $M = 20$ catégories sémantiques qui font aussi bien référence à des émotions au sens psychologique comme *tristesse* ou *colère*, qu'à des états intellectifs comme *valorisation* ou *mésentente*. La colonne gauche du tableau 5.2 donne le contenu de $\mathcal{C}$: pour chacune des étiquettes, la colonne droite décrit la polarité définie pour la catégorie correspondante. Les documents (i.e. les textes ainsi que les paragraphes) du corpus $\mathcal{D}$ sont étiquetés par une ou plusieurs catégories de $\mathcal{C}$. Il faut noter que la tâche considérée étant une tâche de classification, les degrés d'appartenance aux catégories sémantiques (voir section 5.1.2, p. 82) n'ont pas été considérés. Une étiquette *neutre* identifie de plus les documents dépourvus d'étiquetage sémantique puisque dépourvus de charge émotionnelle.

De plus, comme présenté à la section 5.1.3, chacun des documents est lié à un indice de polarité sur l'échelle $\{--, -, \textit{mixte}, +, ++\}$. De manière classique, cette échelle spécialise chacune des polarité en deux niveaux d'intensité, le niveau *mixte* étant réservé aux états pour lesquels il est difficile de décider d'une polarité positive ou négative.

Caractéristiques Les tableaux 5.2(a) et 5.2(b) représentent respectivement les distributions des étiquettes émotionnelles sur les paragraphes et sur les textes. Pour les deux, nous observons un très grand déséquilibre entre les classes : les cinq catégories les plus fréquentes représentent à elles seules presque la totalité des concepts étiquetés ; sur les paragraphes et dans une moins grande mesure sur les textes, l'étiquette *accord* domine de plus le corpus de manière non négligeable.

Catégorie sémantique	polarité
apaisement	+
accord	+
connotation méliorative	+
plaisir	+
valorisation	+
satisfaction	+
surprise positive	+
colère	−
peur	−
déplaisir	−
dévalorisation	−
gêne	−
ennui	−
mépris	−
désaccord	−
surprise négative	−
tristesse	−
connotation péjorative	−
requête	*neutre*
recommandation	*neutre*

TABLE 5.2 – Catégorisation des émotions telle que définie par les partenaires linguistes du projet DoXa.

La distribution de l'échelle de polarité sur les paragraphes et sur les textes est représentée dans le tableau **??**. On observe une répartition des niveaux de polarité plus équilibrée que pour les catégories sémantiques. De plus, tandis que les paragraphes liés à une intensité élevée sont les plus fréquents, cette tendance est inversée pour les textes qui s'avèrent peu intenses en majorité. Dans les deux cas, l'étiquette *mixte* figure parmi les moins représentées.

De manière générale, on constate que pour les textes comme pour les paragraphes, les catégories intellectives comme *accord*, *désaccord* ou *valorisation* sont très fréquentes dans le corpus. De même, on constate une légère dominance des étiquettes positives sur les documents comme sur les paragraphes. Le corpus étant gouverné par la thématique du jeu vidéo, beaucoup de documents constituent des critiques ou des annonces de sorties de jeux : il est possible que ces observations indiquent l'expression de jugements de valeur modérés au sein du corpus.

Enfin nous relevons quelques incohérences d'étiquetage : certains documents ne présentent pas d'étiquette sur $\mathcal{C}$ et disposent néanmoins d'une étiquette sur l'échelle de polarité. En dépit de ces réserves, les annotateurs humains ont fourni un effort conséquent, et il est ici important de relever la difficulté majeure de l'étiquetage manuel de documents selon une catégorisation fine des émotions.

5.4.2 Lexique émotionnel mis à disposition

A partir de grammaires d'extraction semi-automatiques, les partenaires linguistes du projet ont annoté en émotions 9 600 formes sur le corpus. Contrairement à la méthode présentée au chapitre 4, p. 71, ces grammaires constituent des outils linguistiques avancés

(a) Distribution sur les paragraphes

Catégorie sémantique	**# paragraphes**
sans étiquette	269
accord	497
déplaisir	288
désaccord	268
valorisation	120
plaisir	113
satisfaction	94
recommandation	56
mépris	55
ennui	44
apaisement	38
surprise positive	37
colère	36
gêne	30
tristesse	29
surprise négative	12
requête	6
peur	3
connotation méliorative	2
connotation péjorative	1
dévalorisation	0

(b) Distribution sur les textes

Catégorie sémantique	**# textes**
sans étiquette	5
accord	53
déplaisir	36
désaccord	36
plaisir	29
valorisation	15
mépris	8
satisfaction	7
surprise positive	6
colère	4
apaisement	3
ennui	3
tristesse	2
requête	2
surprise négative	1
recommandation	0
gêne	0
peur	0
connotation méliorative	0
connotation péjorative	0
dévalorisation	0

TABLE 5.3 – Distribution des catégories sémantiques dans le corpus d'apprentissage fourni dans le cadre du projet DoXa.

qui mettent notamment en œuvre des méthodes pour l'identification et la résolution des négations (e.g. *pas du tout ravi*) et des modificateurs d'intensité (e.g. *très peu ravi*). Nous constatons que le taux de couverture des documents reste peu élevé : les annotations représentent à peu près 7% des mots employés en moyenne. Comme nous l'avons relevé auparavant, l'extraction de descripteurs sémantiques est sujet aux problèmes de couverture des documents. Nous observons ici certaines difficultés supplémentaires liées à la richesse du modèle et à la constitution manuelle des ressources.

Le lexique constitué pour produire ces annotations regroupe un vocabulaire constitué de mots ou de groupes de mots génériques, tenant toutefois compte du thème central au corpus, à savoir les jeux vidéo. Comme nous le proposons dans nos spécifications (voir section 5.2, p. 84) chacune de ses entrées est associée à une catégorie sémantique, ou plusieurs lorsque l'état décrit nécessite un raffinement. A chacune des catégories annotées est de plus associée une intensité sur une échelle plus précise que celle utilisée pour étiqueter les documents, puisqu'elle définit dix niveaux.

Comme recommandé, la polarité associée aux formes est induite par les catégories sémantiques, cependant contrairement aux spécifications les marqueurs de négation ainsi que les marqueurs d'ambiguïté et les degrés d'appartenance aux catégories ne sont pas implémentés dans les ressources. Il semble que ces informations soient difficiles à récolter manuellement. Nous constatons donc que la représentation fine proposée par le modèle constitue une richesse mais, selon les cas d'utilisation, peut s'avérer difficile à mettre en œuvre.

Le tableau 5.4 présente la distribution des catégories sémantiques sur les annotations

Catégorie sémantique	# annotations
valorisation	3 563
satisfaction	2 999
dévalorisation	1 764
désaccord	1 088
accord	807
requête	807
plaisir	491
tristesse	387
mépris	368
déplaisir	351
surprise négative	281
surprise positive	261
gêne	240
connotation péjorative	138
ennui	128
colère	115
peur	111
apaisement	44
inconfort	0
recommandation	0

TABLE 5.4 – Distribution des catégories sémantiques sur les annotations des formes du corpus d'apprentissage par les partenaires linguistes du projet DoXa.

sémantiques. Comme pour l'étiquetage des textes et des paragraphes nous observons un grand déséquilibre entre les étiquettes.

De plus, nous remarquons que les documents, constituant des pages *web*, ont fait l'objet d'un détourage automatique visant à éliminer le contenu non sémantique comme les menus de navigation ou les segments publicitaires. Néanmoins, au vu des documents filtrés, il semblerait que de nombreux segments aient été manqués et parasitent l'enrichissement sémantique des documents.

5.4.3 Construction d'un espace de représentation enrichi

Comme présenté à la section 5.3, nous proposons d'exploiter les annotations sémantiques décrites précédemment pour constituer un espace de représentation sémantique. Ainsi, pour décrire les documents du corpus $\mathcal{D}$, nous proposons de tenir compte à la fois de descripteurs bas niveau qui permettent la construction de frontières de décision dotées de bonnes propriétés de généralisation, et de descripteurs sémantiques qui offrent une forme de supervision à l'échelle des descripteurs et permettent ainsi d'obtenir des frontières plus précises. Ainsi, pour décrire les documents nous considérons le cadre de la fusion anticipée et nous étudions la construction d'un espace de représentation $\mathcal{X}$ qui repose à la fois sur un espace bas niveau $\mathcal{X}_{\text{raw}}$ et un espace sémantique $\mathcal{X}_{\text{sém}}$.

Descripteurs bas niveau Comme présenté au chapitre 3, nous considérons l'ensemble du vocabulaire associé à trois documents ou plus dans le corpus : il est explicitement tenu compte de la ponctuation et les entrées ne sont pas filtrées selon leur fonction grammaticale.

De plus, nous mettons à contribution une liste d'émoticônes (voir section 2.3.1, p. 45) qui décrit 98 entrées et que nous avons compilée à partir de ressources disponibles librement. Les mots du vocabulaire ainsi constitué sont réduits à leur lemme, nous utilisons pour ce faire le programme *TreeTagger* (Schmid, 1994). Enfin, le vecteur de représentation d'un document est obtenu en appliquant un schéma binaire, ses composantes indiquent de la présence ou de l'absence des mots du vocabulaire.

Comme nous le détaillons dans la suite, plusieurs espaces de représentation bas niveau $\mathcal{X}_{\text{raw}}$ sont considérés : nous examinons en effet les effets d'une spécialisation du vocabulaire tel que proposé au chapitre 3, nous étudions de plus différents mélanges de p-grammes pour des ordres p allant des unigrammes aux trigrammes. Pris de manière isolée, nous avons effectivement observé sur ce corpus une chute systématique des performances pour des ordres plus élevés.

Descripteurs sémantiques A partir des annotations fournies par les partenaires linguistes, nous proposons de former les agrégats présentés à la section 5.3. Néanmoins, ne disposant pas des degrés d'appartenance aux catégories sémantiques ni des marqueurs de négation et d'ambiguïté, pour chacune des catégories sémantiques e nous ne considérons que les 6 descripteurs suivants : le nombre d'annotations qui correspondent à e pour un document d donné et leur proportion par rapport au nombre total d'annotations dans d, la variance de leurs positions, et la moyenne, la somme, le produit ainsi que la variance des intensités correspondantes. Nous normalisons de plus la valeur des caractéristiques étudiées avant de former les agrégats correspondants, pour ce faire nous rapportons ces valeurs dans l'intervalle unité en appliquant une normalisation *min-max*.

Dans l'espace de représentation sémantique ainsi formé $\mathcal{X}_{\text{sém}}$, les $M = 20$ catégories sémantiques identifiés dans les formes du corpus sont chacune décrites par 6 dimensions.

Espace de représentation final : fusion anticipée Nous proposons alors de construire l'espace de représentation final $\mathcal{X}$ comme la concaténation de l'espace sémantique et de l'espace de représentation bas niveau : $\mathcal{X} = \mathcal{X}_{\text{sém}} \oplus \mathcal{X}_{\text{raw}}$. Contrairement à l'espace de représentation que nous proposons au chapitre 3, ici les espaces sont particulièrement hétérogènes : les descripteurs sémantiques sont en nombres bien moins importants que les descripteurs bas niveau. Afin d'homogénéiser les espaces d'origine, nous proposons de normaliser les vecteurs de représentation dans leurs espaces correspondant. Un vecteur de représentation final $\boldsymbol{x} \in \mathcal{X}$ s'exprime alors de la manière suivante :

$$\boldsymbol{x} = \frac{\boldsymbol{x}^{\text{sém}}}{\|\boldsymbol{x}^{\text{sém}}\|_2} \oplus \frac{\boldsymbol{x}^{\text{raw}}}{\|\boldsymbol{x}^{\text{raw}}\|_2}$$

où $\boldsymbol{x}^{\text{sém}} \in \mathcal{X}_{\text{sém}}$ et $\boldsymbol{x}^{\text{raw}} \in \mathcal{X}_{\text{raw}}$ décrivent respectivement l'information sémantique et bas niveau associée au document considéré.

5.4.4 Discrimination de concepts affectifs dans un espace enrichi

Nous avons réalisé des expérimentations visant à discriminer toutes les étiquettes émotionnelles ainsi que l'ensemble des niveaux de polarité : nous avons considéré une stratégie de classification « un contre tous » et nous avons mis en œuvre les espaces de représentation et le protocole décrit dans la suite. Cependant les performances constatées sont très en dessous des résultats escomptés, à la section 5.4.6, nous relevons plusieurs causes potentielles aux difficultés rencontrées. Dans la suite nous ne considérons que l'étiquette *accord* qui est la plus fréquente (ou la moins rare) du corpus. De plus, trop

peu de textes sont fournis pour l'apprentissage : nous ne considérons que le niveau des paragraphes.

Concepts cibles Nous concentrons notre étude sur les paragraphes et sur la catégorie *accord* qui est la plus fréquente du corpus. Ainsi, nous étudions dans la suite l'apprentissage de frontière de décisions pour discriminer entre le concept *accord* (497 documents) et le reste des étiquettes y compris les documents *neutres* (1 325 documents).

Représentations considérées Nous comparons les performances associées à différentes représentations des documents : nous étudions une description bas niveau qui repose de manière isolée sur l'ensemble des unigrammes (uni), des bigrammes (bi) et des trigrammes (tri), ou de manière combinée en considérant à la fois les unigrammes et les bigrammes (uni $\oplus$ bi) ; nous examinons aussi les performances réalisées dans l'espace de représentation sémantique (sém) décrit à la section précédente. Enfin, nous considérons l'espace proposé, à savoir une combinaison de descripteurs bas niveau et de descripteurs sémantiques (sém $\oplus$ uni), pour deux configurations selon que les vecteurs sont normalisés dans leurs espaces d'origine ou non.

Frontières de décision linéaires Comme présenté au chapitre 3 nous mettons en œuvre un apprentissage linéaire dans chacun de ces espaces : nous exploitons à ce titre un algorithme à vaste marge (Fan et al., 2008) et nous utilisons le coût de classification asymétrique présenté au chapitre 3 pour tenir compte du déséquilibre des classes.

Protocole expérimental Pour chacune des représentations étudiées, nous ajustons le coût de classification C par validation croisée sur 10 échantillons du corpus : pour ce faire nous considérons les puissances de 2 pour des exposants entre 0 et 10 et nous retenons la valeur pour laquelle le score $F1$ est en moyenne maximisé. Dans le tableau 5.5 sont reportés les scores moyens de $F1$, de rappel et de précision pour les frontières de décision ainsi obtenues dans chacun des espaces de représentation. La dernière colonne de ce tableau indique de plus la valeur retenue pour le coût de classification qui fournit une indication sur les propriétés de généralisation des frontières correspondantes.

5.4.5 Résultats et discussions

Les performances obtenues sont globalement faibles, ce que nous pouvons entre autres expliquer par la petite quantité de documents disponibles en apprentissage, le grand déséquilibre entre les classes, ainsi que les difficultés rencontrées et résumées à la section suivante.

Dans le tableau 5.5, on observe cependant que les descripteurs bas niveau extraient des documents une information moins pertinente que les descripteurs sémantiques pour décrire la catégorie *accord*. En effet ni les descripteurs génériques représentés par les unigrammes, ni les descriptions qui tiennent compte du contexte pour des ordres plus élevés ne permettent de construire une frontière de décision acceptable. De même, on constate que la spécialisation du vocabulaire reposant sur l'entropie de Shannon (voir chapitre 3) n'apporte pas de gain pour ce corpus : même si un gain négligeable est relevé sur la précision de la frontière correspondante. Pour décrire la catégorie *accord* il semblerait que le vocabulaire identifiable au travers des descripteurs utilisés ne soit pas suffisamment discriminant.

En revanche les descripteurs sémantiques offrent une meilleure caractérisation pour cette catégorie, au vu des scores de précision et de rappel associés : les descripteurs

Espace représentation	F1	Précision	Rappel	C
uni (sans spécialisation)	0.37 ± 0.07	0.37 ± 0.07	0.38 ± 0.08	4
uni	0.37 ± 0.08	$\mathbf{0.38 \pm 0.07}$	0.35 ± 0.08	256
bi	0.34 ± 0.08	0.33 ± 0.08	0.35 ± 0.09	1024
tri	0.32 ± 0.05	0.29 ± 0.05	0.35 ± 0.05	512
uni $\oplus$ bi	0.36 ± 0.05	0.35 ± 0.06	0.36 ± 0.06	1024
sém	$\mathbf{0.45 \pm 0.06}$	0.36 ± 0.05	0.59 ± 0.07	1
sém $\oplus$ uni	0.38 ± 0.04	$\mathbf{0.38 \pm 0.07}$	0.38 ± 0.04	512
sém $\oplus$ uni (normalisé)	$\mathbf{0.45 \pm 0.02}$	0.32 ± 0.02	$\mathbf{0.80 \pm 0.05}$	64

TABLE 5.5 – Performances de frontières de décision linéaires apprises par SVM pour discriminer la catégorie *accord* du reste des étiquettes dans un espace de représentation enrichi : les descripteurs bas niveau correspondent aux unigrammes (uni), aux bigrammes (bi) ainsi qu'aux trigrammes (tri). Un espace de représentation sémantique (sém) décrit pour chacune des $M = 20$ catégories sémantiques 6 agrégats sur leurs annotations.

sémantiques semblent jouer le rôle de marqueurs d'émotions que les descripteurs bas niveau ne peuvent assumer. Les valeurs retenues pour le coût de classification C, sont les plus importantes pour les descripteurs bas niveau (uni $\oplus$ bi, bi, tri), au contraire le coût le plus faible est associé aux descripteurs sémantiques. L'algorithme mis en œuvre est contraint à un sur-apprentissage pour des représentations bas niveau, ce manque de vocabulaire discriminant peut ainsi être compensé par l'exploitation de ressources sémantiques qui implémentent une information plus proche des concepts affectifs.

On remarque de plus que la stratégie de fusion proposée réalise une faible performance lorsque les vecteurs de représentation ne sont pas normalisés dans leurs espaces d'origines : les descripteurs sémantiques, bien moins nombreux que les descripteurs bas niveau, se trouvent submergés dans l'espace de représentation final $\mathcal{X}$ qui ressemble alors plus à $\mathcal{X}_{\text{raw}}$ qu'à $\mathcal{X}_{\text{sém}}$. La repondération qui consiste à combiner des vecteurs de représentation normalisés produit un résultat légèrement meilleur : bien que le score $F1$ ne dépasse pas celui obtenu de manière isolée sur les descripteurs sémantiques, on constate une légère baisse de sa variance (-0.04) liée à une augmentation importante du rappel ($+0.42$). Dans ce contexte, cette stratégie permet d'égaliser la contribution de chacun des espaces d'origine pour la représentation finale qui est faite des documents. De plus, l'espace de représentation bénéficie à la fois de la précision liée aux descripteurs sémantiques ainsi que de la couverture des textes liée aux descripteurs bas niveau.

5.4.6 Limitations et pistes d'enrichissements

Au vu des difficultés rencontrées, les objectifs donnés pour cette mise en œuvre expérimentale ont été révisés. Nous discutons dans la suite des problèmes identifiés et de leurs causes potentielles, nous discutons aussi de la pertinence des résultats obtenus et nous envisageons deux pistes d'enrichissement.

Résumé des difficultés rencontrées Dans les données qui nous sont fournies, nous identifions des problèmes à deux niveaux, à savoir : l'instanciation du modèle pour les enrichissements sémantiques et le corpus d'apprentissage mis à disposition.

Dans le cadre du projet, l'instanciation faite du modèle proposé pour décrire les émotions reste en effet partielle. Les enrichissements sémantiques ne considèrent ni les

degrés d'appartenance aux catégories sémantiques, ni les marqueurs de négation, ni les traits d'ambiguïtés (voir section 5.4.2) et l'information extraite par les descripteurs sémantiques ne jouit pas de toute la richesse du modèle. En particulier, les formes annotées sont sujettes à de nombreuses imprécisions, incertitudes et ambiguïtés qu'il nous est impossible d'identifier.

Les analyses réalisées sur le corpus semblent indiquer que les documents, qui constituent en majorité des critiques et des annonces de sortie de jeux vidéos, expriment des jugements de valeur modérés. Dans ce contexte il est d'autant plus difficile d'identifier des marqueurs d'émotions pour chacune des étiquettes considérées. Son étiquetage présente des taux d'accord inter-annotateurs relativement faibles, la petite quantité de documents réservés à leur apprentissage fait de plus obstacle à la construction de frontières de décision robustes. C'est particulièrement le cas pour les textes entiers et les catégories les moins fréquentes du corpus. Les documents considérés sont des textes *web*, il s'agit de données réelles soumises aux difficultés liées aux pré-traitements appliqués. Nous avons en particulier rappelé la difficulté de leur détourage automatique, en partie responsable du bruit observé dans les espaces de représentation bas niveau et sémantique.

Nous insistons cependant sur l'effort fourni par les partenaires du projet pour constituer le corpus et les ressources. Le modèle proposé pour décrire les émotions offre une représentation fine des états affectifs mais cela peut constituer un désavantage lorsqu'il est fait appel à des sujets humains pour son instanciation. Il serait intéressant d'étudier l'intérêt d'une méthode d'expansion automatique telle que présentée au chapitre 2, p. 39 comme approche alternative.

Pertinence des résultats Les difficultés rencontrées conditionnent la pertinence des résultats présentés. La comparaison proposée entre les différentes représentations pour décrire les documents a été réalisée sur d'autres catégories sémantiques ainsi que sur les textes pris dans leur ensemble, seulement les résultats obtenus, non significatifs et sens apport particulier, sont ici omis.

Enfin, bien que les observations réalisées à la section 5.4.5 demandent des analyses plus avancées, portant notamment sur la nature des descripteurs retenus comme les plus discriminants, les performances constatées font obstacle à de telles analyses.

Hétérogénéité des descripteurs Les descripteurs sémantiques sont de nature intrinsèquement différente par rapport aux descripteurs bas niveau et comme rappelé à la section 1.1.3.2, p. 14, un noyau linéaire n'est pas nécessairement le plus adapté dans un espace de représentation sémantique. Contrairement aux descripteurs bas niveau, sur les descripteurs sémantique peut être caractérisée une information pertinente autour d'une certaine valeur moyenne (par exemple une intensité très faible, ou des annotations positionnées en tout début de document), or le produit scalaire ne s'active que pour des valeurs élevées. L'emploi d'une autre mesure de comparaison permettrait alors de caractériser d'autres types de phénomènes comme par exemple l'accumulation des intensités autour d'une certaine valeur ou des dispersions d'annotations dans un certain rayon.

Pondération des espaces d'origine La stratégie de normalisation que nous avons adoptée dans le cadre d'une fusion anticipée permet d'équilibrer la contribution de chacun des espaces d'origine. Ici les descripteurs sémantiques semblent jouer un rôle important pour discriminer la catégorie *accord* du reste des étiquettes. Il semble alors d'intérêt de considérer des pondérations différentes pour les espaces de représentation d'origine.

5.5 Conclusions et perspectives

Nous avons présenté un modèle pour décrire finement la charge émotionnelle d'un document qui s'inspire des travaux psychologiques et linguistiques sur les émotions. Ce modèle exploite une catégorisation floue des émotions et spécifie, de plus, un niveau d'intensité pour distinguer les états platoniques des états passionnés. Dans notre modèle, un état affectif est décrit par un vecteur d'appartenance aux catégories sémantiques, il peut ainsi dénoter totalement une émotion primaire, en combiner certaines pour décrire une émotion complexe, ou connoter des mélanges et décrire une charge émotionnelle imprécise. La richesse du modèle est adaptable à la tâche considérée : nous avons mis en évidence différentes instanciations qui permettent de formuler les tâches principalement étudiées dans l'état de l'art.

Nous avons de plus présenté un ensemble de spécifications pour la constitution d'enrichissements sémantiques reposant sur le modèle, à ce titre nous préconisons d'adapter la représentation faite des états affectifs à la granularité des mots et des groupes de mots. Nous recommandons notamment l'emploi d'un marqueur d'ambiguïté pour pallier les problèmes de polysémie inhérents au langage, et d'un marqueur de négation pour indiquer des négations non résolues dans les annotations. Un lexique affectif reposant sur le modèle offre alors la possibilité de décrire les états affectifs identifiés au niveau de la phrase de manière fine et graduelle.

En supposant qu'un tel lexique soit disponible et qu'une fonction d'annotation mette en œuvre des outils linguistiques avancés pour l'extraction des mots et des groupes de mots, nous avons également proposé un ensemble de descripteurs sémantiques pour décrire les documents. L'espace considéré décrit des agrégats pour chacune des caractéristiques associées aux annotations comme l'intensité ou les degrés d'appartenance, spécialisés sur chacun des concepts définis pour le problème considéré. Les agrégats sont eux-mêmes le résultat d'une synthèse pertinente : la fonction d'agrégation est définie selon la nature des caractéristiques étudiées et les propriétés que l'on souhaite mettre en évidence, comme par exemple l'accumulation d'intensités élevées ou la dispersion des annotations dans un document.

Le modèle présenté est mis en œuvre dans le cadre du projet DoXa, nous avons à ce titre présenté le corpus étiqueté ainsi que le lexique affectif mis à disposition par les partenaire linguistes du projet. Au vu des difficultés rencontrées, les objectifs de cette mise en œuvre ont été révisés et nous avons concentré notre étude sur de courts documents réduits à 100 mots, et sur la catégorie la plus fréquente du corpus. Les résultats obtenus semblent néanmoins conforter l'hypothèse qu'un enrichissement sémantique des représentations bas niveau est important pour décrire les concepts affectifs : la combinaison de ces deux types de descripteurs nécessite alors une normalisation des représentations faites dans les espaces d'origine.

De nos discussions ressortent deux pistes d'enrichissement : la richesse du modèle est un avantage mais, comme nous l'avons observé, sa complexité peut aussi constituer un obstacle à la constitution de ressources sémantiques. Il serait intéressant d'étudier l'adaptabilité des méthodes d'expansion automatique à des caractéristiques plus fines telles que proposées dans le modèle. De plus, dans nos expérimentations nous avons constaté les bienfaits d'un espace de représentation enrichi tenant compte de descripteurs bas niveau et de descripteurs sémantiques. Pour aller plus loin dans cette voie il serait intéressant de considérer également une repondération adaptée des espaces d'origine, ainsi qu'une mesure de comparaison adaptée aux descripteurs considérés. Pour ce faire, une stratégie de fusion intermédiaire offre un cadre d'étude naturel et prometteur.

Deuxième partie

Informations dynamiques

L'adoption du *web* comme média d'information privilégié ainsi que l'augmentation massive de documents produits quotidiennement sur Internet suscitent l'intérêt dans de nombreux domaines sur l'étude des évolutions temporelles et du dynamisme inhérent à l'information. Cette seconde partie de nos travaux s'ancre dans un tel cadre en considérant toutefois que l'information étudiée est produite par un ensemble de sources dynamiques qui publient des documents textuels à intervalles de temps fréquents. Cette partie s'inscrit ainsi comme une seconde phase de nos travaux, qui vise à l'identification ainsi qu'au suivi de communautés de sources formées autour de thématiques communes. Afin de caractériser les regroupements de sources dans le temps nous tenons spécifiquement compte de la sémantique associée aux communautés, aussi les analyses que nous effectuons se placent dans l'espace de description des documents publiés.

Pour le problème que nous considérons, un ensemble de sources est en mouvement permanent dans un espace de description textuel. A tout moment leur position est donnée à la fois par leur historique de publication et par les documents publiés à l'état courant. De telles données sont dynamiques par nature, leur partitionnement s'éloigne d'une tâche de clustering traditionnelle pour laquelle les données sont supposées stationnaires. Au chapitre 6 nous faisons état des problèmes de partitionnement classiques pour les données dynamiques : pour chacun nous mettons en évidence les différences et les ressemblances avec la tâche de *clustering de sources dynamiques* que nous proposons d'étudier. Au cours du temps, les communautés identifiées se recomposent en permanence au gré des changements d'intérêt des sources : les partitions obtenues sont régies par un fort dynamisme. Nous proposons ainsi d'observer les évolutions encourues dans le temps à la fois sur la sémantique associée à ces regroupements mais aussi sur les renouvellements de populations engendrés. Nous proposons enfin de définir un *thread d'information* comme une communauté temporelle, gouvernée par une thématique remarquablement stable durant un intervalle de temps suffisamment long.

Les sources constituent par ailleurs la donnée du problème que nous considérons or leur définition et leur identification dépendent fortement de la tâche étudiée. Sur Internet par exemple, il apparaît souvent que l'ensemble des documents publiés sur un même domaine représente une information très hétérogène. Au chapitre 7 nous proposons de décomposer un domaine de publication en un sous-ensemble de sources homogènes en exploitant les urls associées aux documents publiés. Nous étudions en particulier un partitionnement hiérarchique des urls représentées comme des ensembles de tokens et nous associons une source à tout regroupement réalisé au sein du dendogramme correspondant. Pour ce faire, nous considérons des dendogrammes de deux types : nous imposons d'une part que le partitionnement respecte un certain critère de cohérence, d'autre part que les regroupements effectués soient compacts. Nous étudions par ailleurs deux modes de partitionnement : d'une part un regroupement par lots qui nécessite la totalité des urls disponibles, d'autre part une construction incrémentale, à mesure que de nouvelles urls sont observées. Nous proposons alors deux méthodes pour l'identification de sources sur Internet qui constituent deux extrêmes pour ces deux axes : nous étudions expérimentalement l'intérêt des deux méthodes proposées sur un corpus de données réelles.

Dans un espace de description textuel, les sources dynamiques sont représentées par un vecteur de très grande dimension, leur partitionnement fait alors face au dilemme du fléau de la dimension. Au chapitre 8 nous proposons un nouvel algorithme de clustering qui étend les K-moyennes sphériques pour effectuer une pondération des descripteurs. Nous définissons ainsi une transformation qui change l'espace d'entrée en un ellipsoïde sur laquelle les dimensions de l'espace originel sont contractées ou dilatées selon leur capacité à mettre en évidence une structure de clusters dans les données. Nous supposons que les

clusters reposent sur des régions denses de l'espace de description et nous associons un ellipsoïde spécifique à chacun. L'aplatissement des ellipsoïdes est ajusté par un paramètre pour lequel nous proposons une procédure de sélection automatique. Nous réalisons alors une étude comparative expérimentale de l'algorithme proposé sur des données synthétiques et sur des données réelles.

L'ensemble des méthodes proposées est mis en œuvre au chapitre 9 dans lequel nous réalisons une étude expérimentale des publications de la presse française sur Internet. Nous avons collecté durant cinq mois les articles publiés quotidiennement sur les fils de syndication des principaux médias d'information en France. Nous avons ainsi constitué un corpus de sources dynamiques correspondant aux principaux médias et blog d'information français. Nous réalisons une étude préliminaire du rôle des paramètres sur le dynamisme régissant les partitions. Nous étudions enfin des évènements précis d'actualité au travers des threads d'information identifiés sur la période d'étude.

Chapitre 6

Clustering de sources dynamiques

Nous étudions le problème de partitionnement d'un ensemble de sources qui émettent régulièrement de nouveaux documents. Sur les réseaux sociaux ces sources peuvent par exemple représenter des utilisateurs qui produisent un contenu textuel en continu, elles peuvent également être associées à des médias d'information qui publient des articles sur Internet. Ce dernier cas est le cas réel étudié au chapitre 9.

Les communautés auxquelles nous nous intéressons caractérisent des regroupements de sources, identifiés au travers des documents qu'elles publient : à toute communauté est ainsi associée une thématique commune qui acquiert un statut fédérateur auprès de ses membres. Pour ce faire nous proposons de décrire une source comme un vecteur dans un espace de représentation textuel. Les composantes de ce vecteur sont alors des termes qui résument l'ensemble des documents publiés par le passé. Une source est ainsi une donnée qui se déplace dans cet espace au gré des documents qu'elle émet au cours du temps. Un ensemble de sources ainsi représentées constitue alors un corpus de données dynamiques.

Le partitionnement de données dynamiques impose de nombreux défis, à la section 6.2 nous présentons trois tâches classiques de clustering qui diffèrent selon le mode de représentation des données mais aussi selon l'objectif considéré. Pour chacune nous faisons apparaître les ressemblances et les différences avec la tâche de clustering de sources dynamiques que nous étudions dans ce chapitre.

A la section 6.3 nous présentons en détail la représentation que nous faisons des sources. A ce titre nous nous plaçons dans le paradigme incrémental : les pré-traitements nécessaires à la constitution des vecteurs sont effectués en ligne, de sorte qu'à tout moment une simple mise à jour des coordonnées soit requise.

Nous formulons à la section 6.4 le problème du clustering de sources dynamiques comme un problème de clustering incrémental : à tout moment, les sources sont regroupées en communautés homogènes, leurs affectations étant décidées d'après les documents publiés jusqu'à l'état courant. Un tel partitionnement encourt de nombreux changements au cours du temps, aussi proposons-nous dans un premier temps d'observer les déplacements des communautés dans l'espace de description. Dans un second temps, nous considérons les transitions effectuées par les sources au sein des communautés. Enfin nous définissons un *thread d'information* comme une communauté qui répond à un certain critère de cohérence dans le temps.

Enfin, les conclusions de ce chapitre sont données à la section 6.5.

6.1 Contexte et motivations

Les données auxquelles nous nous intéressons sont des sources d'information qui publient des documents textuels à intervalles de temps fréquents. Sur Internet par exemple, le média `http://www.lemonde.fr` est une source qui émet de nouveaux articles en continu. Lorsque les sujets abordés par cette dernière deviennent fédérateurs pour un ensemble suffisamment grand de sources, une communauté se crée autour d'une thématique précise. Au sein de sa communauté, une source est précurseur si elle est à l'origine de cette thématique, elle peut alors lui rester fidèle et continuer de publier des articles similaires, elle peut aussi changer d'intérêt et transiter vers une autre communauté.

Etant donné un ensemble de sources observées en continu, des communautés se forment et se désagrègent au gré des documents nouvellement publiés, aussi proposons-nous de représenter une source comme un vecteur dans l'espace de description des documents : à un instant donné, une source est décrite par un vecteur de mots qui résument la totalité des articles produits depuis sa création. Après un certain temps, l'historique de publication d'un média comme `http://www.lemonde.fr` pouvant atteindre une taille conséquente, nous proposons par ailleurs de pondérer l'importance des documents en fonction du temps. Une source ainsi représentée est alors un point qui se déplace (perpétuellement) dans l'espace de description des documents, un ensemble de tels points constitue un corpus de données dynamiques.

Afin de caractériser au plus tôt les évolutions encourues au sein des communautés, nous souhaitons réaliser un partitionnement de ce corpus à tout moment. De plus les sources, telles que nous proposons de les décrire, changent continuellement : il est nécessaire, à tout nouveau pas de temps, de réévaluer la cohérence des communautés précédemment formées. Nous proposons ainsi d'étudier une tâche de clustering de sources dynamiques qui vise, d'une part, à identifier les communautés regroupées autour des sujets les plus fédérateurs, et d'autre part, à suivre l'évolution de ces communautés dans le temps. Les sources étant par ailleurs décrites dans l'espace de représentation des documents, il est possible d'associer une sémantique textuelle aux communautés identifiées.

6.2 Travaux similaires

Lorsqu'un ensemble de données est distribué sur un ensemble de clusters qui réalise une partition du corpus originel, la recherche de cette partition en absence de supervision constitue une tâche de clustering traditionnel. Il est alors supposé que les données sont stationnaires, autrement dit que les clusters n'évoluent pas dans le temps, ni en nombre ni en nature.

Dans un contexte dynamique en revanche, les données évoluent au cours du temps et l'hypothèse de stationnarité ne tient plus. Le corpus est alors vu comme un *flux de données non stationnaires* pour lequel les méthodes classiques de clustering ne peuvent être directement employées. Les données dynamiques imposent par ailleurs de nouveaux défis liés à la temporalité mais aussi à la grande quantité d'information traitée.

Principe général Dans la littérature plusieurs tâches tiennent spécifiquement compte du dynamisme inhérent aux données. Ces dernières peuvent être regroupées au sein du paradigme de l'*analyse de flux de données* qui se distingue sur trois points de l'analyse de données statiques (Muthukrishnan, 2005) : 1) les données sont produites (par un ensemble de sources) de manière continue : à un instant donné, le système n'a pas connaissance des données à traiter aux instants ultérieurs ; 2) la quantité d'information est trop importante

pour être stockée et de fortes contraintes sont imposées sur l'occupation mémoire ainsi que sur les temps de traitement ; 3) enfin, les réponses du système sont effectuées en temps réel, il est désirable que l'information soit traitée en une seule passe.

Un autre défi caractéristique des problèmes d'apprentissage sur des flux de données réside dans l'adaptation des systèmes aux évolutions observées dans le temps. Les données ne sont effectivement pas supposées stationnaires et leur partitionnement subit de nombreux changements au cours du temps. Ces changements peuvent influer sur la taille des clusters, leur nombre ou encore leur sémantique (Tsymbal, 2004; Crespo & Weber, 2005; Núnez et al., 2007). Ces évolutions sont identifiées comme des *dérives de concepts* (*concept drift*) (Widmer & Kubat, 1996) ou selon la rapidité des changements, comme des *basculements de concepts* (*concept shift*) (Bifet & Kirkby, 2009).

Dans la suite nous faisons état des tâches d'apprentissage non supervisées, étudiées de manière classique dans le domaine de l'analyse de flux de données. Nous proposons d'organiser ces tâches selon la nature des données étudiées mais aussi selon les finalités des méthodes proposées. Pour chacune nous mettons en évidence les points de ressemblance et de divergence avec la tâche de *clustering de sources dynamiques* que nous proposons d'étudier au sein de ce chapitre. Dans un premier temps nous présentons une tâche de *clustering sur des flux de données*, dans un second temps nous détaillons une tâche de *clustering sur des graphes dynamiques*, enfin nous décrivons une tâche de *clustering de séries temporelles*.

6.2.1 Clustering sur des flux de données

Représentation des données Une première catégorie d'approche traite des données produites de manière séquentielle qui forment un *flux de données* : les données sont elles-mêmes statiques, leur distribution évolue, elle, au cours du temps. Les articles publiés en continu sur Internet par différents médias d'information composent par exemple un flux de données textuelles, dont les thématiques évoluent avec l'actualité.

Au sein d'un flux, il est nécessaire d'ajuster l'importance de l'historique au travers d'une fonction de vieillissement. A l'instant t, une donnée $\boldsymbol{x}(t)$ est un vecteur qui s'exprime sous la forme suivante :

$$\boldsymbol{x}(t) = \alpha(t - t_0)\boldsymbol{x}$$

où $\alpha : \delta t \mapsto [0, 1]$ est une fonction de vieillissement qui associe à un écart temporel $\delta t = t - t_0$, une importance qui lui est inversement proportionnelle. Il est classique de définir α comme un vieillissement exponentiel paramétré (Aggarwal et al., 2004), ou bien comme un fenêtrage temporel dans quel cas α produit un poids à valeur binaire (Beringer & Hüllermeier, 2006).

Objectif Il est supposé que le flux est distribué sur un ensemble de clusters dont le nombre et la nature sont susceptibles d'évoluer, l'enjeu est alors l'identification (à tout moment) du partitionnement des données le plus fidèle aux derniers changements observés.

La très grande quantité de données à traiter à tout moment est le principal défi étudié : à l'état courant le flux est préalablement résumé avant tout partitionnement. Cette étape de résumé, aussi appelée pré-clustering, constitue souvent une étape indispensable. Aussi les auteurs concentrent-ils leurs efforts sur des structures de résumé efficaces, telle que les micro-clusters par exemple (Aggarwal et al., 2003; Cao et al., 2006; Kranen et al., 2011; Ackermanny et al., 2012).

Méthodes Comme évoqué précédemment, un algorithme de clustering sur des flux doit tenir compte des contraintes liées aux dérives de concepts, à la quantité de données qui ne peuvent être stockées en intégralité, ainsi qu'aux temps limités de traitement. A cet effet, il n'est pas possible d'employer directement les algorithmes de clustering classiques (Muthukrishnan, 2005).

Les méthodes étudiées exploitent pour la plupart le paradigme incrémental. Originellement étudiées pour traiter les très grands jeux de données, elles consistent à partitionner itérativement l'ensemble des données observées jusqu'à l'instant courant. Une partition finale des données est alors obtenue en fusionnant les résultats partiels correspondants. Ces méthodes sont divisées selon le mode de fusion défini (Damez et al., 2012) : elle peut être progressive quand les résultats précédents sont inclus au problème de clustering à l'état courant (clustering en ligne) (Beringer & Hüllermeier, 2006) ; elle peut être finale et faire l'objet d'un clustering a posteriori, une fois tous les résultats partiels obtenus (clustering hors ligne) (Aggarwal et al., 2003).

Discussion Pour une tâche de clustering sur des flux de données, les données elles-mêmes sont supposées statiques : elles ne subissent aucun changement si ce n'est de perdre progressivement de leur influence sur le partitionnement courant. Au contraire, les données que nous traitons sont dynamiques et leur représentation évolue au cours du temps. Pour de telles données il est nécessaire de réexaminer en continu les regroupements effectués au sein des clusters. Bien que la réévaluation des regroupements effectués par le passé soit la motivation première des méthodes hors ligne, elles n'opèrent pas en continu mais sur requête a posteriori de l'utilisateur.

6.2.2 Clustering sur des graphes dynamiques

Représentation des données Une autre catégorie de travaux pour données dynamiques, actuellement traite de graphes. Les données sont présentées au travers d'un graphe par les relations qui les lient les unes aux autres, si ces relations évoluent au cours du temps il s'agit d'un *graphe de relations dynamiques*. L'exemple le plus couramment traité est celui des réseaux sociaux, les interactions continues entre un ensemble d'utilisateurs sont par exemple représentées par un graphe $G(t)$ de la forme suivante :

$$G(t) = \{S, X(t)\}$$

où S est l'ensemble des sommets et $X(t)$ est une matrice de relations qui décrit les arêtes du graphe à l'état courant.

Objectif A tout moment il est supposé que les sources forment des communautés au travers des relations qui les lient. Le problème du clustering de graphes dynamiques consiste alors à identifier un partitionnement des nœuds composé à tout instant des sous-graphes rendant le plus fidèlement compte de ces communautés. Traditionnellement, les méthodes de clustering appliquées aux graphes sont étudiées dans le cadre de l'analyse des systèmes complexes, aussi un second objectif intimement lié à cette tâche porte sur la cohérence et le suivi de ces communautés dans le temps. Nous relevons finalement que pour cette tâche une attention particulière est apportée à l'analyse des communautés identifiées, les contraintes liées à la très grande quantité de données traitées sont, elles, relâchées.

Méthodes Comme précédemment, dans le cas dynamique les méthodes de partitionnement sur des graphes reposent elles aussi sur des adaptations incrémentales des méthodes

de clustering classique (Gaertler et al., 2006). Selon le paradigme incrémental, un graphe dynamique est ainsi partitionné de manière itérative, à chaque nouveau pas de temps est associé un résultat partiel correspondant.

Cependant pour cette tâche les données évoluent au cours du temps et engendrent des changements brusques au sein des communautés. Aussi est-il nécessaire de réévaluer la cohérence d'une communauté à tout nouvel instant.

Les travaux diffèrent ainsi selon le mode de fusion proposé pour agréger les communautés entre intervalles de temps consécutifs. D'une part les méthodes en ligne consistent à définir et à évaluer à tout pas de temps un certain critère de cohérence (Palla et al., 2007; Liu et al., 2008; Rosvall & Bergstrom, 2010). A ce titre certains auteurs étendent explicitement le problème de partitionnement avec pour objectif second l'identification des communautés les plus similaires à celles préalablement obtenues (Yang et al., 2011; Kawadia & Sreenivasan, 2012). D'autre part les méthodes hors ligne consistent à identifier a posteriori un partitionnement des communautés obtenues sur l'échelle du temps (Mucha1 et al., 2010; Yang et al., 2011).

Discussion Les méthodes de clustering appliquées aux graphes examinent les propriétés du graphe de relations comme par exemple le nombre de liens entrant ou sortant des sommets (Zhou & Liu, 2012). Ainsi lorsque la matrice de relations $X(t)$ est interprétée comme une matrice de similarité mesurée dans l'espace de description originel, selon la mesure définie et l'algorithme employé il n'est pas toujours possible d'extraire des communautés un représentant caractéristique dans l'espace de description originel, or celui-ci est nécessaire pour permettre l'interprétation des données.

De plus, comme nous le présentons au chapitre 8 la nature du problème que nous considérons nécessite une mesure de comparaison adaptée aux données considérées, dont le comportement est guidé par leur partitionnement dans l'espace originel.

6.2.3 Clustering de séries temporelles

Représentation des données Une *série temporelle* décrit les états successifs, à intervalles de temps réguliers, d'une même variable. Pour un groupe de m variables, une donnée est ainsi une matrice $X(t)$ dont la $i^{\text{ème}}$ colonne est une série temporelle qui décrit les états successifs de la $i^{\text{ème}}$ variable :

$$\forall i \in [1..m], X_i(t) = (x_i(t_0), \ldots, x_i(t))$$

Cette matrice correspond par exemple à l'évolution du nombre de citations des m articles les plus reconnus d'un même auteur.

Objectif Pour une tâche de clustering de séries temporelles, les données ne sont pas elles mêmes supposées stationnaires mais leur évolution l'est : l'hypothèse est ainsi faite que les données évoluent de manière caractéristique, autrement dit que les changements observés dans l'espace de description originel sont distribués sur un ensemble de clusters. Ainsi, en reprenant l'exemple précédent les clusters constituent à tout moment des regroupements d'auteurs dont les m articles les plus reconnus ont connu une évolution similaire.

Traditionnellement les méthodes de clustering de séries temporelles sont des méthodes hors ligne : un partitionnement des données est réalisé au dernier instant de la période d'étude, en prenant toutefois compte de l'historique par extraction d'attributs de tendances par exemple.

Méthodes Les méthodes proposées pour le clustering de séries temporelles sont divisées en deux approches (Liao, 2005) : la première consiste à étendre un algorithme de clustering classique avec une mesure de similarité adaptée aux données séquentielles. La seconde consiste à appliquer un algorithme de clustering classique sur des vecteurs de caractéristiques, résumant les propriétés des séries.

Discussion Les méthodes de clustering appliquées aux séries temporelles sont peu étudiées dans un cadre incrémental. Les méthodes proposées visent surtout à la prise en compte de la temporalité lors du partitionnement d'un corpus statique.

Pour les données que nous considérons, en mode incrémental les séries décrivent les positions successives ou encore la trajectoire jusqu'à l'état courant d'une source dans l'espace de description originel. Bien qu'il soit intéressant de tirer profit de cette information précise nous soutenons que le grand nombre de dimensions, caractéristique des représentations textuelles, constitue un obstacle à une telle approche. En effet, dans un cadre incrémental les séries analysées décrivent seulement quelques variables, dans notre cas la complexité du problème de clustering nécessiterait de très larges corpus de données et imposerait des temps de traitement très importants.

Nous proposons plutôt de résumer à tout moment les trajectoires des sources dans l'espace de description originel : une source est un vecteur dont les coordonnées agrègent l'historique de publication. La tâche que nous proposons d'étudier constitue ainsi un clustering incrémental de séries, décrites sur un vecteur de caractéristiques. Pour les résumés que nous considérons, l'historique est par ailleurs ajusté au travers d'une fonction de vieillissement.

6.3 Représentation des sources

Une donnée de notre problème est une source décrite d'après les documents qu'elle publie en continu. Dans un premier temps nous proposons de représenter une source comme un vecteur qui se déplace dans un espace de représentation textuel : à l'état courant ses coordonnées résument l'ensemble des documents produits par le passé. Dans un second temps nous proposons d'ajuster l'importance de son historique en imposant à ses documents un vieillissement paramétré par l'utilisateur. Enfin, en vue de leur partitionnement nous proposons de normaliser les vecteurs ainsi obtenus de sorte qu'ils indiquent une direction de l'espace de description. Lorsque l'ensemble des sources évolue lui-même au cours du temps, nous proposons d'ajuster l'importance des sources qui cessent de publier à nouveau en fonction du temps.

Nous nous plaçons dans un cadre incrémental, aussi l'ensemble de ces traitements doit s'effectuer de manière efficace, de sorte qu'un minimum d'information soit gardé en mémoire entre deux instants. A tout nouvel instant, les nouvelles coordonnées des sources nécessitent une simple mise à jour qui requiert une utilisation de la mémoire linéaire avec la taille du corpus.

6.3.1 Espace de représentation

Nous proposons de représenter une source par un *vecteur de publication* dans l'espace d'entrée des documents. Cet espace contient l'ensemble des descripteurs observés depuis l'origine t_0. Plus précisément, à l'instant $t+1$, l'espace d'entrée $\mathcal{X}(t+1)$ correspond à l'espace d'entrée courant enrichi de $\mathcal{V}(t+1)$, l'ensemble des descripteurs nouvellement

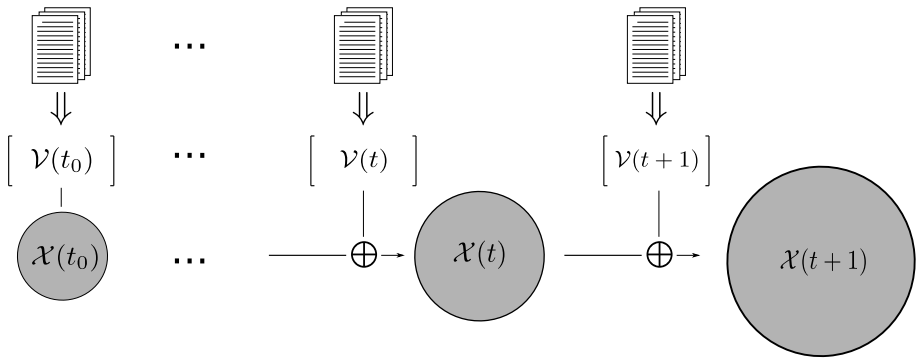


FIGURE 6.1 – Evolution de l'espace de représentation à chaque pas de temps : à l'instant t, l'espace courant est enrichi des termes $\mathcal{V}(t)$ observés dans les documents nouvellement publiés.

observés :

$$\mathcal{X}(t+1) = \mathcal{X}(t) \oplus \mathcal{V}(t+1) \tag{6.1}$$

Comme représenté sur la figure 6.1 les sources sont originellement décrites dans un espace qui correspond à l'ensemble des termes obtenus au premier temps de publication. Ensuite cet espace grandit progressivement, à mesure que de nouveaux documents, utilisant un nouveau vocabulaire et donc de nouveaux termes, sont publiés. Un média comme `http://www.lemonde.fr` qui produit des documents en rapport avec l'actualité enrichit continuellement l'espace de termes spécifiques aux thématiques traitées. Lorsqu'un évènement sportif est en cours par exemple, ce sont entre autres les noms des participants mais aussi l'ensemble du vocabulaire lié à la discipline qui intègrent progressivement l'espace de représentation.

A l'état courant, un partitionnement des sources dans cet espace permet d'extraire un ensemble de représentants. Leurs coordonnées résument alors les termes les plus caractéristiques des sources qu'ils synthétisent. A tout moment, ils apportent ainsi de précieuses indications sur la sémantique associée aux regroupements identifiés.

6.3.2 Vecteur de publication

Pour une source donnée, notons n le nombre de documents produits depuis sa création, décrits sur m dimensions dans $\mathcal{X}(t)$ et notons alors $Z(t)$ la matrice d'entrée, de dimension $n \times m$, qui contient les vecteurs de représentations des documents publiés.

Nous souhaitons former le vecteur de publication $\boldsymbol{x}(t) \in \mathcal{X}(t)$ d'une source comme le point obtenu par agrégation de l'ensemble des documents décrits dans $Z(t)$. Pour ce faire, nous modélisons l'importance des documents par un vecteur de poids $\boldsymbol{\alpha}(t) \in [0,1]^n$. Comme nous le détaillons au paragraphe suivant, ce dernier associe aux documents une importance inversement proportionnelle à leur âge de publication. Dans ce cadre, nous proposons de former le vecteur de publication d'une source à l'état courant comme une combinaison linéaire des documents produits jusqu'alors :

$$\boldsymbol{x}(t) = \sum_{i=1}^{n} \alpha_i(t) \boldsymbol{z}_i \in \mathcal{X}(t) \tag{6.2}$$

où le coefficient $\alpha_i(t) \in [0,1]$ associé au document $\boldsymbol{z}_i$ représente son importance pour la description faite à l'instant t de la source correspondante.

6.3.2.1 Ajustement de l'historique de publication

Une source étant un flux (non stationnaire) qui produit des documents dans le temps, il est nécessaire d'ajuster l'importance de son historique de publication. A mesure que de nouveaux documents sont publiés, l'importance des plus anciens s'estompe ainsi au profit des documents les plus récents.

Pour ce faire nous proposons de manière classique de définir l'importance d'un document au travers d'une fonction de vieillissement exponentielle (Aggarwal et al., 2003). Ainsi à l'état courant en notant $\delta t(i) = t - t_0(i)$ l'âge du document $\boldsymbol{z}_i$ mesuré en intervalles de temps écoulés depuis sa publication, nous proposons de définir $\boldsymbol{\alpha}(t)$ l'importance de l'historique d'une source comme :

$$\boldsymbol{\alpha}(t) = \left(\sigma^{\delta t(i)}, \ldots, \sigma^{\delta t(n)}\right) \in [0,1]^n \tag{6.3}$$

où $\sigma \in [0,1]$ est un paramètre d'ajustement qui contrôle le taux de vieillissement imposé à l'historique. Lorsque $\sigma = 1$, les documents ont une importance uniforme à tout instant, pour $\sigma = 1/2$ les documents perdent 50% de leur importance à chaque pas de temps, enfin à mesure que $\sigma \to 0$, l'historique est à tout moment négligeable par rapport aux documents nouvellement publiés à t.

Viellissement et demi-vie Le vieillissement exponentiel décrit ci-dessus s'exprime également sous la forme équivalente $(1/2)^{-\sigma\delta t(i)} \in [0,1]$. Sous cette forme, le paramètre σ contrôle la durée après laquelle l'historique perd la moitié de son importance initiale. Cette expression est classique dans la littérature (Aggarwal et al., 2004; Leskovec et al., 2009), nous l'employons par exemple au chapitre 9, dans lequel son utilisation semble plus naturelle.

6.3.2.2 Vecteur de publication normalisé

Nous proposons de normaliser les vecteurs de publication des sources de sorte qu'ils soient de norme 1, c'est-à-dire projetés sur l'hypersphère unité de l'espace de représentation. Cette opération est classique pour les représentations textuelles (voir section. 1.1). Elle autorise une comparaison terme à terme des sources à tout instant, sans qu'il soit nécessaire de garder en mémoire davantage que les composantes elles-mêmes des vecteurs.

Ainsi pour une source, son vecteur de publication normalisé décrit une direction à l'état courant dans $\mathcal{X}(t)$, donnée par l'ensemble des documents publiés depuis sa création :

$$\tilde{\boldsymbol{x}}(t) = \frac{\boldsymbol{x}(t)}{\|\boldsymbol{x}(t)\|} \in \mathcal{X}(t) \tag{6.4}$$

6.3.2.3 Règle de mise à jour

Dans un cadre incrémental, les vecteurs de publication sont recalculés à tout pas de temps : à l'état courant ils tiennent alors compte de l'historique de publication mais aussi de l'ensemble des documents nouvellement publiés. En écrivant l'expression normalisée d'un vecteur de publication comme une récurrence sur le temps, nous montrons qu'à l'instant courant ses nouvelles coordonnées peuvent être obtenues de manière efficace, d'après une simple règle de mise à jour.

Etant donné un document $\boldsymbol{z}_i$ dont la date de publication $t_0(i)$ est antérieure à t, nous remarquons dans un premier temps que son importance s'exprime sous la forme récursive suivante : $\alpha_i(t) = \sigma^{t-t_0(i)} = \sigma^{(t-1)-t_0(i)}\sigma = \alpha_i(t-1)\sigma$.

Sans perte de généralité, nous supposons alors que les lignes de $Z(t)$ sont ordonnées de telle sorte que les n_t premières lignes correspondent à l'historique de publication ; et que les $n - n_t$ dernières correspondent aux documents publiés à l'instant t. En notant par ailleurs $\eta(t)$ le terme de normalisation de l'équation (6.4), le vecteur de publication d'une source peut s'exprimer sous la forme récursive suivante :

$$\tilde{\boldsymbol{x}}(t) = \frac{1}{\eta(t)} \left[\sum_{i=1}^{n_t} \alpha_i(t-1)\sigma \boldsymbol{z}_i + \sum_{i=n_t+1}^{n} \boldsymbol{z}_i \right] = \frac{1}{\eta(t)} \left[\boldsymbol{x}(t-1)\sigma + \sum_{i=n_t+1}^{n} \boldsymbol{z}_i \right] \tag{6.5}$$

A l'instant t le vecteur de publication normalisé $\tilde{\boldsymbol{x}}(t)$ est ainsi obtenu comme suit :

1. le vieillissement paramétré par σ est dans un premier temps appliqué au vecteur non normalisé $\boldsymbol{x}(t-1)$.
2. Les coordonnées du vecteur ajusté sont alors ajoutées à celles des documents nouvellement produits.
3. Le résultat est finalement projeté sur l'hypersphère unité.

Entre deux pas de temps consécutifs cette règle de mise à jour nécessite seulement les composantes courantes du vecteur de publication non normalisé. Aussi pour un corpus de sources dynamiques, l'utilisation mémoire est-elle linéaire avec le nombre total de sources étudiées ainsi que le nombre total de descripteurs composant l'espace de représentation.

6.3.2.4 Ensemble dynamique de sources

Jusqu'à présent nous avons supposé que l'ensemble des sources étudiées était fixe et que seule leur description évoluait au cours du temps. Lorsque de nouvelles sources intègrent cet ensemble, ou, de manière équivalente, lorsque des sources cessent de publier, l'ensemble est lui-même dynamique et forme un flux.

Dans ce cas nous proposons, tout comme pour les documents, d'ajuster l'influence de l'historique sur le flux en associant à une source $\boldsymbol{x}$ une représentation $\underline{\tilde{\boldsymbol{x}}}(t)$ pondérée par $\alpha_{\boldsymbol{x}}(t) \in [0, 1]$ telle que :

$$\underline{\tilde{\boldsymbol{x}}}(t) = \alpha_{\boldsymbol{x}}(t)\tilde{\boldsymbol{x}}(t)$$

Initialement une source jouit d'une importance maximale, elle se situe alors sur l'hypersphère unité : $\|\underline{\tilde{\boldsymbol{x}}}(t)\| = \alpha_{\boldsymbol{x}}(t) = 1$. A mesure que son importance décline, sa norme diminue jusqu'à devenir insignifiante : $\|\underline{\tilde{\boldsymbol{x}}}(t)\| = \alpha_{\boldsymbol{x}}(t) \to 0$.

Pour ce faire nous proposons, comme pour les documents, de définir cet importance au travers d'une fonction de vieillissement exponentielle de même paramètre σ : $\alpha_{\boldsymbol{x}}(t) = \sigma^{\delta t(\boldsymbol{x})}$, où $\delta t(\boldsymbol{x}) = t - t_0(\boldsymbol{x})$ est l'âge d'une source mesuré en intervalle de temps. $t_0(\boldsymbol{x})$ peut alors correspondre à la date de création d'une source, ou, comme c'est le cas au chapitre 9, à sa date de publication la plus récente.

Pour chacune des sources d'un ensemble dynamique de sources, la règle de mise à jour décrite dans (6.5) nécessite une nouvelle étape finale qui consiste à appliquer au résultat normalisé son vieillissement correspondant.

6.4 Clustering incrémental pour des sources dynamiques

6.4.1 Formulation du problème

Notons $X(t)$ la matrice de représentation d'un ensemble de sources, représentées dans $\mathcal{X}(t)$ comme décrit à la section précédente. A l'instant t, nous cherchons à obtenir un

partitionnement des sources, composé des K clusters les plus homogènes. Pour alléger la notation, quand faire se peut, nous omettons dans la suite la dépendance des quantités étudiées à la variable t. Plus formellement, à l'instant t, étant donné $\Pi = \{\pi_1, \ldots, \pi_K\}$ un partitionnement de $X(t)$ composé de K communautés et $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$ l'ensemble des représentants associés dans l'espace de description, aussi appelés *centroïdes*, nous proposons de considérer le problème suivant :

$$\max_{(C,\Pi)} F(C, \Pi, X) \quad \text{avec } F(C, \Pi, X) = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in \pi_k} \text{sim}(\boldsymbol{x}, \boldsymbol{c_k}) \tag{6.6}$$

où la fonction objectif F mesure l'homogénéité intra-cluster par une fonction de similarité $\text{sim} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, dont le choix et la définition est central au problème (son choix est notamment discuté au chapitre 9, p. 157). Ainsi, une solution (C^*, Π^*) du problème (6.6) à l'état courant correspond au partitionnement des sources le plus homogène à l'instant t : il s'agit d'un problème de clustering incrémental.

Méthodes Dans le cas statique, une solution à l'état courant pour le problème (6.6) est NP-difficile, les méthodes proposées consistent alors à identifier une solution approchée. L'algorithme des K-moyennes, dont la complexité est linéaire avec la quantité de données, raffine par exemple une partition initiale de manière itérative jusqu'à obtenir une solution satisfaisante.

Dans le cas dynamique, le partitionnement courant évolue à chaque pas de temps. Comme décrit à la section 6.2, les méthodes incrémentales exploitent le partitionnement précédent pour identifier un partitionnement courant. Les approches « purement » en ligne examinent ainsi les données de manière itérative et raffinent ce partitionnement de manière progressive (Bottou & Bengio, 1995). Une autre approche, dite par « mini-lots » (*mini-batch*) (Sculley, 2010) consiste à exploiter l'intégralité des observations faites à l'état courant (Beringer & Hüllermeier, 2006). Au chapitre 9 nous mettons en œuvre une méthode de clustering par mini-lots pour le clustering de sources dynamiques publiant des documents sur Internet.

Influence du paramètre de vieillissement Le paramètre $\sigma \in [0, 1]$ qui ajuste l'influence de l'historique sur la représentation des sources, est une donnée du problème (6.6).

Sa valeur conditionne les données présentées en entrée : pour une valeur proche de 0, les sources sont essentiellement décrites sur les publications les plus récentes. Un fort dynamisme régit à tout moment les partitions. Pour une valeur proche de 1, l'influence des nouveaux documents devient négligeable à mesure que l'historique de publication grandit. Dans ce cadre les sources sont quasi-stationnaires et le problème (6.6) se rapproche d'une tâche de clustering classique.

6.4.2 Dynamisme des partitions

Entre deux pas de temps l'ensemble des documents nouvellement publiés engendre un mouvement dans l'espace d'entrée. A l'état courant, une source apparentée à une communauté peut ainsi lui rester fidèle ou, selon ses changements d'intérêt, transiter vers une communauté nouvelle. Nous observons ainsi deux types d'évolutions : d'une part les communautés dérivent dans l'espace d'entrée au gré des thématiques qui les gouvernent, d'autre part elles absorbent ou perdent des individus selon leurs changements d'intérêt correspondants.

Nous proposons de mesurer ces deux types de dérive en étudiant, dans l'espace d'entrée, les mouvements des représentants associés aux communautés. Au sein des partitions, les transitions effectuées par les sources. Nous définissons alors un *thread d'information* comme une communauté temporelle, qui respecte un certain critère de cohérence dans le temps.

6.4.2.1 Déplacements des communautés

Le représentant d'une communauté, aussi appelé *centroïde*, est le point le plus similaire à ses individus dans l'espace d'entrée. Dans notre cas, la thématique qui gouverne une communauté π_k est résumée au travers des coordonnées de son centroïde $\boldsymbol{c}_k$. Nous proposons de caractériser cette dernière, à l'instant t, d'après les $M \leq m$ directions les plus importantes données par son centroïde $\boldsymbol{c}_k$:

$$\boldsymbol{d}(k,t) = \frac{\boldsymbol{c}_k(t) \circ \boldsymbol{\delta}_k^M(t)}{\|\boldsymbol{c}_k(t) \circ \boldsymbol{\delta}_k^M(t)\|} \tag{6.7}$$

où $\circ$ est le produit de Hadamard et $\boldsymbol{\delta}_k^M(t)$ est un vecteur binaire, indicateur des M composantes les plus importantes du $k^{\text{ième}}$ centroïde à l'instant t, l'importance étant définie par la norme.

En effet les vecteurs de publications des sources étant décrits dans un espace textuel qui évolue au cours du temps, de nombreux termes sont abandonnés puis repris au cours du temps. De nombreuses régions de cet espace représentent alors du bruit pour la description faite des sources, à plus forte raison quand leurs historiques de publication subissent de plus un vieillissement non négligeable. Le paramètre M agit ainsi comme un facteur de grossissement qui contrôle la résolution souhaitée pour décrire les thématiques. Pour de grandes valeurs un certain nombre de termes non pertinents risque cependant d'émerger.

Nous proposons alors de mesurer les déplacements d'une communauté, ou encore le glissement de sa thématique, en mesurant entre intervalles de temps consécutifs l'angle $\theta(k,t)$ formé entre ses directions les plus importantes :

$$\cos\ \theta(k,t) = \boldsymbol{d}_k(t-1)^\intercal \boldsymbol{d}_k(t) \in [0,1] \tag{6.8}$$

Le déplacement $1 - \cos\theta(k,t)$ est alors nul quand les directions données sont égales entre deux pas de temps. Il est maximal et vaut 1 quand ces directions sont orthogonales.

6.4.2.2 Transitions au sein des communautés

Une communauté dont la thématique dérive au cours du temps voit ses sources adopter progressivement un nouveau sujet consensuel, mais peut aussi correspondre à l'absorption ou à l'abandon de sources selon leurs changements d'intérêt correspondants. Les sources transitent effectivement de communauté en communauté au gré de leur déplacements dans l'espace de représentation : leurs regroupements représentés par les communautés évoluent ainsi dans le temps.

La mesure de la stabilité d'une communauté est une tâche classique pour un problème de clustering de graphes dynamiques. Palla et al. (2007) examinent à cet effet la population d'une communauté et mesurent plus précisément son auto-corrélation entre intervalles de temps successifs. Entre deux pas de temps consécutifs, nous proposons plutôt de mesurer la stabilité $\tau(k,t)$ d'une communauté π_k comme la proportion d'individus qui lui est restée fidèle :

$$\tau(k,t) = \frac{|\pi_k(t-1) \cap \pi_k(t)|}{|\pi_k(t-1)|} \in [0,1] \tag{6.9}$$

Le taux de transition $1-\tau(k,t)$ est alors nul quand la population précédente est restée fidèle à l'état courant. Il est maximal et vaut 1 quand un changement complet de population se produit à l'état courant.

6.4.3 Communautés temporelles : threads d'information

Au cours du temps les communautés se déplacent et évoluent au gré des documents nouvellement publiés : prise à deux instants différents, il est possible qu'une communauté ne soit plus gouvernée par la même thématique et/ou qu'elle ne compose plus la même population.

En vue des analyses qui sont réalisées dans le temps sur un partitionnement incrémental, nous pouvons ainsi nous intéresser d'une part aux périodes de temps durant lesquelles une communauté est associée à une thématique précise, d'autre part à celles pendant lesquelles le regroupement qu'elle constitue forme un ensemble de sources solidaires. Durant ces périodes de stabilité, une communauté est en effet associée à une sémantique ou à une population qui demeure cohérente dans le temps. Dans la suite nous définissons un critère de cohérence qui exploite la caractérisation faite des communautés dans l'espace de description : nous privilégions en effet la sémantique et nous déléguons la solidarité des sources aux analyses postérieures.

Nous proposons alors de définir un *thread d'information* comme une communauté temporelle dont l'alignement dans l'espace de description est remarquable durant un intervalle minimum de temps :

$$\exists\ \delta \geq \delta_{\min} \mid \quad \forall t' \in [t \dots t+\delta], \quad \cos\ \theta(k,t') \geq \frac{1}{K}\sum_{l=1}^{K} \cos\ \theta(l,t') \tag{6.10}$$

où $\delta_{\min}$ est un paramètre utilisateur qui ajuste la durée minimale de publications après laquelle une communauté cohérente, dont l'alignement est supérieur à l'alignement moyen des communautés, forme un thread d'information.

6.5 Conclusions

Nous considérons une tâche de partitionnement de sources dynamiques qui publient des documents textuels à intervalles de temps fréquents. Contrairement aux données supposées stationnaires, caractéristiques d'une tâche de clustering traditionnel, les données que nous considérons évoluent au cours du temps et leur partitionnement relève d'un défi particulier.

Nous avons comparé la tâche que nous étudions à trois problèmes classiques dans le domaine de l'analyse de flux de données non stationnaires. Nous proposons de résumer les documents produits à tout instant selon les contraintes de publication imposées par les sources, aussi les contraintes liées à la très grande quantité d'information sont relâchées par rapport à une tâche de clustering sur flux de données. De plus, contrairement à une tâche de clustering sur graphes dynamiques, nous proposons de représenter les sources dans un espace de représentation textuel et ainsi de caractériser les communautés identifiées au travers des thématiques qui les gouvernent. Enfin nous soutenons que la trajectoire complète des sources dans l'espace de représentation est une information riche qui imposerait au problème de partitionnement correspondant de très grandes quantités de données ainsi que d'importants temps de traitements. Aussi la tâche que nous considérons s'apparente-t-elle à une tâche de clustering de séries temporelles, résumées à tout moment par un vecteur de caractéristiques.

Nous proposons de produire le vecteur de publication d'une source comme une combinaison des documents qu'elle a produite depuis sa création. Afin d'ajuster l'importance de l'historique nous modélisons l'effet du temps comme une fonction exponentielle paramétrée. Enfin, de manière classique pour des représentations textuelles nous normalisons ces vecteurs de sorte qu'ils indiquent une direction de l'espace de représentation. Dans un cadre incrémental, nous avons montré que l'ensemble de ces traitements peut être réalisé de manière efficace et requiert une simple règle de mise à jour.

La tâche que nous étudions consiste alors à effectuer un partitionnement incrémental de sources dynamiques : ces dernières sont regroupées en communautés qui reflètent à tout moment des derniers changements observés. Dans le temps le partitionnement ainsi obtenu encourt lui même de nombreux changements. Tandis que les déplacements des sources engendrent des glissements de thématiques pour les communautés correspondantes, leurs transitions au sein de la partition provoquent des renouvellements de populations au sein des communautés. Nous proposons d'observer ces deux types de dérive sur une communauté à chaque nouveau pas de temps. D'une part, nous mesurons l'angle formé entre les directions les plus importantes de son représentant. D'autre part, nous évaluons la proportion d'individus qui lui est resté fidèle. Nous définissons alors un thread d'information comme une communauté temporelle qui respecte un critère de cohérence évalué sur ces deux types de changements.

Chapitre 7

Identification de sources d'information sur Internet

Sur Internet, un média d'information est une source dynamique qui produit des documents en continu en fonction de l'actualité. En vue d'un problème de clustering de sources dynamiques tel que défini au chapitre précédent, l'ensemble des documents publiés par un même média est exploité pour former son vecteur de publication. Néanmoins pour un média comme `www.lemonde.fr` qui couvre aussi bien des affaires politiques que des évènements sportifs par exemple, l'ensemble des documents qu'il publie représente un contenu très hétérogène. Son vecteur de publication couvre alors rapidement l'espace de représentation de manière uniforme, et le partitionnement d'un tel ensemble de sources consisterait rapidement en un ensemble de regroupements arbitraires.

Nous proposons de décomposer un média d'information ou de manière plus générale un domaine de publication en une hiérarchie de sources spécialisées, qui publient des documents de manière plus homogène.

Pour ce faire nous proposons d'exploiter les urls associées au contenu produit sur un même domaine et en particulier d'identifier des regroupements d'urls qui tiennent compte des contraintes de publication de ce contenu.

A la section 7.2 nous présentons d'abord les règles de syntaxe des urls sur Internet, nous proposons ensuite de représenter une url d'après l'ensemble des éléments qui la structure. Nous définissons ainsi une source d'information comme un regroupement d'urls qui répond soit à un certain critère de cohérence soit à un critère de compacité. Motivés par la représentation faite des urls, nous proposons également une mesure évaluant l'homogénéité d'une source, et nous formulons le problème considéré à partir du dendogramme des urls ainsi représentées.

En plus des propriétés de cohérence et de compacité évoquées précédemment, nous considérons un deuxième axe d'étude pour obtenir ce dendogramme : il est soit formé en une fois étant donné l'ensemble des urls disponibles, soit construit de manière incrémentale à mesure que de nouvelles urls sont observées. Nous proposons alors deux algorithmes qui correspondent respectivement à deux extrêmes pour ces deux axes.

Le premier est présenté à la section 7.3, il consiste à réaliser un partitionnement hiérarchique et cohérent des urls disponibles. Pour ce faire nous montrons qu'en imposant un ordre aux éléments structurant les urls il est possible d'interpréter l'arbre préfixe induit sur le corpus comme un dendogramme dont les niveaux sont des partitions cohérentes.

Le second est donné à la section 7.4, il consiste à maintenir, de manière incrémentale, un dendogramme qui respecte un certain critère de compacité dit historique : à tout moment une url nouvellement présentée peut spécialiser une source existante ou en créer

une nouvelle, mais ne peut invalider un partitionnement précédent.

A la section 7.5, nous réalisons une étude comparative expérimentale dans laquelle nous comparons les deux méthodes proposées à deux méthodes de référence. Enfin des travaux similaires sont présentés à la section 7.6, et les conclusions ainsi que les perspectives de ce chapitre sont données à la section 7.7.

7.1 Contexte et motivations

Au chapitre précédent nous avons formulé le problème du partitionnement d'un ensemble de sources dynamiques qui publient des documents à intervalles de temps fréquents. Pour ce problème une source représente une donnée, sur les réseaux sociaux par exemple elle correspond au compte d'un utilisateur qui dépose régulièrement des documents en ligne. Ici, nous considérons le cas de médias d'information qui publient en continu des articles sur Internet.

Problème considéré Sur Internet une source d'information est un domaine, identifié par un *dns*, sur lequel sont déposés des documents accessibles au travers d'une url (Berners-Lee, 2005). Or il apparaît souvent que les documents publiés sur un même domaine sont très hétérogènes. Par exemple le contenu des documents publiés au *dns* `www.lemonde.fr` varie entre autres selon les services proposés (par exemple *blogging*, *forums*, *actualité*), la ligne éditoriale adoptée (par exemple *politique*, *économie*, *internationale*), ou encore l'audience visée. Pour de telles sources, leur partitionnement consisterait alors à effectuer des regroupements arbitraires de données : en reprenant l'exemple ci-dessus, le vecteur de publication (voir section 6.3.2, p. 109) associé à la source `www.lemonde.fr` couvre effectivement l'espace de représentation de manière uniforme.

Sources d'information homogènes Lorsqu'une source publie des documents de manière cohérente, on dit que cette source est homogène. Il est d'usage de décomposer un domaine de publication en un sous-ensemble de sources homogènes en exploitant les règles de structuration de ses urls. Pour ce faire une approche classique consiste à reproduire manuellement la hiérarchie de fichiers fournie au travers de leurs éléments (voir section 7.2.1.1). Un nœud de cette hiérarchie est effectivement un répertoire sur le serveur dans lequel sont déposés des documents de manière cohérente. Un ensemble de sources plus homogènes est ainsi obtenu en définissant une collection de règles manuelles : l'expression régulière `www.lemonde.fr/politique/*` identifie par exemple une source *politique* sur le domaine correspondant. La définition de ces règles pour un domaine nécessite néanmoins des connaissances coûteuses sur sa structuration. De plus elles demandent à être révisées régulièrement de manière à tenir compte des restructurations effectuées au cours du temps.

Notre approche Nous proposons plutôt d'identifier un sous-ensemble de sources homogènes sur un domaine en réalisant un partitionnement hiérarchique de ses urls. Pour ce faire nous proposons de représenter une url d'après l'ensemble des éléments qui la constituent et qui sont décrits dans la *rfc 3986* (Berners-Lee, 2005). Le dendogramme ainsi obtenu est une hiérarchie dont la racine est le domaine lui-même, les nœuds internes sont des sources homogènes et les feuilles correspondent aux urls publiées. Afin de tenir compte des structurations qui évoluent, nous étudions également l'élaboration d'un dendogramme dans un cadre incrémental.

segment	délimiteur	sous-délimiteurs
authority	`://`	`]:/?#@!$&'()*+;[`
path	`/`	`/`
query	`?`	`&=`
fragment	`#`	(non permis)

TABLE 7.1 – Délimiteurs et sous-délimiteurs spécifiés dans la *rfc 3986* pour la structuration des différents segments composant une url.

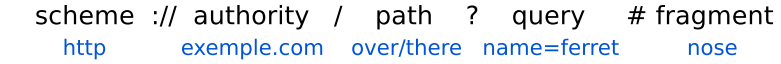


FIGURE 7.1 – Structuration des urls telle que définie dans la *rfc 3986*. Une url est composée de quatre types de segments : l'*authority*, le *path*, le *query* et le *fragment*.

7.2 Caractérisation de sources d'information sur Internet

Dans un premier temps nous décrivons les règles de syntaxe d'une url, puis nous détaillons la représentation que nous faisons de ces dernières. Dans un second temps nous définissons une source d'information comme un ensemble d'urls, caractérisé sur les éléments les plus représentatifs du regroupement qu'elle constitue. Motivé par la représentation faite des urls, nous proposons alors une mesure de l'homogénéité d'une source. Enfin nous décrivons la tâche considérée, à savoir l'identification d'une hiérarchie de sources homogènes sur un domaine de publication.

7.2.1 Représentation des urls

7.2.1.1 Règles de syntaxe

Une url est une liste ordonnée de segments dont la sémantique ainsi que les règles de structuration sont définies dans la *rfc 3986* (Berners-Lee, 2005). Comme représenté sur la figure 7.1, en plus du nom de domaine fourni dans le segment *authority*, une url est structurée selon trois types de segments qui respectent l'ordre suivant : le *path* qui décrit le chemin physique sur le serveur, le *query* qui est composé d'arguments valués représentant des données transmises au serveur, et le *fragment* qui est composé d'arguments binaires précisant par exemple la position demandée dans le document. Le tableau 7.1 présente l'ensemble des délimiteurs ainsi que des sous-délimiteurs réservés pour la structuration des différents segments. Il faut noter que les ambiguïtés possibles entre délimiteurs et sous-délimiteurs sont levées à l'aide de règles contextuelles non présentées ici mais disponibles dans la *rfc 3986*.

Bien qu'une sémantique particulière soit associée à chacun des segments, la ré-écriture d'urls (*url rewriting*) est un procédé répandu qui en autorise une redéfinition. Dans la suite nous supposons ainsi que tous les segments qui composent une url sont nécessaires à l'adressage d'un document publié sur Internet. Nous définissons alors un *token* d'url comme tout élément séparé par un délimiteur. Quand un élément est lui même composé de sous-éléments nous proposons de les concaténer en un token unique. Ces notions sont illustrées dans la section suivante.

7.2.1.2 Représentation

En notant $\mathcal{T}$ l'ensemble des tokens observés sur un domaine et $\mathbb{P}[\mathcal{T}]$ l'ensemble de ses parties, nous proposons de représenter une url u comme l'ensemble (non ordonné) des tokens la constituant :

$$u = \{t_1, ..., t_{|u|}\} \in \mathbb{P}[\mathcal{T}]$$

Par exemple, les urls suivantes :

- `http://dns/xx.php?type=blog&cat=politics#article`
- `http://dns/yy.php?type=news&cat=politics#dossier`

sont respectivement décrites par les ensembles suivants :

- {dns, xx.php, type_blog, cat_politics, article}
- {dns, yy.php, type_news, cat_politics, dossier}

Dans cet exemple les recommandations de la *rfc 3986* ne sont pas respectées puisqu'il semble que les documents publiés au travers de ces deux urls soient déposés au point d'entrée artificiellement construit *dns/cat_politics*. Comme c'est ici le cas, les documents sont décrits par les noms des fichiers physiques sur le serveur. La représentation non ordonnée que nous proposons pour décrire les urls est motivée par des structurations plus exotiques pour lesquelles la sémantique des différents segments aurait été redéfinie.

7.2.1.3 Propriétés

Dans la suite nous faisons l'hypothèse que tout corpus d'urls $\mathcal{U} = \{u_1, \ldots, u_n\}$ publiées sur un domaine *dns*, vérifie les trois propriétés suivantes :

$$\forall\ u, \forall\ t, t' \in u, \quad t \neq t' \tag{7.1}$$

$$\forall\ u, \exists\ t \in u,\ \forall u' \neq u, \quad t \notin u' \tag{7.2}$$

$$\{\text{dns}\} = \bigcap_{u \in \mathcal{U}} u \tag{7.3}$$

qui énoncent que le corpus $\mathcal{U}$ ne contient pas de doublons (7.1), que toute url possède un token qui lui est unique (7.2), et que le nom du domaine de publication est l'unique token commun à l'ensemble des urls (7.3). Si la seconde propriété n'est pas respectée, nous considérons qu'un identifiant artificiel peut être ajouté comme un token supplémentaire à chacune des urls.

7.2.2 Source d'information et homogénéité

7.2.2.1 Définition d'une source d'information

Nous définissons une *source d'information* comme une url, c'est-à-dire un ensemble de tokens $u \in \mathbb{P}[\mathcal{T}]$, à laquelle est publié un ensemble de documents $U = \{u_1, \ldots, u_{|U|}\} \in \mathbb{P}[\mathcal{U}]$. Cette url peut exister sur un domaine, elle peut aussi être construite de manière artificielle : nous proposons de caractériser une source d'après les tokens les plus représentatifs des documents qui lui sont associés. Pour ce faire nous définissons la fonction suivante :

$$S: \quad \mathbb{P}[\mathcal{U}] \to \mathbb{P}[\mathcal{T}] \tag{7.4}$$

$$U \mapsto \bigcap_{u \in U} u \tag{7.5}$$

qui extrait les tokens communs aux urls décrites dans U. On dit alors que la source associée à l'ensemble d'urls U est caractérisée par l'ensemble de tokens $S(U)$. Dans la suite, pour désigner une source, nous faisons aussi bien référence à son ensemble d'urls qu'à l'ensemble de tokens qui la caractérise.

D'après les propriétés (7.3) et (7.2) l'ensemble de toutes les urls publiées sur un domaine est associé à exactement une source $S(\mathcal{U}) = \{\text{dns}\}$, de plus toute url u est associée à une source qui lui est propre $S(\{u\}) = u$.

7.2.2.2 Homogénéité

Nous proposons de définir l'homogénéité $h(U)$ d'une source U de la manière suivante :

$$\begin{aligned} h(U): \quad & \mathbb{P}[\mathcal{U}] \to]0,1] \\ & U \mapsto \frac{1}{|U|} \sum_{u \in U} \frac{|S(U)|}{|u|} \end{aligned} \tag{7.6}$$

qui compare le nombre de tokens de la source $|S(U)|$ au nombre de tokens $|u|$ des urls qu'elle représente, ce nombre étant rapporté au nombre $|U|$ d'urls représentées.

L'homogénéité ainsi définie est maximale et vaut 1 pour une source associée à une unique url, elle est minimale et s'approche de 0 lorsque dans U les urls ne sont pas représentées par les mêmes tokens. L'ensemble de tokens $S(\mathcal{U}) = \{\text{dns}\}$ caractérise par exemple la source la moins homogène qu'il est possible d'identifier sur un domaine.

A partir de l'homogénéité peut être défini $1 - h(u)$, le taux de compression moyen réalisé par $S(U)$ sur ses urls. Cette quantité est maximale sur un domaine, nulle sur une url.

Dans la suite nous souhaitons décomposer $S(\mathcal{U})$ en un ensemble de sources plus homogènes qui réalisent toutes un taux de compression non nul sur les urls qu'elles publient.

7.2.3 Hiérarchie de sources : dendogramme des urls

Objectif Nous souhaitons décomposer un domaine en un sous-ensemble Π de sources (plus) homogènes. Pour ce faire nous supposons qu'il est possible d'identifier un sous-ensemble $U \subset \mathcal{U}$ du corpus associé à une source $S(U)$ plus homogène que $S(\mathcal{U})$. Cette nouvelle source représente alors une spécialisation du domaine dans le sens où U, l'ensemble des urls qu'elle publie, est plus spécifique que $\mathcal{U}$, l'ensemble des urls publiées sur le domaine.

Principe Nous proposons de fractionner l'ensemble des urls publiées sur un domaine de manière récursive. Nous obtenons alors une hiérarchie dont les feuilles sont des urls et les branches sont des regroupements d'urls ordonnés par spécialisation. A tout regroupement d'urls correspondant une source, chacun des niveaux contient un ensemble de sources qui spécialisent celles des niveaux supérieurs.

Méthode Nous proposons ainsi de décomposer un domaine en une hiérarchie de sources en réalisant un partitionnement hiérarchique de ses urls. Autrement dit à partir d'un corpus d'urls, nous formons un arbre étiqueté sur $\mathcal{T}$, dont les feuilles représentent des urls et dont les nœuds internes constituent des sources.

Comme représenté sur la figure 7.2, les sources sont décrites par l'ensemble des tokens rencontrés depuis la racine, la propriété (7.3) impliquant que celle-ci corresponde elle-même à la source $S(\mathcal{U}) = \{\text{dns}\}$. Cet arbre est un dendogramme de largeur $|\mathcal{U}|$, chacun de ses niveaux réalise un partitionnement des urls du corpus. De plus les sources des niveaux

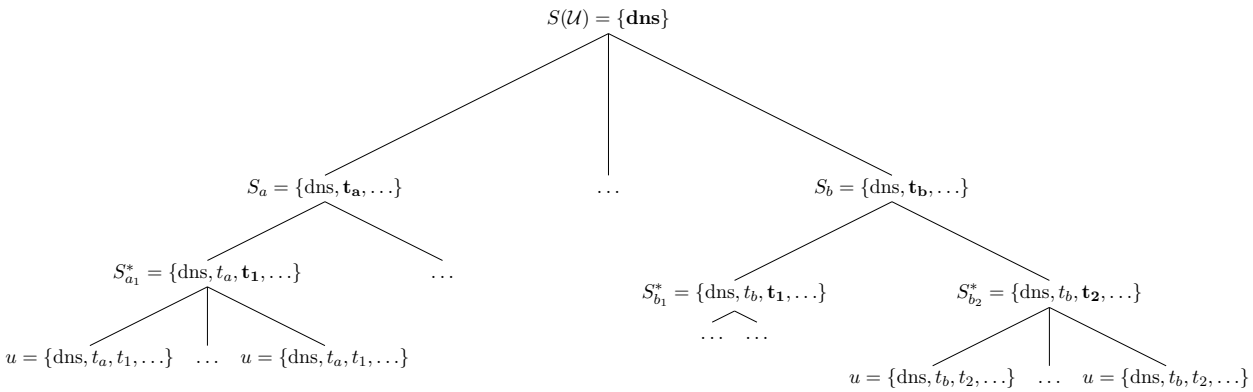


FIGURE 7.2 – Hiérarchie de sources sur le domaine dns à partir d'un partitionnement hiérarchique de ses urls $\mathcal{U}$. Les nœuds de l'arbre sont étiquetés sur $\mathcal{T}$, les feuilles correspondent à chacune des urls de $\mathcal{U}$, les nœuds internes sont des sources caractérisées par les tokens rencontrés en remontant à la racine. Les tokens mis en gras, indiquent des différences avec le nœud parent.

inférieurs sont par définition toutes plus homogènes que celles des niveaux supérieurs. En particulier, les nœuds internes les plus profonds (notés par une étoile sur la figure) identifient les sources qui présentent un taux de compression non nul les plus homogènes.

Dans la suite nous définissons plus précisément la notion de partition sur un corpus d'urls et nous étudions deux types de regroupements hiérarchiques : pour le premier nous imposons aux hiérarchies une notion de cohérence, pour le second, tout niveau de la hiérarchie constitue un regroupement compact d'urls.

7.2.3.1 Définition des partition d'urls

Nous définissons une partition Π sur $\mathcal{U}$ comme un ensemble de regroupements d'urls tel que :

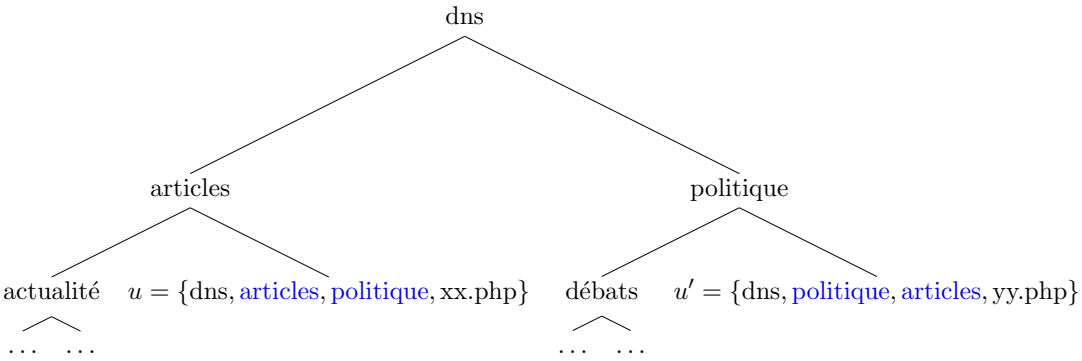
$$\begin{aligned} \forall\, U, U' \in \Pi, &\quad S(U) \neq S(U') \\ \forall\, U \in \Pi, &\quad |U| \geq 1 \\ \forall\, u \in \mathcal{U}, &\quad \exists\, U \in \Pi,\ S(U) \subset u \end{aligned}$$

La source $S(\mathcal{U})$ associée au domaine de publication engendre par exemple une partition de $\mathcal{U}$. Comme représenté sur la figure 7.2, tout partitionnement hiérarchique de $\mathcal{U}$ produit un dendogramme dont les niveaux réalisent différents partitionnements du corpus. Dans la suite les partitions que nous considérons sont toutes issues d'un tel arbre, nous exploitons en particulier la partition composée des nœuds internes les plus profonds afin d'identifier les sources les plus homogènes sur un domaine.

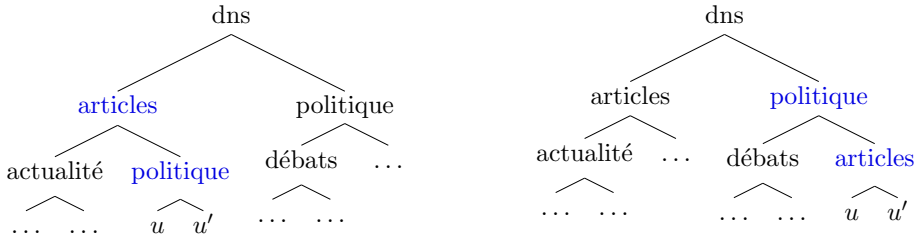
7.2.3.2 Partition cohérente

Nous munissons $\mathbb{P}[\mathcal{U}]$ de la fonction suivante :

$$\begin{aligned} T: \quad & \mathbb{P}[\mathcal{U}] \to \mathbb{P}[\mathcal{T}] \\ & U \mapsto \bigcup_{u \in U} u \end{aligned} \tag{7.7}$$



(a) Partitionnement non cohérent du corpus



(b) Deux partitionnements cohérents du corpus

FIGURE 7.3 – Exemples de partitionnements non cohérents et cohérents

qui extrait d'un ensemble d'urls U l'ensemble de ses tokens $T(U)$. Sur le corpus $\mathcal{U}$, nous avons par exemple $T(\mathcal{U}) = \mathcal{T}$, la totalité des tokens observés sur le domaine. On définit la relation d'ordre dite relation de *cohérence* $\prec$ comme :

$$U \prec U' \qquad \text{ssi} \qquad S(U) \subset T(U')$$

Un hiérarchie de sources H est cohérente si et seulement si, pour tout couple (U, U') de regroupements d'urls associés à des nœuds internes distincts de H, la propriété suivante est vérifiée :

$$U \prec U' \Rightarrow U' \not\prec U \tag{7.8}$$

en particulier les regroupements définissant une partition Π basée sur une hiérarchie cohérente vérifient la propriété attendue, on dit aussi que Π forme une partition cohérente de $\mathcal{U}$.

Sur la figure 7.3(a) nous avons représenté une hiérarchie pour laquelle la partition non cohérente $\{\{dns, articles\}, \{dns, politique\}\}$ viole la propriété (7.8). Si on considère que U regroupe tout les urls sous le nœud *articles* et U' toutes celles sous le nœud *politique*, alors $T(U')$ contient *politique* et *articles*, en particulier $S(U) \subset T(U')$; et inversement $T(U)$ contient également *politique* et *articles*, en particulier $S(U') \subset T(U)$. Comme représenté sur la figure 7.3(a) un réarrangement de la partition consistant à regrouper les urls u et u' rétablit la relation de cohérence, à condition qu'aucune url sous le nœud *politique* (resp. *articles*) ne contienne *articles* (resp. *politique*).

7.2.3.3 Partition compacte

Une partition Π de $\mathcal{U}$ est compacte si et seulement si pour tout u dans $\mathcal{U}$, Π vérifie :

$$u \in U \Rightarrow U = \text{argmax}_{U' \in \Pi} |u \cap S(U')|$$

Une url u est alors affectée au cluster U dont la source $S(U)$ (le représentant du cluster) lui est la plus similaire. La compacité est une propriété classique en vue du partitionnement d'un ensemble de données.

7.3 Identification de sources par lots

Nous réalisons ici un partitionnement cohérent des urls publiées sur un domaine. Dans ce contexte nous proposons d'obtenir la hiérarchie de sources associée à un domaine en exploitant l'intégralité des connaissances disponibles sur ce dernier. Plus précisément en munissant l'ensemble des tokens observés $\mathcal{T}$ d'une relation d'ordre, nous proposons de construire l'arbre préfixe qu'elle induit sur l'ensemble $\mathcal{U}$ des urls publiées sur le domaine. Nous interprétons alors cet arbre comme un dendogramme.

Dans un premier temps nous rappelons la définition d'un arbre préfixe puis nous montrons que toute partition identifiée au travers d'un arbre préfixe est une partition cohérente. Dans un second temps nous étudions l'ordre particulier qui ordonne les tokens selon leur fréquence d'apparition sur le domaine. Nous proposons alors un algorithme d'identification de sources par lots dont la complexité est linéaire avec le nombre d'urls traitées.

7.3.1 Arbre préfixe et partitions cohérentes

Nous munissons $\mathcal{T}$, l'ensemble des tokens observés sur $\mathcal{U}$, d'une relation d'ordre α telle que :

$$\forall t, t' \in \mathcal{T}, \qquad t \leq_{\mathcal{T}}^{\alpha} t' \text{ ou } t' \leq_{\mathcal{T}}^{\alpha} t$$

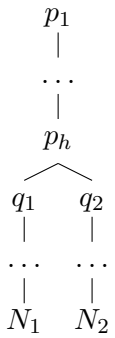
Définition : arbre préfixe Un arbre préfixe, ou *trie d'information*, est une structure de données organisant un ensemble de chaînes en une hiérarchie dans laquelle il existe un nœud pour chaque préfixe commun. Les chaînes sont, elles, stockées dans les feuilles de l'arbre (Black, 2004).

Etant donné un corpus d'urls $\mathcal{U}$ et u une url de ce corpus, nous dénotons par $[u]_\alpha$ la chaîne composée des éléments de u ordonnés selon α et par H_α l'arbre préfixe induit sur le corpus.

Partitions cohérentes à partir d'un arbre préfixe Nous nous intéressons à l'ensemble des partitions identifiables sur le corpus $\mathcal{U}$ à partir de l'arbre préfixe H_α. Comme décrit à la section précédente, à tout nœud interne de la hiérarchie correspond une source qui est associée aux urls de son sous-arbre. En particulier, prises dans leur ensemble les sources d'un même niveau forment une partition d'urls.

Proposition 1 *Pour tout ordre α, soit H_α l'arbre préfixe qu'il induit sur $\mathcal{U}$, alors tout partitionnement Π de $\mathcal{U}$ basée sur H_α est cohérent.*

Preuve Soit α un ordre et soit U_1 et U_2 les regroupements d'urls associés à deux nœuds internes distincts N_1 et N_2 de H_α l'arbre préfixe induit de α sur $\mathcal{U}$. On veut montrer que si $U_1 \prec U_2$ alors $U_2 \not\prec U_1$. Par définition d'un arbre préfixe, H_α est de la forme suivante :



où $p_h \in \mathcal{T}$ est le plus profond token de la hiérarchie commun aux éléments de U_1 et de U_2 : $\{p_1, \ldots, p_h\}$ est l'ensemble des tokens communs aux éléments des deux ensembles d'urls. Comme représenté sur le schéma, nous notons de plus q_1 (resp. q_2) le token commun aux éléments de U_1 (resp. U_2), le moins profond après p_h.

Supposons sans perte de généralité que $q_1 \leq^{\alpha}_{\mathcal{T}} q_2$,

a) par définition de q_1 on a $q_1 \in S(U_1)$,

b) de plus, q_1 n'apparaît dans aucune url de U_2. En effet s'il apparaissait dans $u \in U_2$, $[u]_\alpha$ devrait s'écrire $[p_1, \ldots, p_h, q_2, \ldots, q_1, \ldots]$ ce qui contredit l'ordre α. On a donc $q_1 \notin T(U_2)$

c) q_1 permet donc de montrer que $S(U_1) \not\subset T(U2) \Rightarrow U_1 \not\prec U_2$.

On montre ainsi que $q_1 \leq^{\alpha}_{\mathcal{T}} q_2 \Rightarrow U_1 \not\prec U_2$.

Maintenant si $U_1 \prec U_2$, alors d'après la contraposée du résultat précédent $q_1 \not\leq^{\alpha}_{\mathcal{T}} q_2$.

Or par définition de $\leq^{\alpha}_{\mathcal{T}}$ et comme $q_1 \neq q_2$ puisque $U_1 \neq U_2$, on a $q_2 \leq^{\alpha}_{\mathcal{T}} q_1$.

Donc d'après le résultat précédent, $U_2 \not\prec U_1$.

Donc on a bien si $U_1 \prec U_2$ alors $U_2 \not\prec U_1$ quels que soient U_1 et U_2 les regroupements d'urls associés à deux nœuds distincts d'un arbre préfixe.

De toute relation d'ordre sur $\mathcal{T}$ est ainsi déduit un partitionnement cohérent des urls. Une perspective théorique intéressante consisterait à étudier l'autre sens d'inclusion, à savoir si à tout partitionnement cohérent des urls est associé un ordre sur $\mathcal{T}$.

7.3.2 Tokens fréquents

La relation (1) implique qu'à tout ordre α sur $\mathcal{T}$ correspond un arbre préfixe dont les niveaux constituent des partitions cohérentes d'urls. Nous proposons d'exploiter un ordre particulier pour lequel les sources obtenues sont identifiées au travers des tokens les plus fréquents. Soit l'ordre occ défini comme :

$$\forall t, t' \in \mathcal{T}, \qquad t \leq^{\text{occ}}_{\mathcal{T}} t' \qquad \text{ssi} \qquad \text{occ}_{\mathcal{U}}(t) \leq \text{occ}_{\mathcal{U}}(t') \tag{7.9}$$

où la fonction $\text{occ}_{\mathcal{U}} : \mathcal{T} \to [1..|\mathcal{U}|]$ associe à un token son nombre d'occurences dans le corpus $\mathcal{U}$. L'arbre préfixe H_{occ} induit de la relation occ sur $\mathcal{U}$ est alors un *fp-tree* de support nul, tel qu'employé par l'algorithme *fp-growth* (Han et al., 2000) pour la recherche de motifs fréquents dans un ensemble de données. Parmi les propriétés d'un tel arbre, sa taille mesurée en nombre de nœuds est toujours bornée par $\sum_{t \in \mathcal{T}} \text{occ}_{\mathcal{U}}(t)$ la somme des occurences des tokens observés sur le domaine.

7.3.3 Algorithme proposé

Nous supposons que l'ensemble $\mathcal{U}$ des urls associées à un domaine est connu, nous exploitons alors l'ordre occ afin d'obtenir la hiérarchie de sources H_{occ}. Ses nœuds internes les plus profonds forment une partition d'urls, et constituent le sous-ensemble de sources, dotées d'un taux de compression non nul, les plus homogènes de la hiérarchie obtenue.

Principe L'algorithme que nous proposons calcule dans un premier temps l'ordre occ sur le corpus $\mathcal{U}$ et établit dans un second temps la hiérarchie de sources H_{occ}.

L'algorithme 1 produit la hiérarchie H_{occ} à partir de l'ensemble des urls $\mathcal{U}$ publiées sur un même domaine. Une première étape consiste ainsi à extraire la relation d'ordre occ sur $\mathcal{T}$ et à ordonner les urls selon cette dernière (lignes 4-5). Dans un second temps, l'arbre préfixe H_{occ} est construit par insertion successive des urls (lignes 8-10). Les nœuds internes les plus profonds du trie ainsi retourné (ligne 11) peuvent alors être examinés afin de constituer Π_{occ}. Les fonctions countTokensOccurences, sort ainsi que chaîne sont supposées fournies.

Algorithm 1 *Construction de H_{occ} sur un corpus d'urls $\mathcal{U}$*

```
1:  input : U
2:  output : H_occ
3:  // Définition et application de l'ordre occ sur U
4:  occ ← countTokensOccurences(U)
5:  U ← sort(U, occ)
6:  H_occ ← ∅
7:  // Construction du trie
8:  for u ∈ U do
9:      traversée(H_occ, u)
10: end for
11: return H_occ
12:
13: function traversée(T, u)
14: // Traversée du niveau courant si l'un des nœuds le permet
15: for T_fils ∈ fils(T) do
16:     u ← next(u); t ← label(T)
17:     if u = t then
18:         return traversée (T_fils, u \ u)
19:     end if
20: end for
21: // Insertion d'une chaîne composée des tokens restants après la traversée
22: chaîne(T, u)
23: end function
```

Complexité Munis d'une structure adaptée, l'ordonnancement des tokens ainsi que la définition de occ requièrent $|\mathcal{U}|$ opérations dans le pire cas. La création de l'arbre préfixe H_{occ} nécessite, elle, une nouvelle lecture des urls, ce qui requiert également $|\mathcal{U}|$ opérations dans le pire cas. La complexité dans le pire cas de l'algorithme 1 est donc en $O(|\mathcal{U}|)$.

Motivations pour une approche incrémentale En supposant que $\mathcal{U}$ contienne l'ensemble des urls publiées sur un domaine, il est possible de définir l'ordre occ sur l'ensemble

des tokens observés sur ce dernier. Or en pratique, s'il s'avère que la structuration des urls évolue au cours du temps, il serait nécessaire de redéfinir occ et donc H_{occ}. Pour une telle situation, une approche consiste à relancer l'algorithme 1 à intervalles de temps réguliers (Leung & Khan, 2006). Outre le nombre important de traitements nécessaires à cette approche, les hiérarchies produites à différents intervalles de temps sont susceptibles d'invalider les sources identifiées par le passé. En effet, la relation d'ordre occ étant susceptible d'évoluer, on ne saurait garantir qu'une même url soit toujours associée à une même source. Une approche qui nécessite moins de traitements consiste à maintenir une hiérarchie dynamique au travers d'une relation d'ordre dynamique (Tanbeer et al., 2008), néanmoins cette dernière souffre du même problème d'urls affectées à des sources différentes dans le temps.

En relâchant la contrainte d'ordre sur l'ensemble des tokens, une solution consisterait à définir un ordre partiel, correspondant par exemple à l'ordre d'arrivée des tokens. Dans la suite, nous décrivons plutôt un algorithme incrémental qui maintient une partition compacte respectant l'ordre d'arrivée des urls présentées. Cette méthode garantit la validité des partitions constituées par le passé : en particulier entre deux insertions, une nouvelle source peut être identifiée ou une ancienne source peut être spécialisée mais en aucun cas invalidée.

7.4 Identification incrémentale de sources

Nous supposons dans cette section que le corpus d'urls $\mathcal{U}$ est présenté de manière séquentielle : à l'instant courant nous n'avons pas connaissance des urls observées aux instants ultérieurs. Dans ce cadre nous considérons un certain critère de compacité qui tient compte de l'ordre d'apparition des urls. Nous proposons alors un algorithme incrémental pour l'identification de sources homogènes sur un domaine.

7.4.1 Compacité et incrémentalité

Dans un cadre incrémental les urls sont présentées de manière séquentielle et nous souhaitons maintenir l'ensemble des sources qui partitionne les urls observées de manière la plus compacte possible.

Néanmoins en mode incrémental, l'ordre d'apparition des urls contraint le problème : en supposant en effet qu'une url u soit affectée à une source $S(U)$ et qu'une nouvelle url u' soit présentée à l'état courant. Si u est l'url la plus similaire à u' alors le principe de similarité maximale affecte u' à la source $S(U)$ indépendamment de la compatibilité entre les tokens de u' et de $S(U)$. Ce cas peut être illustré par l'exemple suivant :

$$S(U) = \{\text{dns}, \text{news}, \text{today}\} \text{ et } \begin{cases} u = \{\text{dns}, \text{news}, \text{today}, \text{sport}, \text{volley}, \text{match}, \text{xx.php}\} \\ u' = \{\text{dns}, \text{sport}, \text{volley}, \text{match}, \text{yy.php}\} \end{cases}$$

où l'on suppose que u, affectée à U, est l'url préalablement publiée la plus similaire à u' publiée à la date courante. Maximiser la compacité classique telle que rappelée à la section 7.2.3.3 conduit à associer u' à $S(U)$ ce qui serait contraire à la définition d'une source, en effet on aurait alors $S(U) \not\subset u'$ (voir section 7.2.2), et il deviendrait nécessaire de réduire $S(U)$ à $\{\text{dns}\}$.

Cette situation est directement liée à l'ordre d'apparition des urls : en levant la contrainte de séquentialité, la totalité du corpus serait connue et effectivement les sources seraient d'abord identifiées sur les regroupements les plus compacts. En reprenant l'exemple, l'ensemble $\{\text{dns}, \text{sport}, \text{volley}, \text{match}\}$ caractériserait ainsi une source qui publie à la fois les

urls u et u'. En mode incrémental, en revanche cette source n'est identifiée que lorsque ces deux urls ne peuvent être affectées de manière indépendante à d'autres sources.

Dans la suite nous décrivons un algorithme incrémental qui maintient une partition dotée d'un critère de compacité dit historique : toute nouvelle url présentée est affectée à la source qui représente un compromis entre la compacité et l'ordre d'arrivée des urls.

7.4.2 Algorithme proposé

La méthode que nous proposons consiste à maximiser l'homogénéité des sources après observation de chacune des urls. Ainsi, le dendogramme résultant contient à tout moment les sources les plus compactes qu'il est possible d'identifier étant donné l'ordre d'arrivée des urls présentées.

Principe Nous proposons de former le dendogramme de manière descendante : lors de la traversée d'une url, nous étudions l'intérêt de tout token en cours de descente qui permette d'atteindre le niveau inférieur.

Au niveau courant, le nœud qu'il est possible de traverser et qui mène à la source la plus compacte pour l'url présentée est sélectionné. S'il n'est pas possible de traverser le niveau courant, une nouvelle feuille est insérée au niveau courant.

L'algorithme que nous proposons maintient ainsi une hiérarchie de sources de manière incrémentale, ces dernières peuvent être spécialisées mais en aucun cas invalidées.

L'algorithme 2 décrit notre approche : une hiérarchie H est mise à jour d'après l'ensemble de tokens u. Si H est vide alors la hiérarchie est initialisée avec u (ligne 4), sinon les tokens sont propagés au travers de la hiérarchie (ligne 6) jusqu'à ce qu'une nouvelle source soit identifiée ou qu'une nouvelle feuille soit insérée. Le chemin parcouru dans H est alors celui qui maximise une compacité historique (lignes 12-16). Dans l'algorithme 2, les fonctions *tokens*, *nbTokens*, *nœud*, *fils* et *nouveauFils* sont supposées définies.

Complexité La traversée de la hiérarchie par chacun des $|\mathcal{U}|$ ensembles de tokens traités jusqu'à l'instant courant est bornée par $O(|\mathcal{U}|^2)$ qui donne la complexité de l'algorithme 2. Bien que ce dernier soit plus exigeant en calculs que la méthode par lots proposée à la section précédente, elle permet, dans un contexte temporel, de tenir compte des dernières évolutions observées au sein de la structuration d'un domaine et ce, sans invalider les sources identifiées par le passé.

Par ailleurs, l'algorithme 2 présente une implémentation naïve de la méthode que nous proposons, une implémentation consistant à enregistrer, dès la racine de H, le chemin le plus similaire serait associée à une complexité en moyenne, bien moins élevée.

7.5 Etude comparative expérimentale

Nous décrivons ici, une expérience réalisée sur des données réelles, issues du domaine `www.lemonde.fr`. Nous comparons les sources extraites par la méthode d'identification par lots à celles obtenues par la méthode d'identification incrémentale. Nous étendons cette comparaison à deux méthodes naïves qui jouent ici le rôle de référence. Nous commençons par décrire les données ainsi que les approches étudiées, ensuite nous décrivons les mesures d'évaluation employées pour attester de la qualité des résultats. Enfin, nous discutons des performances de chacune des méthodes.

Algorithm 2 *Mise à jour incrémentale de H après observation de u*

input : $H,\ u$
output : H
if $H = \emptyset$ **then**
  $H \leftarrow$ nœud$(\emptyset)$; chaîne$(H,\ u)$
else
  $(T_{\text{parent}},\ u_{\text{restant}}) \leftarrow$ traversée $(H,\ u)$; chaîne$(T_{\text{parent}},\ u_{\text{restant}})$
end if
return H

function traversée $(T,\ u)$
$h_{\max} \leftarrow 0; T_{\max} \leftarrow \emptyset$
for $T_{\text{match}} \in \{T_{\text{fils}} \in \text{fils}(T) \mid S(T_{\text{fils}}) \subset u\}$ **do**
  $h \leftarrow$ tauxCompression$(T_{\text{match}},\ u)$
  $T_{\max} \leftarrow h_{\max} < h\,?\ T_{\text{match}} : T_{\max}$
  $h_{\max} \leftarrow \max(h_{\max}, h)$
end for
// Insertion d'une chaîne formée des tokens restants
if $T_{\max} = \emptyset$ **then**
  return $(T,\ u)$
end if
// Traversée du niveau courant
return traversée $(T_{\max},\ u \setminus S(T_{\text{fils}}))$

7.5.1 Description des données

Nous avons recueilli un peu plus de 11 000 urls associées à des articles couvrant des thématiques variées, publiées sur le domaine `www.lemonde.fr`. Ces urls ont fait l'objet d'une collecte durant quatre mois auprès de fils de syndication interrogés quotidiennement.

Sur la partie gauche de la figure 7.4 nous avons représenté le nombre d'urls observées chaque jour de la collecte. Nous constatons une baisse de publications systématique en fins de semaine qui correspond aux jours chômés. Par ailleurs sur la partie droite de la figure sont représentées les publications cumulées sur l'ensemble de la période. Nous constatons une croissance linéaire de la taille du corpus et ainsi une répartition équitable des urls présentées sur l'ensemble de la période.

A partir de l'ensemble des urls disponibles, nous formons un corpus séquentiel $\mathcal{U}$ de la manière suivante : les urls sont dans un premier temps représentées comme des ensembles non ordonnés de tokens tels que décrits à la section 7.2. Pour satisfaire les propriétés (7.2) et (7.3), un token artificiel identifiant chacune des urls est ajouté dans un second temps à chacune des descriptions. Enfin, les tokens numériques étant susceptibles de représenter des dates de publication, nous les écartons des ensembles ainsi obtenus.

Les urls sont alors ordonnées par date de publication et présentées de manière séquentielle.

7.5.2 Protocole expérimental

Nous proposons de comparer les partitions de sources obtenues par quatre approches. Les deux premières correspondent à deux méthodes naïves et jouent ici le rôle de référence : pour la méthode *dns* l'unique source composant la hiérarchie est le domaine `www.lemonde.`

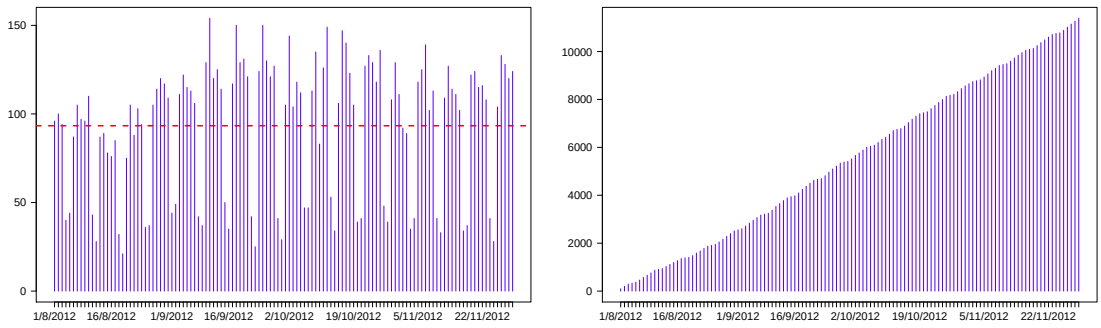
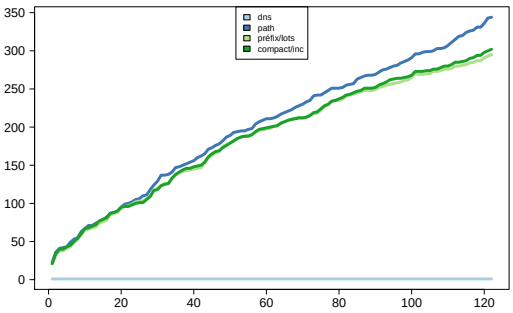
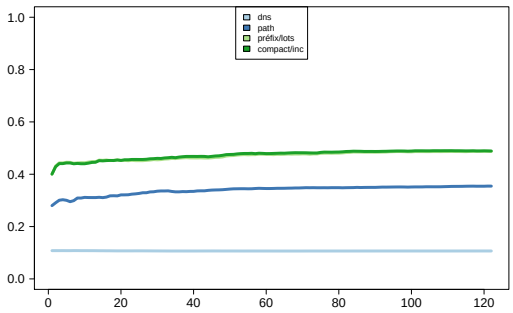


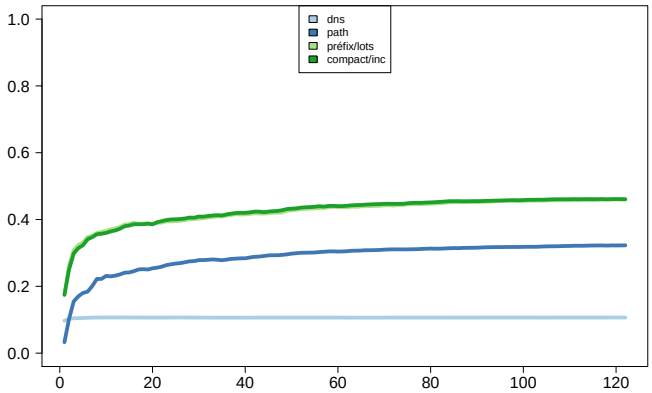
FIGURE 7.4 – Nombre d'urls sur le domaine `www.lemonde.fr` observées quotidiennement durant quatre mois (gauche). Publications cumulées sur la période d'observation (droite)



(a) Evolution du nombre de sources identifiées.



(b) Evolution de l'homogénéité moyenne des sources.



(c) Evolution du score Θ qui mesure un compromis entre les deux mesures de qualité ci-dessus.

FIGURE 7.5 – Mesures de qualité des sources identifiées par quatre méthodes : *dns*, *path*, *préfixe/lots* et *compact/inc*.

`fr`. La méthode *path* est, elle, conforme aux recommandations de la *rfc 3986*. Elle consiste à reproduire la hiérarchie donnée dans le segment *path* des urls présentées. Enfin la méthode *préfixe/lots* identifie un arbre préfixe par lots et la méthode incrémentale *compact/inc* maintient une hiérarchie respectant le critère de compacité dit historique.

Il faut noter que les méthodes proposées sont organisées autour de deux axes : d'une part les urls sont traitées par lots ou de manière incrémentale, d'autre part les partitions identifiées respectent un critère de cohérence ou de compacité. Nous étudions ainsi deux extrêmes pour cette organisation qui sont respectivement les méthodes *préfixe/lots* et *compact/inc*.

Pour chacune des méthodes la partition évaluée est composée des sources qui présentent un taux de compression non nul les plus homogènes. Nous proposons de réaliser une évaluation quotidienne de ces sources, pour la méthode *préfixe/lots* c'est l'intégralité des urls publiées depuis l'origine qui est présentée à toute nouvelle date.

Nous employons les mesures de qualité suivantes pour évaluer une partition : le nombre de sources qui la composent, le score d'homogénéité donné dans (7.6) et réalisé en moyenne par ses sources. Nous mesurons de plus le score suivant qui représente un compromis entre les deux :

$$\Theta(\Pi) = \frac{1}{|\Pi|} \sum_{U \in \Pi} h(U) - \frac{|\Pi|}{|\mathcal{U}|}$$

Les scores obtenus quotidiennement par chacune des méthodes sur ces trois mesures sont reportés sur la figure 7.5.

7.5.3 Résultats et discussion

Pour chacune des partitions, la figure 7.5(a) représente le nombre de sources qui la composent quotidiennement, la figure 7.5(b) décrit l'évolution de son homogénéité moyenne, et la figure 7.5(c) représente l'évolution du score Θ qui réalise un compromis entre ces deux derniers.

Comme nous l'observons sur la figure 7.5(b), la méthode *dns* représente naturellement la partition de sources la moins homogène à tout intervalle de temps. Les trois autres méthodes en revanche produisent des partitions remarquablement plus compactes. Sur le domaine `www.lemonde.fr` nous constatons alors qu'un certain nombre de sources rendent plus fidèlement compte de ses comportements de publication.

Durant les premiers jours de publications, nous observons sur la figure 7.5(a) que chacune des méthodes, hormis l'approche *dns*, maintient un nombre comparable de sources. Durant cette période il semblerait que trop peu de tokens soient connus et qu'aucune des méthodes ne puisse se distinguer : pour chacune des méthodes, cette période semble constituer une phase d'initialisation. C'est au jour 20 que l'on observe une divergence : à partir de cette date la méthode *path* maintient un ensemble de sources toujours plus important que les deux méthodes proposées. Conformément à la *rfc 3986* cette méthode reproduit la hiérarchie donnée dans le segment *path* des urls présentées, il semblerait que l'arbre obtenu soit anormalement grand. En effet sur la figure 7.5(b), nous constatons que les sources produites par la méthode *path* sont en moyenne moins homogènes que celles identifiées au travers des deux méthodes proposées.

Enfin sur la figure 7.5(c), nous constatons que les deux méthodes proposées réalisent les plus grands scores de compromis entre le nombre de sources identifiées et leur homogénéité moyenne. Nous pouvons ainsi conclure que les méthodes correspondantes maintiennent un ensemble de sources de plus grande qualité que ceux proposés par les deux méthodes

de référence. Néanmoins sur les données étudiées il est impossible de distinguer les deux méthodes proposées en terme de performances.

7.6 Travaux similaires

Traditionnellement, les méthodes proposant d'étudier la structuration des urls reposent sur un ensemble de règles manuellement définies, comme Yossef et al. (2009) qui proposent par exemple une collection de règles pour la normalisation d'urls. Bien que les auteurs étudient un ensemble de règles simples pour le problème qu'ils considèrent, pour une tâche d'identification de sources la définition de règles manuelles est un processus à la fois coûteux et permanent quand la structuration des domaines évolue dans le temps.

D'autres travaux proposent d'inférer ces règles à partir d'un corpus d'urls : toujours en vue de leurs normalisation, Lei et al. (2010) proposent par exemple de tirer parti des tokens d'urls et proposent un algorithme identifiant une hiérarchie proche d'un arbre de décision. Leur approche diffère de la nôtre notamment de part la représentation faite des urls : les auteurs proposent de les décrire sur un ensemble de variables correspondants par exemple aux valeurs prises à chacune des positions du *path* ou aux variables présentes dans le segment *query*. Leur méthode nécessite par ailleurs la connaissance de l'ensemble des urls publiées sur un domaine et ne permet pas un traitement séquentiel de ces dernières.

Dans le cadre d'une tâche de compression d'urls en vue de leur stockage, Michel et al. (2000) proposent également d'étudier un arbre préfixe sur un corpus d'urls. Telle que recommandé par la *rfc 3986*, leur approche se limite néanmoins à reproduire la hiérarchie donnée dans le segment *path* des urls traitées. Enfin, pour cette même tâche, Koht-arsa et Sanguanpong (2001) représentent un ensemble d'urls au travers d'un arbre *avl* (Adelson-Velskii & Landis, 1963). Leur méthode aussi se base sur les éléments du segment *path*. De plus les hiérarchies formées sont communes à plusieurs domaines de publication, aussi l'algorithme qu'ils proposent est plus proche d'une méthode de comparaison de chaînes que d'une méthode d'identification de sources sur Internet.

7.7 Conclusions et perspectives

Sur Internet la *rfc 3986* définit une source d'information comme le domaine de publication de ses urls. Or il apparaît souvent que les documents publiés sur un même domaine sont très hétérogènes : ils varient entre autres selon la ligne éditoriale définie, l'audience visée ou encore les thématiques abordées. Dans le cadre d'analyses visant à regrouper un ensemble de sources selon les thématiques qu'elles publient il est ainsi nécessaire de décomposer un domaine en un sous-ensemble de sources, toutes plus homogènes que ce dernier.

En représentant une url comme l'ensemble des tokens qui la composent, nous avons proposé de définir une source comme un regroupement d'urls et de la caractériser d'après l'ensemble de tokens qui représente au mieux ce regroupement. Nous avons ainsi proposé de décomposer un domaine en une hiérarchie de sources en produisant un dendogramme de ses urls.

Pour former ce dendogramme nous avons étudié des partitionnements de deux types : pour le premier les partitions respectent un critère de cohérence, pour le second elles doivent constituer des regroupements compacts. Pour le partitionnement nous avons de plus considéré deux modes de traitement : un mode par lots qui examine dans leur ensemble l'intégralité des urls présentées, et un mode incrémental qui traite les urls de manière

séquentielle. Nous avons alors proposé deux méthodes qui correspondent à deux extrêmes pour ces deux axes d'étude.

En supposant d'une part que l'ensemble des urls publiés est connu, nous avons proposé de définir une relation d'ordre sur l'ensemble des tokens structurant un domaine. Nous avons alors interprété l'arbre préfixe induit de cet ordre sur le corpus comme un dendogramme et nous avons montré qu'il réalise un partitionnement cohérent des urls.

Dans un contexte incrémental d'autre part, nous avons considéré le dendogramme qui correspond à un critère de compacité historique : nous avons proposé de maintenir, sous la contrainte de l'ordre d'apparition des urls, la hiérarchie de sources la plus compacte.

Au travers d'une étude comparative expérimentale réalisée sur un jeux de données réelles, nous avons par ailleurs montré l'intérêt des deux méthodes proposées par rapport à deux approches naïves.

Néanmoins sur les données étudiées il n'est pas possible de distinguer ces deux extrêmes en terme de performances. Aussi serait-il intéressant d'étudier des données qui permette de mettre en évidence les différences entre ces deux extrêmes, et alors de considérer les cas intermédiaires qui consistent respectivement à construire un partitionnement cohérent incrémental et un partitionnement compact par lots.

Chapitre 8

K-moyennes ellipsoïdales pour le clustering de documents

Nous considérons une tâche de clustering de documents décrits dans un espace de représentation textuel. Dans un tel espace, un algorithme de clustering reposant sur une mesure de comparaison classique, tel que l'algorithme des K-moyennes traditionnelles, ne produit pas de résultat satisfaisant. En effet le *fléau de la dimension* mais aussi la nature particulière des descripteurs utilisés font obstacle à la constitution d'une partition homogène et significative. Ainsi que nous le présentons à la section 8.1, lorsque, de plus, le nombre de documents n est largement inférieur au nombre de descripteurs m, les clusters recherchés ne sont pas correctement représentés et l'identification d'un bon partitionnement des données est d'autant plus difficile.

De nombreux travaux portant sur l'algorithme des K-moyennes étudient le problème du partitionnement d'ensembles réduits de données, représentés dans des espaces de très grande dimension. Comme nous le discutons à la section 8.2 les méthodes proposées exploitent les K-moyennes classiques et se heurtent ainsi aux particularités des données textuelles telles que rappelées au chapitre 1.

Nous proposons une extension des K-moyennes sphériques destinées au clustering de données textuelles, pour le cas où le nombre de documents est largement inférieur au nombre de dimensions. Notre proposition repose sur une transformation qui change l'hypersphère unité, exploitée par les K-moyennes sphériques, en un ellipsoïde sur lequel certains descripteurs exhibent une plus grande importance que d'autres. Nous supposons par ailleurs que les données sont situées dans des régions denses de l'espace de représentation et nous proposons de définir un ellipsoïde spécifique à chacun des clusters. A la section 8.3 nous proposons ainsi une nouvelle fonction objectif et nous dérivons les solutions analytiques pour le calcul des centroïdes et des ellipsoïdes associés. Par ailleurs, nous montrons que la complexité de l'algorithme que nous proposons est d'ordre égal à celle de l'algorithme des K-moyennes classiques. Une évaluation expérimentale comparative est présentée à la section 8.4 : les résultats obtenus sur des données réelles et synthétiques montrent la pertinence de la méthode que nous proposons. Enfin, la section 8.5 présente les conclusions de ce chapitre.

Ces travaux ont été publiés dans une conférence (Dzogang et al., 2012b).

8.1 Contexte et motivations

L'algorithme des K-moyennes est une méthode classique en apprentissage non supervisé (Jain, 2010). Il consiste en la recherche d'un partitionnement des données qui minimise la dissimilarité intra-clusters mesurée par la distance euclidienne. Pour ce faire, K clusters sont formés autour de K représentants aussi appelés *centroïdes* ; dans le cas euclidien, du fait de la fonction de coût utilisée, ces représentants sont alors obtenus comme la moyenne arithmétique des données affectées aux clusters.

Lorsque les données sont des documents textuels représentés d'après les termes qui les composent, l'espace de description $\mathcal{X}$ est en très grande dimension et contient de nombreuses régions vides. De plus étant donnée la spécificité des descripteurs textuels (Strehl et al., 2000), l'algorithme des K-moyennes euclidiennes est peu adapté. Pour ce type de données, Dhillon et Modha (2001) ont introduit les K-moyennes sphériques, qui comparent les vecteurs de représentation des documents au travers de leur similarité angulaire et maintiennent les centroïdes sur l'hypersphère unité. Les résultats expérimentaux obtenus montrent l'efficacité de la méthode proposée pour le partitionnement de données représentées dans un tel espace. Pour ce faire les auteurs exploitent des corpus dans lesquels les données présentent un taux de parcimonie en moyenne supérieur à 0.98, la parcimonie étant définie pour un document comme le rapport du nombre de composantes égales à zéro sur le nombre total de descripteurs.

Un autre problème classique pour le clustering de données textuelles vient de ce que le problème de partitionnement posé est souvent mal conditionné pour les jeux de données considérés, c'est notamment le cas lorsque le nombre de documents n est largement inférieur au nombre de descripteurs m. Comme rappelé à la section 1.2.1.2, p. 21, la *sélection de dimensions* est une tâche qui vise à éliminer les dimensions non pertinentes et, ainsi à produire des résultats qui réalisent un meilleur compromis entre le biais et la variance (Hastie et al., 2001). Pour une tâche de clustering, il en résulte alors des partitions plus stables, moins dépendantes du bruit présent dans les données : elle permet de rendre plus robuste un algorithme comme celui des K-moyennes qui dépend d'une étape d'initialisation aléatoire. De plus, en inhibant l'influence des descripteurs non pertinents, la sélection de descripteurs permet à la fois d'obtenir des partitions dont l'interprétation est plus aisée et de pallier le problème du fléau de la dimension. Des exemples de tâches pour lesquelles $n \ll m$ incluent le clustering de données de micro-puces à ADN (Witten & Tibshirani, 2010; Hanczar & Nadif, 2011) ou le clustering de documents en présence de peu de données (Kalogeratos & Likas, 2011). Toujours dans le cadre du texte, un autre exemple est celui du clustering de données qui évoluent au cours du temps. Comme le problème du partitionnement en K clusters de n sources qui publient des documents dans le temps. Les sources sont décrites par les documents qu'elles publient au cours du temps (elles sont par exemple des flux de données, des *blogs*, ou des utilisateurs qui génèrent régulièrement un contenu textuel sur Internet, voir chapitre 7, p. 117). A tout instant, la partition courante est mise à jour d'après les documents nouvellement publiés et l'espace de description (composé de l'ensemble des descripteurs observés jusqu'alors) grandit de manière à tenir compte des descripteurs nouvellement observés (voir section 6.3.2, p. 109). Ces sources sont alors décrites par un vecteur de très grande dimension dans un espace de représentation creux et le problème de clustering correspondant tombe rapidement dans le cas $n \ll m$. Bien entendu pour une tâche de clustering de textes plus classique, il arrive que trop peu de documents soient disponibles. Bien que dans ce chapitre nous ne distinguions pas ces deux tâches, notre proposition est principalement motivée par la première.

Nous proposons une méthode à l'intersection du clustering sphérique et de la sélection

de dimensions, plus précisément, extension des K-moyennes sphériques pour effectuer une sélection de descripteurs. Plus spécifiquement, nous proposons de construire un partitionnement des données sur des ellipsoïdes plutôt que sur l'hypersphère unité. Notre motivation première est d'infléchir la mesure de similarité angulaire vers les descripteurs les plus pertinents : nous caractérisons un ellipsoïde par un vecteur de poids positifs dont l'effet est de dilater ou de contracter les dimensions de l'espace d'entrée selon leur pertinence pour le partitionnement. Bien que ces poids puissent être ajustés manuellement, nous proposons de les obtenir en identifiant les ellipsoïdes qui maximisent la similarité intra-cluster. De plus, nous supposons que les clusters se situent dans des régions denses de l'espace de représentation et nous proposons de définir un ellipsoïde spécifique à chacun.

8.2 Travaux similaires

Pour une tâche d'apprentissage non supervisé, la sélection de descripteurs consiste en la recherche d'un sous espace dans lequel des structures de clusters sont plus facilement identifiables. Comme présenté à la section 1.2, p. 19, la sélection de descripteurs diffère de l'extraction de descripteurs qui consiste, elle, à exploiter la matrice de représentation d'origine X afin d'identifier une nouvelle description des données, condensée sur $l \ll m$ nouveaux descripteurs. Cependant, dans sa forme traditionnelle, l'extraction de descripteur n'a pas pour objectif d'inhiber l'influence des descripteurs les uns par rapport aux autres. De plus les nouveaux descripteurs s'expriment comme des combinaisons des descripteurs originels et nécessitent une étape d'interprétation additionnelle. L'explication des partitions formées peut alors constituer un processus fastidieux.

Comme rappelé à la section 1.2.1.1, p. 20, un exemple classique de sélection de descripteurs dans le cadre des données textuelles consiste en une étape indispensable de prétraitement visant à éliminer les mots vides, rares ou fréquents dans le corpus d'étude. Récemment, Witten et Tibshirani (2010) ont proposé un cadre théorique pour la sélection de descripteurs en apprentissage non supervisé. Pour l'algorithme des K-moyennes en particulier, les auteurs proposent d'étendre la fonction objectif avec une contrainte de type *lasso* : un vecteur de poids positifs sur l'espace d'entrée est assujetti à une contrainte $l1$. Il est alors possible de tirer parti des propriétés de parcimonie offertes par cette norme pour mettre à zéro les descripteurs non pertinents. Bien que ce type de régularisation nécessite généralement des algorithmes de recherche gourmands en calculs, un intérêt certain réside dans ses propriétés de parcimonie.

Jusqu'à présent nous avons évoqué une réduction globale des dimensions de l'espace de représentation $\mathcal{X}$ et nous avons considéré le cas où les données sont toutes projetées dans un nouvel espace $\mathcal{X}'$ de taille réduite. Une approche différente est motivée par l'hypothèse que les clusters se situent dans différentes régions de l'espace d'entrée. Aggarwal et al. (1999) proposent par exemple un algorithme itératif de *projected clustering* : leur méthode consiste à projeter chacun des clusters dans un sous-espace formé des descripteurs les plus corrélés au centroïde correspondant. Cette approche réalisant une recherche exhaustive de l'ensemble des solutions, se pose le problème du passage à l'échelle pour les données décrites sur un très grand nombre de dimensions. Une autre approche consiste à exploiter un vecteur de pondération spécifique à chacun des clusters et, lors de l'optimisation de la fonction objectif, à définir un ensemble de contraintes sur ces poids. L'algorithme *cosa* (Friedman & Meulman, 2004) pénalise par exemple la fonction objectif des K-moyennes par l'entropie associée à chacun des vecteurs de poids. A partir de la solution associée au nouveau problème, les auteurs obtiennent une matrice de similarité calculée dans un espace de taille réduite et réalisent alors un partitionnement hiérarchique des données. De manière

très similaire, les algorithmes *ewkm* (Jing et al., 2007) et *lac* (Domeniconi et al., 2007) constituent explicitement deux extensions à l'algorithme des K-moyennes. Pour chacun, un vecteur de poids est associé à chaque cluster, leurs mesures d'entropie pénalisant la nouvelle fonction objectif.

Néanmoins, ces méthodes reposent toutes sur la distance euclidienne des données, or des travaux ont montré que cette dernière n'est pas adaptée au cas du texte (Strehl et al., 2000). Une approche plus appropriée est proposée par Wang et Domeniconi (2008) : elle consiste à substituer à la distance euclidienne exploitée par l'algorithme *lac*, une fonction noyau qui réalise un produit scalaire dans un espace sémantique latent (voir section 1.1.3.3, p. 18). Cependant cette approche nécessite d'important calculs et, telle que présentée par les auteurs, repose sur l'exploitation de ressources externes. De plus, elle réalise une extraction plutôt qu'une sélection des descripteurs : comme souligné ci-dessus, les données sont représentées dans un espace dense qui inhibe les propriétés de parcimonie de l'espace originel.

Kalogeratos et Likas (2011) exploitent une mesure de similarité angulaire et cherchent des structures locales dans différentes régions de l'espace d'entrée en étudiant les médoïdes [1] associés à chacun des clusters. Cependant l'algorithme qu'ils proposent introduit un certain nombre de paramètres supplémentaires et repose fortement sur leur ajustement manuel. De plus, cet algorithme souffre d'une complexité supérieure à l'algorithme des K-moyennes traditionnelles.

Nous proposons une extension de l'algorithme des K-moyennes sphériques pour effectuer une sélection de descripteurs textuels : notre proposition consiste en l'identification de sous espaces denses sur l'hypersphère unité et s'inscrit dans un cadre similaire aux algorithmes *cosa* et *ewkm*. A notre connaissance, l'unique extension de l'algorithme des K-moyennes sphériques destinée à la sélection de descripteurs est proposée par Modha et Spangler (2003) : les auteurs proposent une extension de l'algorithme des K-moyennes qui autorise en particulier l'emploi de toute fonction de distance convexe. Dans ce cadre, les auteurs effectuent une sélection d'espaces de représentation en introduisant un vecteur de poids positifs qui ajuste l'influence de chacun des espaces originels. Bien que la sélection d'espaces de représentation puisse être vue comme une extension de la sélection de descripteurs, les auteurs ne proposent pas de règles de mise à jour automatique pour dériver ces poids.

8.3 K-moyennes ellipsoïdales

Nous rappelons dans un premier temps l'algorithme des K-moyennes sphériques proposé pour le clustering de données textuelles. Nous décrivons dans un second temps l'algorithme des K-moyennes ellipsoïdales que nous proposons dans le même cadre, pour réaliser une pondération des descripteurs.

8.3.1 Rappel du clustering sur la sphère

L'algorithme des K-moyennes sphériques ou *spherical K-means* (*spkm*) (Dhillon & Modha, 2001) produit une partition des documents formée de K sous-classes π_k, en maximisant la similarité angulaire intra-cluster. Dans ce contexte, le centroïde $\boldsymbol{c}_k$ associé au cluster π_k est le vecteur de l'hypersphère unité, autrement dit de la sphère, qui minimise l'angle

1. Un médoïde est une donnée du corpus étudié, à la différence d'un centroïde qui représente une donnée artificielle. Par exemple, l'algorithme des K-médianes identifient des médoïdes et non des centroïdes.

formé avec l'ensemble des données de π_k. Plus formellement, notons $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$ l'ensemble des centroïdes et $\Pi = \{\pi_1, \ldots, \pi_K\}$ une partition. Soit n documents associés à un vecteur de représentation de norme 1 (i.e. décrits sur l'orthant positif de la sphère), l'algorithme *spkm* maximise la fonction objectif suivante sur les variables (C, Π) :

$$F_{\text{spkm}}(C, \Pi) = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in \pi_k} \boldsymbol{x}^\intercal \boldsymbol{c}_k \quad \text{s.t} \quad \forall k, \|\boldsymbol{c}_k\| = 1 \tag{8.1}$$

Un maximum local de F_{spkm} est atteint en (C^*, Π^*) qui, pour tout k dans l'intervalle $[1..K]$, est défini par :

$$\boldsymbol{c}_k^* = \frac{\bar{\boldsymbol{x}}_{\boldsymbol{k}}}{\|\bar{\boldsymbol{x}}_{\boldsymbol{k}}\|}$$
$$\pi_k^* = \{\boldsymbol{x} |\ k = \text{argmax}_{l=1}^{K}\ \boldsymbol{x}^\intercal \boldsymbol{c}_l\}$$

où $\bar{\boldsymbol{x}}_k = \sum_{\boldsymbol{x} \in \pi_k} \boldsymbol{x}/|\pi_k|$ est équivalent au centroïde produit par l'algorithme des K-moyennes traditionnelles qui repose sur la distance euclidienne. Il faut noter que $\boldsymbol{c}_k^*$ représente la projection de ce dernier sur la sphère et qu'il ne dépend pas explicitement de la taille des clusters $|\pi_k|$.

L'algorithme effectue des itérations successives sur Π et sur C jusqu'à ce qu'une partition satisfaisante soit trouvée, c'est-à-dire lorsqu'une valeur d'homogénéité suffisante est atteinte ou qu'un nombre d'itération maximum a été effectué. Puisque la mesure de similarité angulaire accorde un poids égal à chacune des dimensions de l'espace de description, nous qualifions cet algorithme de *pleinement dimensionnel*.

8.3.2 Principe du clustering sur des ellipsoïdes

Nous faisons l'hypothèse que les clusters existent dans des régions denses de l'espace de représentation, nous souhaitons alors contracter ou dilater les dimensions de cet espace selon leur pertinence pour évaluer la similarité entre les documents. Considérons un vecteur de poids strictement positifs $\boldsymbol{\lambda} \in]0, 1]^m$ tel que $\mathbf{1}^\intercal \boldsymbol{\lambda} = 1$ et dont les composantes attestent de la pertinence de chacune des dimensions.

Notons alors $\tilde{\boldsymbol{x}} = \boldsymbol{\lambda} \circ \boldsymbol{x}$ la version pondérée d'un vecteur $\boldsymbol{x}$ sur la sphère et considérons l'ellipsoïde $\mathcal{E}_{\boldsymbol{\lambda}} = \{\boldsymbol{z} / \|\boldsymbol{\lambda}^{-1} \circ \boldsymbol{z}\| = 1\}$, avec $\circ$ qui représente le produit de Hadamard ou encore le produit par composantes entre ses vecteurs arguments. $\mathcal{E}_{\boldsymbol{\lambda}}$ est centré à l'origine et ses axes correspondent aux axes de l'espace d'entrée. Pour tout $\boldsymbol{x}$ sur la sphère on a alors $\|\boldsymbol{\lambda}^{-1} \circ \tilde{\boldsymbol{x}}\| = \|\boldsymbol{x}\| = 1$ ce qui implique que le projeté $\tilde{\boldsymbol{x}}$ repose sur l'ellipsoïde $\mathcal{E}_{\boldsymbol{\lambda}}$. Ainsi, $\boldsymbol{\lambda}$ définit une transformation qui change la sphère en l'ellipsoïde $\mathcal{E}_{\boldsymbol{\lambda}}$.

Par ailleurs, pour $\boldsymbol{x}$ sur la sphère, considérons sa mesure de similarité avec son projeté $\tilde{\boldsymbol{x}}$, on a $\tilde{\boldsymbol{x}}^\intercal \boldsymbol{x} = \|\tilde{\boldsymbol{x}}\| \|\boldsymbol{x}\| \cos\alpha = \|\tilde{\boldsymbol{x}}\| \cos\alpha$ où α est l'angle formé entre $\boldsymbol{x}$ et $\tilde{\boldsymbol{x}}$. Il en découle que le calcul de $\tilde{\boldsymbol{x}}$ peut être exprimé comme la transformation composée de la mise-à-échelle $\|\tilde{\boldsymbol{x}}\|\mathbf{I}$ suivie de la rotation $\mathbf{R}$ d'angle α, et que l'on peut écrire $\tilde{\boldsymbol{x}} = (\mathbf{R}(\|\tilde{\boldsymbol{x}}\|\mathbf{I}))\boldsymbol{x}$. En particulier, lorsque pour tout j, $\lambda_j = 1/m$ alors $\tilde{\boldsymbol{x}} = \frac{1}{m}\boldsymbol{x}$ et $\mathbf{R}$ est réduit à la matrice identité $\mathbf{I}$. Dans ce cas, $\mathcal{E}_{\boldsymbol{\lambda}}$ est une hypersphère sur laquelle le projeté de $\boldsymbol{x}$ est contracté dans toutes les directions par un facteur $1/m$. Dans le cas général, ce facteur de contraction dépend de la quantité d'information contenue à la fois dans la donnée $\boldsymbol{x}$ et dans le vecteur de poids $\boldsymbol{\lambda}$: lorsque λ_j représente le poids le plus important alors $\boldsymbol{x}$ subit une contraction d'autant plus importante sur la composante j que sa valeur x_j est négligeable par rapport à ses autres composantes. De manière analogue, la rotation qui suit rapproche $\boldsymbol{x}$ d'autant

plus près du $j^{\text{ème}}$ axe de l'espace que sa composante x_j est importante par rapport à ses autres composantes.

8.3.3 Mesure de similarité sur des ellipsoïdes

Tandis que pour des vecteurs reposant sur la sphère le produit scalaire représente une mesure d'angle entre deux vecteurs, sur un ellipsoïde sa sémantique diffère. Soit $\tilde{\boldsymbol{x}}$ et $\tilde{\boldsymbol{z}}$ deux vecteurs sur $\mathcal{E}_{\boldsymbol{\lambda}}$, le produit scalaire classique s'écrit :

$$\tilde{\boldsymbol{x}}^\intercal \tilde{\boldsymbol{z}} = \|\tilde{\boldsymbol{x}}\|\|\tilde{\boldsymbol{z}}\| \cos\tilde{\alpha} \tag{8.2}$$

où $\tilde{\alpha}$ est l'angle formé entre $\tilde{\boldsymbol{x}}$ et $\tilde{\boldsymbol{z}}$. La mesure de similarité entre ces deux vecteurs dépend à la fois de cet angle et de leurs normes respectives : en fixant $\tilde{\boldsymbol{x}}$ et en autorisant $\tilde{\boldsymbol{z}}$ à se déplacer librement sur $\mathcal{E}_{\boldsymbol{\lambda}}$ on observe que plus $\tilde{\alpha}$ est petit, ou plus $\|\tilde{\boldsymbol{z}}\|$ est grand, plus $\tilde{\boldsymbol{z}}$ est similaire à $\tilde{\boldsymbol{x}}$. Par ailleurs la norme des vecteurs étant contrainte sur un ellipsoïde, cette double dépendance exprime un compromis. Ainsi selon la position de ces vecteurs sur $\mathcal{E}_{\boldsymbol{\lambda}}$, leur similarité est plus grandement influencée par leurs normes respectives ou par l'angle qu'ils forment.

Nous proposons d'établir une borne supérieure sur la similarité entre deux vecteurs $\tilde{\boldsymbol{x}}$ et $\tilde{\boldsymbol{z}}$ de l'ellipsoïde $\mathcal{E}_{\boldsymbol{\lambda}^{\frac{1}{2}}}$: on a $\tilde{\boldsymbol{x}}^\intercal\tilde{\boldsymbol{z}} = (\boldsymbol{\lambda}^{\frac{1}{2}} \circ \boldsymbol{x})^\intercal(\boldsymbol{\lambda}^{\frac{1}{2}} \circ \boldsymbol{z}) = (\boldsymbol{\lambda} \circ \boldsymbol{x})^\intercal \boldsymbol{z} \leq \|\boldsymbol{\lambda} \circ \boldsymbol{x}\|\|\boldsymbol{z}\|$ d'après l'inégalité de Cauchy-Schwartz. De plus, puisque $\tilde{\boldsymbol{x}} = \boldsymbol{\lambda}^{\frac{1}{2}} \circ \boldsymbol{x}$ et que $\boldsymbol{z}$ est sur la sphère, i.e. $\|\boldsymbol{z}\| = 1$,

$$\|\boldsymbol{\lambda} \circ \boldsymbol{x}\|\|\boldsymbol{z}\| = \frac{(\boldsymbol{\lambda} \circ \boldsymbol{x})^\intercal(\boldsymbol{\lambda} \circ \boldsymbol{x})}{\|\boldsymbol{\lambda} \circ \boldsymbol{x}\|} = \frac{\tilde{\boldsymbol{x}}^\intercal(\boldsymbol{\lambda}^{\frac{3}{2}} \circ \boldsymbol{x})}{\|\boldsymbol{\lambda} \circ \boldsymbol{x}\|}$$

en notant $\tilde{\boldsymbol{c}} = \boldsymbol{\lambda}^{\frac{1}{2}} \circ \frac{\boldsymbol{\lambda} \circ \boldsymbol{x}}{\|\boldsymbol{\lambda} \circ \boldsymbol{x}\|}$ on obtient $\tilde{\boldsymbol{x}}^\intercal\tilde{\boldsymbol{z}} \leq \tilde{\boldsymbol{x}}^\intercal\tilde{\boldsymbol{c}}$. De manière similaire quand $\tilde{\mathbf{X}}$ décrit n vecteurs sur $\mathcal{E}_{\boldsymbol{\lambda}^{\frac{1}{2}}}$, il s'ensuit que

$$\sum_{\tilde{\boldsymbol{x}} \in \tilde{\mathbf{X}}} \tilde{\boldsymbol{x}}^\intercal\tilde{\boldsymbol{z}} \leq \sum_{\tilde{\boldsymbol{x}} \in \tilde{\mathbf{X}}} \tilde{\boldsymbol{x}}^\intercal\tilde{\boldsymbol{c}}, \qquad \text{avec } \tilde{\boldsymbol{c}} = \boldsymbol{\lambda}^{\frac{1}{2}} \circ \frac{\boldsymbol{\lambda} \circ \bar{\boldsymbol{x}}}{\|\boldsymbol{\lambda} \circ \bar{\boldsymbol{x}}\|}$$

On appel $\tilde{\boldsymbol{c}}$ le *centroïde* associé aux vecteurs de $\tilde{\mathbf{X}}$, induit par le produit scalaire sur un ellipsoïde.

8.3.4 Formulation du problème

Nous proposons d'employer la mesure de similarité sur l'ellipsoïde telle que décrite à la section précédente à la place de la similarité angulaire mesurée sur la sphère. De plus, dans le but d'identifier des structures locales de clusters dans l'espace de description, nous associons à chacun des clusters π_k un ellipsoïde défini par un vecteur de poids positifs $\boldsymbol{\lambda}_k \in [0,1]^m$: nous supposons que cet ellipsoïde repose dans le sous espace défini par l'ensemble des descripteurs pour lesquels dans π_k il existe un vecteur à valeur non nulle. Dans la suite nous notons Λ l'ensemble $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K\}$, ce dernier peut être défini manuellement, selon la tâche considérée et les données étudiées, nous proposons plutôt de le déterminer en maximisant une mesure de similarité intra-cluster.

Puisque, sur une unique dimension, il est en général plus aisé d'obtenir un partitionnement homogène des données, l'ensemble des solutions Λ^* est en général trivial : il est composé de vecteurs de poids dont toutes les composantes, sauf une, sont proches de zéro. Nous proposons d'introduire un paramètre d'ajustement, s, qui s'exprime dans l'intervalle $[0,1[$ et dont l'effet est d'ajuster la forme des ellipsoïdes $\mathcal{E}_{\boldsymbol{\lambda}_k^s}$: lorsque $s = 0$ alors pour

tout k, $\boldsymbol{\lambda}_k^s = \mathbf{1}$ et $\mathcal{E}_{\boldsymbol{\lambda}_k^s}$ devient la sphère. A mesure que s tend vers 1, Λ^* décrit un ensemble d'ellipsoïdes dont le taux d'*aplatissement* est correspondant : dans le cas limite, il s'agit de l'ensemble de solutions trivial précédemment présenté. Ainsi, le problème que nous formulons peut être vu comme une généralisation du cas sphérique.

Soit un ensemble de n vecteurs sur l'orthant positif de la sphère $\boldsymbol{x} \in \mathcal{X}$, un entier K et un réel $s \in [0, 1[$, la fonction objectif à maximiser sur l'ensemble des centroïdes C, l'ensemble des vecteurs de poids Λ et sur le partitionnement Π s'écrit sous la forme suivante :

$$F_{\text{ellkm}}(C, \Lambda, \Pi) = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in \pi_k} \left(\boldsymbol{\lambda}_k^{\frac{s}{2}} \circ \boldsymbol{x}\right)^\intercal \left(\boldsymbol{\lambda}_k^{\frac{s}{2}} \circ \boldsymbol{c}_k\right) \tag{8.3}$$
$$s.t \begin{cases} \forall k, & \mathbf{1}^\intercal \boldsymbol{\lambda}_k = 1 \\ \forall k, & \|\boldsymbol{c}_k\| = 1 \end{cases}$$

Le théorème 1 ci-dessous donne la solution de (8.3). Lorsque s vaut zéro l'équation (8.3) se réduit à l'équation (8.1) et la mesure de similarité induite est la mesure de similarité angulaire, pleinement dimensionnelle. Une plus grande valeur de s réalise une pondération plus déséquilibrée, pour laquelle un poids élevé est accordé aux dimensions qui participent grandement à la similarité entre un centroïde et les données qu'il représente. Puisque les composantes des vecteurs de poids somment à 1, les dimensions non pertinentes sont réduites à des valeurs proches de zéro.

Théorème 1 *F_{ellkm} atteint un maximum local en (C^*, Λ^*, Π^*) défini par :*

$$\boldsymbol{c}_k^* = \frac{\boldsymbol{\lambda}_k^s \circ \bar{\boldsymbol{x}}_k}{\|\boldsymbol{\lambda}_k^s \circ \bar{\boldsymbol{x}}_k\|} \tag{8.4}$$

$$\boldsymbol{\lambda}_k^* = \frac{(\bar{\boldsymbol{x}}_k \circ \boldsymbol{c}_k)^{\frac{1}{1-s}}}{\mathbf{1}^\intercal (\bar{\boldsymbol{x}}_k \circ \boldsymbol{c}_k)^{\frac{1}{1-s}}} \tag{8.5}$$

$$\pi_k^* = \{\boldsymbol{x} | k = argmax_{l=1}^{K} \ (\boldsymbol{\lambda}_l^{\frac{s}{2}} \circ \boldsymbol{x})^\intercal (\boldsymbol{\lambda}_l^{\frac{s}{2}} \circ \boldsymbol{c}_l)\} \tag{8.6}$$

De plus, $\forall s \in [0, 1[$, C^* est un maximum global de F_{ellkm} évalué en C et Λ^* est un maximum global de F_{ellkm} évalué en Λ.

Preuve Soit Λ fixé, il découle de l'expression du centroïde sur un ellipsoïde (voir section 8.3.3) que

$$\tilde{\boldsymbol{c}}_k^* = \boldsymbol{\lambda}^{\frac{s}{2}} \circ \frac{\boldsymbol{\lambda}_k^s \circ \bar{\boldsymbol{x}}_k}{\|\boldsymbol{\lambda}_k^s \circ \bar{\boldsymbol{x}}_k\|} = \boldsymbol{\lambda}^{\frac{s}{2}} \circ \boldsymbol{c}_k^*$$

D'un point de vue analytique, le même résultat est obtenu (preuve omise) en fixant à zéro la dérivée du Lagrangien de (8.3) évalué en C.

De plus, en maintenant C fixé, et en notant γ_k les multiplicateurs de Lagrange associés aux contraintes sur les $\boldsymbol{\lambda}_k$, le Lagrangien de (8.3) évalué en Λ est :

$$L(\Lambda) = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in \pi_k} \sum_{j=1}^{m} \lambda_{kj}^s x_j c_{kj} - \left[\sum_{k=1}^{K} \gamma_k \left(\sum_{j=1}^{m} \lambda_{kj} - 1\right)\right]$$

Les dérivées de L pour la $j^{\text{ème}}$ composante du $k^{\text{ème}}$ cluster ainsi que pour le multiplicateur γ_k sont alors :

$$\frac{\partial L}{\partial \lambda_{kj}} = \sum_{\boldsymbol{x} \in \pi_k} s\lambda_{kj}^{s-1} x_j c_{kj} - \gamma_k \tag{8.7}$$

$$\frac{\partial L}{\partial \gamma_k} = 1 - \sum_{j=1}^{m} \lambda_{kj} \tag{8.8}$$

En mettant (8.7) à zéro, on obtient :

$$\lambda_{kj} = \left(\frac{\gamma_k}{\sum_{\boldsymbol{x} \in \pi_k} x_j c_{kj} s} \right)^{\frac{1}{s-1}}$$

De plus en mettant (8.8) à zéro, il s'ensuit que :

$$\gamma_k = s \left(\sum_{l=1}^{m} \frac{1}{\left(\sum_{\boldsymbol{x} \in \pi_k} x_l c_{kl} \right)^{\frac{1}{(s-1)}}} \right)^{-(s-1)}$$

Finalement on obtient

$$\lambda_{kj} = \frac{1}{\sum_{l=1}^{m} \left(\frac{\sum_{\boldsymbol{x} \in \pi_k} x_j c_{kj}}{\sum_{\boldsymbol{x} \in \pi_k} x_l c_{kl}} \right)^{\frac{1}{s-1}}}$$

On observe que $\boldsymbol{\lambda}_k$ s'exprime comme le produit de Hadamard entre $\bar{\boldsymbol{x}}$ et $\boldsymbol{c}_k$, mis à la puissance $1/(1-s)$ et normalisé par leur produit scalaire, ce que l'on réécrit comme l'expression donnée dans l'équation (8.5).

Finalement, L étant concave en Λ pour tout s dans $[0, 1]$ et Λ^* n'étant pas défini pour $s = 1$, la solution est garantie optimale pour tout s dans l'intervalle $[0, 1[$. Par ailleurs, à mesure que s tend vers 1, $\boldsymbol{\lambda}_k^*$ dégénère en une solution triviale pour laquelle toutes les composantes sauf une sont proches de zéro.

8.3.5 Algorithme proposé

L'algorithme 3 présente l'algorithme des K-moyennes ellipsoïdales que nous proposons. Une étape supplémentaire est ajoutée à l'algorithme des K-moyennes sphériques pour calculer les ellipsoïdes sur lesquels sont projetés les centroïdes à l'état courant (ligne 14). A l'initialisation, les centroïdes sont tirés aléatoirement et les ellipsoïdes sont définis comme des hypersphères de rayon $1/m$. Les formules de mise à jour sont fournies dans les équations (8.4-8.6) du théorème 1. L'algorithme s'arrête lorsque le nombre d'itérations maximum est atteint ou lorsque la partition courante est suffisamment stable ce qui est traduit par un gain de valeur objectif inférieur à un seuil prédéfini η (pour les expériences que nous présentons à la section 8.4, η est fixé empiriquement à 10^{-8}).

Algorithm 3 *K-moyennes ellipsoïdales*

1: *input :* $\mathbf{X}, K, s$
2: *output :* (Π, C)
3: $C_0 \leftarrow K$ centroïdes aléatoires
4: $\Lambda_0 \leftarrow K$ vecteurs de poids uniformes fixés à $(\frac{1}{m}, \ldots, \frac{1}{m})$
5: $\Pi_0 \leftarrow$ partition aléatoire
6: $t \leftarrow 0$
7: **while** $t < \max T$ **and** $\Delta F > \eta$ **do**
8: **for all** $\boldsymbol{x} \in \mathcal{X}$ **do**
9: $l = \text{argmax}_{k=1}^{K} \left(\boldsymbol{\lambda}_k^{\frac{s}{2}} \circ \boldsymbol{x}\right)^{\mathsf{T}} \left(\boldsymbol{\lambda}_k^{\frac{s}{2}} \circ \boldsymbol{c}_k\right)$
10: mise à jour de π_l avec $\boldsymbol{x}$
11: **end for**
12: **for all** $k \in [1..K]$ **do**
13: mise à jour de $\boldsymbol{c}_k$ selon l'éq. (8.4)
14: mise à jour de $\boldsymbol{\lambda}_k$ selon l'éq. (8.5)
15: **end for**
16: $C_{t+1} \leftarrow \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$
17: $\Lambda_{t+1} \leftarrow \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K\}$
18: $\Pi_{t+1} \leftarrow \{\pi_1, \ldots, \pi_K\}$
19: $\Delta F = F_{\text{ellkm}}(C_{t+1}, \Lambda_{t+1}, \Pi_{t+1}) - F_{\text{ellkm}}(C_t, \Lambda_t, \Pi_t)$
20: $t \leftarrow t + 1$
21: **end while**

Convergence Nous montrons la convergence de l'algorithme 3 en montrant d'abord que chacune des étapes de l'algorithme augmente la valeur de F_{ellkm}.

Soit F_t la valeur de la fonction objectif F_{ellkm} à l'étape t de l'algorithme 3. Notons alors Π_t la partition, C_t l'ensemble des centroïdes et Λ_t l'ensemble des vecteurs de poids à l'étape t. D'après le théorème 1,

$$\begin{aligned} F_t = F_{\text{ellkm}}(C_t, \Lambda_t, \Pi_t) &\leq F_{\text{ellkm}}(C_t^*, \Lambda_t^*, \Pi_t) \\ &= F_{\text{ellkm}}(C_{t+1}, \Lambda_{t+1}, \Pi_t) \end{aligned}$$

De plus, par définition de Π^* nous avons,

$$\begin{aligned} F_{\text{ellkm}}(C_{t+1}, \Lambda_{t+1}, \Pi_t) &\leq F_{\text{ellkm}}(C_{t+1}, \Lambda_{t+1}, \Pi_t^*) \\ &= F_{\text{ellkm}}(C_{t+1}, \Lambda_{t+1}, \Pi_{t+1}) = F_{t+1} \end{aligned}$$

Ainsi on a bien,

$$F_t \leq F_{t+1} \tag{8.9}$$

En observant finalement que F_{ellkm} possède une borne supérieure positive et constante, on remarque que ΔF tend vers zéro à mesure que $\max_T$ tend vers l'infini, ce qui montre la convergence de l'algorithme.

Complexité Pour l'algorithme 3, l'étape de réaffectation (ligne 9) qui conduit à la mise à jour de Π (ligne 10) nécessite le calcul de la matrice de similarité entre les données et les centroïdes, cette étape effectue nKm opérations. Le calcul des nouveaux centroïdes et des nouveaux ellipsoïdes requiert nm opérations. Dans le pire cas, l'algorithme effectue $\max T$ itérations.

La complexité globale de l'algorithme est donc $O(nKm \times \max_T)$ ce qui est d'ordre égal à la complexité de l'algorithme des K-moyennes traditionnelles.

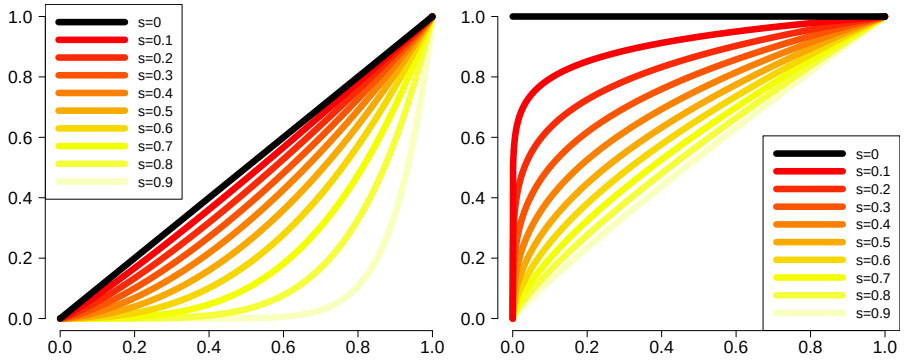


FIGURE 8.1 – Influence du paramètre s sur les partitions produites par les K-moyennes ellipsoïdales. L'influence du paramètre sur la forme des ellipsoïdes est représentée par la grandeur $(\bar{\boldsymbol{x}} \circ \boldsymbol{c})_j^{\frac{1}{1-s}}$ en fonction de $(\bar{\boldsymbol{x}} \circ \boldsymbol{c})_j$ (gauche). L'influence du paramètre sur les coordonnées des centroïdes est représentée par la quantité $\boldsymbol{\lambda}_j^s$ en fonction de $\boldsymbol{\lambda}_j$.

8.3.6 Paramètre de parcimonie s

Dans la section 8.3, nous avons introduit un paramètre d'ajustement s pour contrôler la forme des ellipsoïdes. Dans un premier temps, nous étudions le rôle de ce paramètre pour l'algorithme des K-moyennes ellipsoïdales. Dans un second temps, nous proposons une procédure de sélection pour son ajustement automatique.

8.3.6.1 Sémantique du paramètre

Nous étudions le rôle du paramètre s dans l'algorithme des K-moyennes ellipsoïdales, pour ce faire nous observons son influence sur les ellipsoïdes et les centroïdes dont les expressions respectives sont données dans les équations (8.4) et (8.5). Sur la figure 8.1 est représenté son influence durant l'étape de mise à jour. La partie gauche de la figure représente le numérateur de l'expression (8.5) sur la composante j, en fonction de $(\bar{\boldsymbol{x}} \circ \boldsymbol{c})_j$ et décrit l'influence de s sur les ellipsoïdes. Nous observons que pour des valeurs de s proches de 1, les petites valeurs de $(\bar{\boldsymbol{x}} \circ \boldsymbol{c})_j$ contribuent peu à la mise à jour des ellipsoïdes : ils exhibent un aplatissement plus important sur les composantes correspondantes, et ce de manière proportionnelle à s. Pour l'expression (8.4), la partie droite de la figure représente $\boldsymbol{\lambda}_j^s$ en fonction de $\boldsymbol{\lambda}_j$ et décrit l'influence de s sur les centroïdes. Lorsque s vaut 0, toutes les composantes reçoivent un poids égal et les centroïdes peuvent s'y déplacer librement. En remarquant que les composantes de $\boldsymbol{\lambda}$ somment à 1 et s'expriment de manière relative, on observe qu'à mesure que s tend vers 1, les déplacements des centroïdes se limitent aux seules composantes auxquelles un poids élevé est accordé. Ces dernières représentent les axes sur lesquels l'ellipsoïde correspondant présente un aplatissement minimal.

Ainsi, pour une valeur de s proche de 0, il est moins tenu compte du vecteur de pondération $\boldsymbol{\lambda}_k$ et l'ellipsoïde associé approche une forme sphérique. Pour une valeur de s proche de 1, les ellipsoïdes approchent une forme rectiligne, autrement dit $\boldsymbol{\lambda}_k$ est un vecteur dont toutes les composantes sont proches de zéro sauf une : il s'agit de l'axe sur lequel la similarité entre un centroïde et les données qu'il représente est maximale. C'est pourquoi, lorsqu'il est préférable de tenir compte d'un grand nombre de descripteurs, une

valeur de s proche de zéro est appropriée ; dans les autres cas, lorsque les clusters se situent dans des régions extrêmement locales et denses de l'espace de description, de plus grandes valeurs de s sont préférables.

Il faut par ailleurs noter que pour l'algorithme 3, de grandes valeurs de s tendent à produire de brusques changements à chacune des itérations tandis que de petites valeurs tendent à produire des changements plus progressifs. Par conséquent, s influence également la sensibilité des partitions en phase d'apprentissage.

8.3.6.2 Procédure de sélection automatique

Nous décrivons une procédure générale pour l'ajustement de s qui repose sur le principe du gap statistique que nous rappelons ci-dessous.

Pour l'ajustement du paramètre s, on pourrait envisager à priori une méthode naïve consistant à évaluer la similarité intra-cluster pour différentes valeurs et à retenir celle pour laquelle les partitions produites sont les plus homogènes. Cependant sur l'ellipsoïde $\mathcal{E}_{\boldsymbol{\lambda}^{\frac{s}{2}}}$, la mesure de similarité dépend directement du paramètre s. En conséquence, modifier s revient à modifier la fonction objectif et il devient impossible de comparer la qualité des partitions produites pour différentes valeurs de s.

Méthode du gap statistique Inspiré par la *méthode du coude* (*elbow method*) (Ketchen & Shook, 1996), la méthode du gap statistique (Tibshirani et al., 2001) a été proposée pour estimer le nombre K de clusters structurant un jeu de données $\mathbf{X}$. Cette procédure consiste à normaliser la fonction objectif pour éliminer l'effet de l'influence de K sur la mesure d'homogénéité des partitions.

Witten et Tibshirani (2010) proposent d'étendre cette procédure pour estimer un paramètre continu que nous avons ici noté s puisque nous l'employons à cet effet. Soit $\mathbf{X}_{b\in[1..B]}$, B variantes aléatoires de $\mathbf{X}$, par exemple obtenues en permutant l'ensemble des données sur chacune des dimensions de manière répétée. On note F l'évaluation de la fonction objectif, paramétrée par s, sur le jeu de données $\mathbf{X}$ et F_b son évaluation sur les B variantes aléatoires $\mathbf{X}_{b\in[1..B]}$. La mesure de gap associée à s est alors définie comme :

$$\text{gap}(s) = \log F - \frac{1}{B}\sum_{b=1}^{B} \log F_b$$

La mesure de gap associée à tout s d'un ensemble $\mathcal{S}$, défini par l'utilisateur, est dans un premier temps calculée. Dans un second temps, la valeur $s_0^* = \text{argmax}_{s\in\mathcal{S}}\text{gap}(s)$ est retenue comme le meilleur candidat pour le jeu de données $\mathbf{X}$.

Principe de la méthode proposée Pour une méthode de clustering qui dépend d'une étape d'initialisation aléatoire comme l'algorithme des K-moyennes, il est important, lorsque $n \ll m$, d'attester de la qualité des partitions sur un ensemble répété de simulations. A cet effet, Witten et Tibshirani (2010) proposent d'obtenir s^* en agrégeant l'ensemble des candidats s_i^* associés à différents lancements de l'algorithme $i \in [1..N]$ par vote majoritaire. D'une manière différente, nous proposons de définir l'heuristique suivante pour le choix de s^* :

$$s^* = \text{argmax}_{s\in\mathcal{S}} \sum_{i=1}^{N} \text{gap}_i(s) - \tau_s \tag{8.10}$$

où τ_s est l'écart-type associé à $(\text{gap}_i(s))_{i\in[1..N]}$ et $\text{gap}_i(s)$ est la mesure de gap associée à s pour la $i^{\text{ème}}$ simulation. Cette heuristique est motivée par l'observation que pour N simulations, le choix de s^* est moins sensible aux valeurs extrêmes de $\text{gap}_i(s)$ qu'un vote majoritaire réalisé sur les s_i^*.

Il faut par ailleurs noter que la méthode du gap repose sur l'hypothèse que les clusters sont bien séparés dans l'espace de description (Tibshirani et al., 2001). Lorsque les clusters n'exhibent pas un tel comportement, par exemple en présence de bruit dans les données ou lorsque la partition recherchée n'est tout simplement pas bien représentée dans $\mathcal{X}$, cette procédure est employée dans les K-moyennes ellipsoïdales afin d'identifier les sous espaces sur lesquels le partitionnement du jeu de données $\mathbf{X}$ est le plus homogène comparé aux partitionnements réalisés sur les variantes aléatoires $\mathbf{X}_{b\in[1..B]}$.

L'heuristique que nous proposons requiert de plus que ce partitionnement soit le plus stable. Si d'après l'équation (8.10), c'est $s^* = 0$ qui est retenu, cela signifie qu'aucun des sous espaces considérés ne permet de réaliser un meilleur partitionnement que l'espace de description originel.

Procédure proposée La procédure complète pour la sélection de s est la suivante :

- B jeux de données $\mathbf{X}_b$ de référence sont générés en permutant aléatoirement les composantes de chacune des données dans $\mathbf{X}$. Pour un corpus de documents, cela revient à interchanger aléatoirement tous les mots pour chacun des documents et ainsi à défaire tout structure de cluster susceptible d'exister.
- Pour chacun des $s \in \mathcal{S}$, incluant le cas sphérique ($s = 0$), la fonction objectif est évaluée sur le jeu de données d'étude ainsi que sur les B jeux de données de référence. Cette étape est répétée pour N simulations correspondant chacune à une initialisation aléatoire différente.
- Pour chacun des $s \in \mathcal{S}$ la mesure de gap est évaluée pour chacune des simulations.
- s^* qui satisfait l'heuristique définie dans l'équation (8.10) est retenu.

En tenant compte de la complexité de l'algorithme des K-moyennes ellipsoïdales, la complexité de cette procédure est en $O(N|\mathcal{S}|(B+1) \times nKm \times \max T)$ où $|\mathcal{S}|$ est le nombre de valeurs testées pour l'ajustement de s et N est le nombre de simulations effectuées.

8.4 Evaluation comparative expérimentale

Cette section présente l'évaluation expérimentale réalisée de l'algorithme proposé. Dans un premier temps, nous comparons les partitions obtenues d'après la mesure de similarité sur un ellipsoïde avec le partitionnement réalisé par son homologue pleinement dimensionnel sur une sphère. Pour ce faire nous considérons spécifiquement le cas $n \ll m$ et nous évaluons les performances réalisées par *ellkm* et *spkm* sur un corpus de données synthétiques. Dans un deuxième temps, nous confrontons l'algorithme proposé à quatre autres méthodes issues de l'état de l'art. Nous exploitons alors un corpus de données réelles, le classique *20-newsgroup*, à partir duquel nous constituons différents jeux de données. Ces derniers sont divisés en deux catégories : la première correspond au cas où $n \ll m$, la seconde constitue un cadre plus classique dans lequel un nombre plus grand de documents sont disponibles.

8.4.1 Données synthétiques

Nous comparons les performances de *spkm* et de *ellkm*, en générant, de manière artificielle, un corpus de données selon une partition fixée. Dans un tel cadre, nous réalisons

id.	n	m	n / m	parcimonie
1	30	1088	0.027	0.98
2	60	1736	0.035	0.98
3	90	2239	0.04	0.98
4	120	2514	0.047	0.98
5	150	2608	0.057	0.98
6	180	2756	0.065	0.98
7	210	2799	0.075	0.98
8	240	2879	0.083	0.98
9	270	2923	0.092	0.98
10	300	2941	0.1	0.98

TABLE 8.1 – Description des 10 jeux de données synthétiques.

une analyse des caractéristiques des deux algorithmes étudiés.

8.4.1.1 Génération des données

Les données synthétiques simulent un corpus de documents décrits sur $M = 3000$ mots et pour lequel un schéma de pondération à valeurs dans l'intervalle $[0, 1]$ est utilisé. Ces données sont générées de manière à ce que les vecteurs de représentation correspondants exhibent un taux de parcimonie moyen de 0.98. Les documents du corpus sont équitablement répartis sur $K = 3$ clusters. Nous définissons de plus un vocabulaire spécifique à chacun des clusters π_k en réservant un ensemble de 100 mots spécifiques à chacun. Pour ce corpus, le processus de génération d'un document $\boldsymbol{x}$ est le suivant :

- Un facteur de parcimonie q est tiré aléatoirement à partir d'une distribution normale centrée en 0.98 et d'écart-type 10^{-2}.
- Le document $\boldsymbol{x}$ contient ainsi $p = M(1 - q)$ composantes non nulles, tirées uniformément dans $[\epsilon, 1]$, où ϵ est une constante positive proche de zéro, ici fixée à 10^{-3}. 40% de ces composantes sont échantillonnées uniformément parmi l'ensemble des descripteurs spécifiques à π_k, le reste correspond à des descripteurs non spécifiques à π_k et représente 60% de bruit dans les données.

Jeux de données Dix jeux de données sont ainsi générés en variant le nombre n de documents dans l'intervalle $[30, 300]$. Le nombre m de dimensions de l'espace de représentation associé à chacun des jeux de données est obtenu comme le nombre de composantes pour lesquelles il existe au moins un document à valeur non nulle. De petites valeurs de n correspondent au cas $n \ll m$, pour lequel les composantes spécifiques à chacun des clusters sont minoritaires dans l'espace de représentation. A mesure que n croît, le rapport entre le nombre de documents n et le nombre de dimensions m augmente et les clusters sont mieux séparés dans l'espace de représentation. Le tableau 8.1 présente entre autres la quantité m et le rapport n/m pour chacun des dix jeux de données.

8.4.1.2 Protocole expérimental

Nous comparons l'algorithme des K-moyennes ellipsoïdales (*ellkm*) avec celui des K-moyennes sphériques (*spkm*). Pour chacun des jeux de données, différentes valeurs du paramètre d'ajustement des ellipsoïdes sont évaluées. De plus la procédure de sélection automatique décrite à la section 8.3.6 est également évaluée. Pour ce faire $B = 10$ jeux de

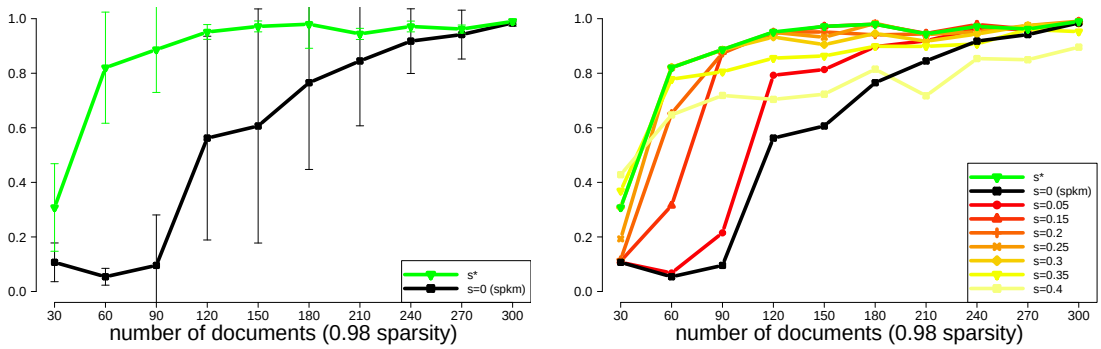


FIGURE 8.2 – Scores moyens de *nmi* associés au cas sphérique et à la procédure de sélection automatique sur 20 lancements réalisés pour 8 jeux de données. L'axe des abscisses représente le nombre n de documents qui composent les jeux de données. Les écarts-types sont donnés dans le cas sphérique et pour la procédure de sélection automatique (gauche), différentes valeurs de s sont de plus évaluées dans l'intervalle $[0.05, 0.4]$ (droite).

données de référence sont considérés et 10 valeurs différentes de s sont testées dans l'intervalle $[0, 0.4]$ (d'après des expériences qui ne sont pas reportées ici, nous avons observé que des valeurs de s supérieures à 0.4 ne permettent pas d'obtenir de partitions satisfaisantes sur ce corpus).

Bien que les étiquettes de classes décrivant les partitions générées ne soient pas exploitées par les algorithmes évalués, nous en tenons compte lors du processus d'évaluation, et nous attestons de la qualité des partitions de manière supervisée. Nous utilisons à ce titre la mesure d'*information mutuelle normalisée* (Strehl & Ghosh, 2002) :

$$\mathrm{nmi}(\Pi, \boldsymbol{y}) = \frac{I(\Pi, \boldsymbol{y})}{\sqrt{H(\Pi)H(\boldsymbol{y})}}$$

où $\boldsymbol{y}$ est le vecteur de $[1..K]^n$ qui attribue à tout document son étiquette de classe. Π est la partition évaluée, $I(\Pi, \boldsymbol{y})$ est l'information mutuelle de Π et de $\boldsymbol{y}$; $H(\Pi)$ et $H(\boldsymbol{y})$ représentent respectivement l'entropie de Shannon de Π et de $\boldsymbol{y}$. Pour cette variante normalisée, $\mathrm{nmi}(\Pi, \boldsymbol{y}) \in [0, 1]$ doit être maximisée.

Pour chacun des jeux de données, 20 simulations sont effectuées : pour un même lancement, chacun des algorithme est initialisé avec une partition aléatoire identique.

8.4.1.3 Résultats et discussions

Sur la figure 8.2 sont présentés les scores moyens de *nmi* obtenus par l'algorithme des K-moyennes ellipsoïdales pour différentes valeurs de s, incluant le cas sphérique ($s = 0$), ainsi que pour la procédure de sélection automatique (s^*). Par souci de clarté, nous avons reporté les écarts-types associés aux scores de *nmi* uniquement pour le cas sphérique et la procédure de sélection automatique sur la partie gauche de la figure.

Sur la partie gauche de la figure 8.2, nous observons que lorsque le rapport n/m est petit, notre proposition produit des partitions significativement meilleurs que l'algorithme sphérique, pleinement dimensionnel. Plus précisément, lorsque $n \ll m$ les données sont mieux partitionnées sur un nombre réduit de dimensions. Tandis que dans l'espace d'entrée originel les dimensions spécifiques à chacun des clusters ne sont pas suffisamment représentées par les données, l'algorithme des K-moyennes ellipsoïdales projette ces

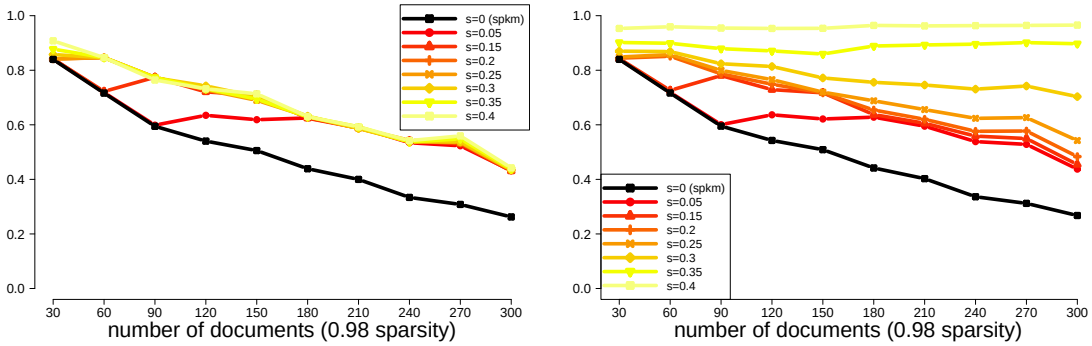


FIGURE 8.3 – Taux moyen de parcimonie des centroïdes sur 20 lancements (gauche). Pourcentage moyen de composantes inférieures à 10^{-3} (droite). L'axe des abscisses représente le nombre n de documents qui composent les jeux de données.

derniers sur des sous espaces pour lesquels une structure de clusters est plus facilement identifiable. La sélection de descripteurs vise à mettre en évidence les descripteurs pertinents pour décrire les données, dans ce cadre la procédure de sélection automatique retient systématiquement des ellipsoïdes exhibant un aplatissement non nul (i.e. des valeurs non nulles de $s < 1$). Par ailleurs, comme l'indiquent les écarts-types associés aux scores de *nmi*, nous constatons que les partitions produites par *ellkm* sont plus stables que celles obtenues par *spkm*.

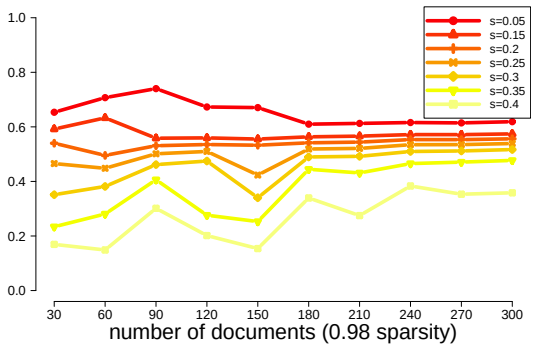


FIGURE 8.4 – Entropie moyenne des vecteurs de poids sur 20 lancements. L'axe des abscisses représente le nombre n de documents qui composent les jeux de données.

Sur la partie droite de la figure 8.2, nous avons représenté les performances moyennes associées aux ellipsoïdes pour différents allongements. Comme le montrent les scores obtenus pour $s = 0.4$, de plus grands ratios n/m nécessitent plus de dimensions (chacune des 100 dimensions spécifiques à chacun des clusters) afin de retrouver, dans les données, les clusters générés. Nous observons en particulier que les grands ratios sont bien traités par *spkm* et par *ellkm* pour de petites valeurs de s. A l'opposé pour les ratios plus petits qui correspondent au cas $n \ll m$, *spkm* éprouve de grandes difficultés tandis que les ellipsoïdes associés à de grands degrés d'aplatissement réalisent les meilleures performances.

Afin d'observer l'influence du ratio n/m sur les centroïdes obtenus d'après différents degrés d'aplatissement, nous avons représenté, sur la partie gauche de la figure 8.3, le

taux moyens de parcimonie des centroïdes (c'est-à-dire le nombre de composantes nulles) pour différentes valeurs de s, incluant le cas sphérique ($s = 0$). Comme attendu, pour des valeurs de s non nulles, *ellkm* favorise des centroïdes plus parcimonieux que *spkm*, en conséquence les partitions induites sont présumées plus stables puisque mon dépendantes aux faibles variations des données sur les composantes non pertinentes. Néanmoins, tandis que *ellkm* force les descripteurs non pertinents à des valeurs proches de zéro, il n'annule pas complètement leur importance. En effet, bien qu'un nombre significatif de composantes soit fixé à zéro, cet effet n'est pas accentué pour des ellipsoïdes exhibant de plus grands degrés d'aplatissement. Sur la partie droite de la figure 8.3 sont représentés les pourcentages moyens de composantes strictement inférieures à 10^{-3} (qui est la plus petite valeur observée dans les données). Nous constatons alors qu'à mesure que le degré d'aplatissement des ellipsoïdes augmente, pour les centroïdes résultants le nombre de composantes négligeables par rapport aux données augmente également. Ainsi, bien que les descripteurs non pertinents ne soient pas tous associé à un poids nul, leur importance est rapportée à une valeur négligeable par rapport aux données.

Sur la figure 8.4 sont représentées les entropies moyennes associées aux vecteurs de poids qui caractérisent les ellipsoïdes. De manière similaire, nous constatons que l'entropie moyenne décroît en fonction de s. Nous relevons une importante chute pour de grandes valeurs de s sur le jeu de données composé de $n = 150$ documents : les jeux de données sont générés de manière indépendante et contiennent tous un bruit non négligeable, nous soupçonnons qu'ici les centroïdes restent bloqués sur quelques dimensions.

Enfin, la figure 8.5 représente les vecteurs de poids associés à chacun des trois centroïdes pour $n = 120$ et $s^* = 0.2$. Nous observons qu'un poids important est correctement assigné à chacune des 100 dimensions spécifiques à chacun des clusters.

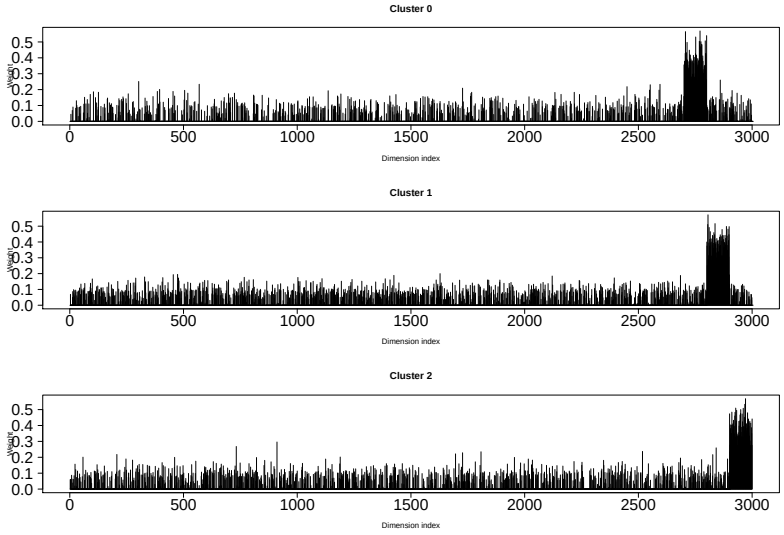


FIGURE 8.5 – Vecteurs de poids $\boldsymbol{\lambda}^s$ caractérisant chacun des ellipsoïdes correspondant à un lancement de *ellkm* pour $s^* = 0.2$ sur un jeu de données synthétiques de taille $n = 120$.

8.4.2 Données réelles : *20-newsgroup*

Nous présentons ici les résultats d'une évaluation comparative expérimentale de *ellkm* avec quatre autres algorithmes issus de l'état de l'art. Les expériences ont été réalisées

id.	catégories	n	m	n/m	parcimonie
1.1	soc.religion/comp.graphics	272	2455	0.1	0.98
2.1	comp.graphics/rec.sport.baseball/sci.space	250	1699	0.1	0.98
3.1	talk.politics.guns/talk.politics.mideast	260	3164	0.1	0.98
1.2	soc.religion/comp.graphics	1772	8895	0.2	0.99
2.2	comp.graphics/rec.sport.baseball/sci.space	2574	10368	0.2	0.99
3.2	talk.politics.guns/talk.politics.mideast	1790	10712	0.2	0.99

TABLE 8.2 – Description des trois corpus extraits depuis les données *20-newsgroup*. Deux jeux de données sont construits à partir de chacun des corpus : 1.1, 2.1 et 3.1 font référence au cas $n \ll m$ tandis que 1.2, 2.2 et 3.2 font référence au cas classique pour lequel n est plus grand.

sur trois corpus de données réelles, échantillonnés à partir de six catégories des données *20 newsgroup*, à savoir : *soc.religion*, *comp.graphics*, *rec.sport.baseball*, *sci.space*, *talk.politics.guns*, *talk.politics.mideast*. Pour chacun des trois corpus, deux jeux de données sont extraits : le premier contient très peu de documents de sorte que $n \ll m$, le second est composé de bien plus de documents et correspond à un cadre plus classique. Une description des jeux de données est fournie dans le tableau 8.2. Pour chacun, les documents sont équitablement répartis sur les K clusters correspondants.

8.4.2.1 Pré-traitement des données

L'ensemble des méta-informations liées au format des données comme les en-têtes de messages sont dans un premier temps filtrées. La casse des mots est ensuite normalisée de sorte que tous soient représentés par des caractères minuscules. Les mots sont alors réduits à leur forme lemmatique et filtrés selon leur type grammatical : l'ensemble des verbes auxiliaires (e.g. *être, avoir*), des déterminants (e.g. *le, un*) ainsi que des conjonctions (e.g. *ou, mais*) sont ainsi écartés. La lemmatisation ainsi que l'étiquetage grammatical sont tous deux effectués par le programme *TreeTagger* (Schmid, 1994). Pour constituer la matrice de représentation des données $\mathbf{X}$, nous adoptons le schéma de pondération *tf/idf*. Finalement pour chacun des jeux de données, les termes présents dans plus de 20% des documents ainsi que ceux qui apparaissent dans moins de deux documents sont écartés ; les documents qui comportent moins de 10 mots sont écartés. Les documents décrits sont alors projetés sur la sphère.

8.4.2.2 Algorithmes évalués

Nous avons sélectionné quatre algorithmes issus de l'état de l'art et nous comparons leurs performances à celles obtenues par *ellkm*. D'une part, *sparcl* (Witten & Tibshirani, 2010) et *ewkm* (Jing et al., 2007) sont deux extensions des K-moyennes classiques qui effectuent une sélection de descripteurs dans l'espace euclidien. Tandis que *ewkm* tente d'obtenir des sous espaces spécifiques à chacun des clusters, *sparcl* identifie un unique sous espace pour décrire l'ensemble des données. D'autre part, *spkm*, utilisé à la section précédente, et *plsa* (Hofmann, 2001) sont deux méthodes classiques pour le clustering données représentées par des matrices creuses, en grande dimension et de type textuel. Elles constituent toutes deux des méthodes pleinement dimensionnelles. Hormis *spkm* qui est a été présenté en détail à la section précédente, l'ajustement des paramètres pour chacun des algorithmes est effectué comme suit :

sparcl inclut une procédure basée sur la méthode du gap pour la sélection du paramètre d'ajustement qui contrôle la taille du sous espace de projection (Witten & Tibshirani, 2010). Dans nos expériences, 10 valeurs du paramètre sur $B = 10$ jeux de données de référence sont testées pour chacun des jeux de données.

ewkm emploie un paramètre d'ajustement γ pour contrôler l'entropie des vecteurs de poids qui définissent les sous espaces de projection. Dans leurs expérimentations, les auteurs fixent γ manuellement (Jing et al., 2007), nous évaluons 10 valeurs différentes et nous retenons la valeur pour laquelle l'algorithme obtient le plus grand score moyen de *nmi*. Il faut noter que cette procédure d'ajustement tient compte des étiquettes de clusters et *ewkm* est ici avantagé par rapport à ses concurrents.

plsa est une autre méthode classique pour la recherche d'information au sein de données représentées dans des espaces creux, en grande dimension (Hofmann, 2001). Ici, *plsa* est vu comme un algorithme de clustering pleinement dimensionnel. Dans nos expériences le document $\boldsymbol{x}$ est associé au cluster $\pi_k = \text{argmax}_{k=1}^{K} p(\boldsymbol{z}_k|\boldsymbol{x})$, où le thème $\boldsymbol{z}$ est vu comme un centroïde.

Pour chacun des algorithmes, le nombre K de clusters est fixé au nombre réel de classes composant les jeux de données.

8.4.2.3 Protocole expérimental

Sur chacun des six jeux de données, nous effectuons 20 simulations pour chacun des algorithmes. Comme précédemment (voir section 8.4.1), les partitions sont évaluées par le score de *nmi*. Nous utilisons de plus l'indice classique de Rand (Rand, 1971) noté *rand*, ainsi que la fréquence de la classe qui domine les clusters, aussi appelée score de *pureté*. Tandis que le score de *nmi* mesure la dépendance entre les partitions évaluées et les vraies classes, le score de pureté atteste de la cohérence du partitionnement, et l'indice de Rand donne une mesure de l'appariement entre un partitionnement des données et leur vrai distribution.

8.4.2.4 Résultats

Le tableau 8.3 présente les résultats obtenus par chacun des algorithmes dans le cas où $n \ll m$. Ceux obtenus dans un cas plus classique pour lequel un plus grand nombre de documents est disponible sont fournis dans le tableau 8.4. Les *n/a*, lorsqu'ils figurent, signifient que l'algorithme n'a pas fourni de résultats sous 24 heures.

Nous remarquons que *sparcl* et *ewkm* qui projettent les données dans des sous espaces de l'espace euclidien rencontrent des difficultés quel que soit le corpus et quel que soit le nombre de documents présentés. Comme souligné à la section 8.2, ces résultats confirment l'inadéquation de l'algorithme des K-moyennes classiques pour de telles données. Au contraire, nous constatons que *spkm* et *plsa*, qui partitionnent les documents dans un espace non euclidien, obtiennent des résultats meilleurs en moyenne, et ce, malgré l'emploi qui est fait d'une mesure de similarité pleinement dimensionnelle.

Dans le tableau 8.3 nous constatons que sur les jeux de données 1.1 et 3.1, composés tous deux de $K = 2$ classes, les partitions obtenues par *sparcl* sont proches de l'aléatoire comme l'indiquent leur indice de Rand et leur score de pureté. Ces résultats s'opposent à ceux obtenus sur le jeu de données 2.1 qui est lui composé de $K = 3$ classes : il semblerait que sur celui-ci, la procédure d'ajustement automatique trouve de meilleures indications d'une structure de clusters. Au contraire, *ewkm* obtient son meilleur résultat sur le jeu de données 1.1, cependant ses performances restent inférieures à *sparcl* dans tous les autres cas. Cela peut entre autres s'expliquer par le fait que *ewkm* tente d'obtenir K sous espaces

id.	méthode	nmi	rand	pureté
1.1	ellkm	0.57 ± 0	0.84 ± 0	0.91 ± 0
	spkm	0.26 ± 0.07	0.66 ± 0.02	0.72 ± 0.03
	plsa	0.38 ± 0.04	0.73 ± 0.02	0.81 ± 0.02
	sparcl	0.08 ± 0	0.50 ± 0	0.55 ± 0
	ewkm	0.11 ± 0.01	0.56 ± 0	0.65 ± 0.01
2.1	ellkm	0.36 ± 0.01	0.72 ± 0	0.7 ± 0.01
	spkm	0.14 ± 0.02	0.62 ± 0	0.52 ± 0.02
	plsa	0.12 ± 0.01	0.61 ± 0	0.53 ± 0.01
	sparcl	0.17 ± 0	0.44 ± 0	0.47 ± 0
	ewkm	0.09 ± 0	0.58 ± 0	0.50 ± 0
3.1	ellkm	0.12 ± 0.01	0.58 ± 0	0.68 ± 0.01
	spkm	0.02 ± 0	0.51 ± 0	0.56 ± 0
	plsa	0.02 ± 0	0.51 ± 0	0.58 ± 0
	sparcl	0.09 ± 0	0.50 ± 0	0.56 ± 0
	ewkm	0.04 ± 0	0.52 ± 0	0.59 ± 0

TABLE 8.3 – Comparaisons sur trois jeux de données extraits du corpus *20-newsgroup* et pour lesquels $n \ll m$.

de projection et qu'il ne peut estimer correctement l'ensemble de ses paramètres sur des données textuelles.

Dans les cas où $n \ll m$ (jeux de données 1.1, 2.1, 3.1), *ellkm* obtient des performances manifestement supérieures sur l'ensemble des critères d'évaluation. De plus, sur le jeu de données 1.1 pour lequel les méthodes pleinement dimensionnelles produisent des partitions significatives, nous constatons que les partitions produites par *ellkm* sont plus stables.

Pour de plus grandes valeurs de n (jeux de données 1.2, 2.2, 3.2), *plsa* et *spkm* obtiennent des résultats équivalents. Il faut noter que *plsa* produit de moins bonnes partitions sur le jeu de données 2.2, ce qui peut être en partie attribué à la configuration de l'algorithme : nous avons fixé un nombre d'itérations maximal à 80, une plus grande valeur pourrait conduire à de meilleurs résultats au prix de temps de calculs plus importants et d'une plus grande variance des résultats. Pour de plus grandes valeurs de n, la procédure de sélection automatique utilisée par *ellkm* ne trouve pas de sous espaces plus pertinents que l'espace d'entrée originel, excepté sur le jeu de données 1.2 pour lequel la valeur $s^* = 0.05$ est retenue. A mesure que le nombre n de documents croît, leurs vraies catégories sont mieux décrites sur plus de dimensions et *ellkm* est réduit à *spkm*. Il faut noter que *sparcl* ne termine pas pour les jeux de données 2.2 et 2.3. L'avantage d'une pénalisation de type *lasso* employée par l'algorithme est de fixer à exactement zéro l'importance de certains des descripteurs, néanmoins, au prix d'importants calculs.

Pour chacun des algorithmes, les meilleures performances sont obtenues sur le premier corpus (jeux de données 1.1 and 1.2). En effet, ce corpus exhibe deux clusters bien séparés, exceptés *sparcl* et *ewkm*, chacun des algorithmes tend à retrouver les vraies catégories associées aux documents. Le second corpus (jeux de données 2.1 et 2.2) est composé de trois clusters bien séparés et tous composés de moins de documents. Bien qu'ils présentent une performance inférieure, nous observons que *spkm* et *plsa* tendent à produire des partitions cohérentes. En revanche, le troisième corpus (jeux de données 3.1 et 3.2) est composé de deux sous-classes d'une même catégorie (*politics*). Il est ainsi attendu que les clusters partagent beaucoup de vocabulaire commun et en effet, toutes les méthodes éprouvent des

id.	méthode	nmi	rand	pureté
1.2	ellkm	0.76 ± 0	0.92 ± 0	0.96 ± 0
	spkm	0.77 ± 0	0.93 ± 0	0.96 ± 0
	plsa	0.75 ± 0	0.92 ± 0	0.96 ± 0
	sparcl	0.15 ± 0	0.53 ± 0	0.62 ± 0
	ewkm	0.09 ± 0.01	0.55 ± 0	0.63 ± 0.01
2.2	ellkm	0.67 ± 0	0.88 ± 0	0.90 ± 0
	spkm	0.67 ± 0	0.88 ± 0	0.90 ± 0
	plsa	0.53 ± 0.01	0.81 ± 0	0.82 ± 0.01
	sparcl	n/a	n/a	n/a
	ewkm	0.07 ± 0	0.56 ± 0	0.47 ± 0
3.2	ellkm	0.25 ± 0.04	0.64 ± 0.02	0.72 ± 0.02
	spkm	0.25 ± 0.04	0.64 ± 0.02	0.72 ± 0.02
	plsa	0.18 ± 0.03	0.61 ± 0.01	0.71 ± 0.01
	sparcl	n/a	n/a	n/a
	ewkm	0.03 ± 0	0.52 ± 0	0.57 ± 0

TABLE 8.4 – Comparaisons sur trois jeux de données extraits du corpus *20-newsgroup* et pour lesquels le nombre n de documents présentés est plus classique.

id.	ellkm	spkm
1.1	0.36 ± 0.02	0.25 ± 0
1.2	0.38 ± 0	0.29 ± 0
2.1	0.52 ± 0.01	0.28 ± 0
2.2	0.39 ± 0.01	0.39 ± 0.01
3.1	0.28 ± 0.02	0.20 ± 0.02
3.2	0.21 ± 0	0.21 ± 0

TABLE 8.5 – Comparaisons des taux moyens de parcimonie pour l'ensemble des centroïdes.

difficultés sur ce corpus.

Le tableau 8.5 présente les taux moyens de parcimonie pour les centroïdes produits par *ellkm* et *spkm*. Pour chacun des jeux de données, *ellkm* favorise des centroïdes plus parcimonieux et de la sorte, des partitions mieux interprétables. Dans une situation pour laquelle les temps de calculs constituent une contrainte importante, une implémentation efficace de la méthode proposée permet de tirer profit des composantes fixées à zéro lors du calcul de similarité entre paires de documents.

8.5 Conclusion

Nous avons proposé une extension de l'algorithme des K-moyennes sphériques pour effectuer une sélection de descripteurs sur des jeux de données textuelles, représentées dans des espaces creux en très grande dimension. Nous faisons l'hypothèse que les clusters se situent dans des régions denses de l'espace de description et nous exploitons une transformation qui consiste à transformer l'hypersphère unité en un ellipsoïde. Une étape additionnelle est ajoutée à l'algorithme originel des K-moyennes pour mettre à jour les ellipsoïdes spécifiques à chacun des clusters. L'algorithme résultant produit un ensemble

de centroïdes et d'ellipsoïdes associés qui constituent un optimum local pour l'homogénéité intra-clusters mesurée par la fonction de similarité proposée. Un paramètre d'ajustement s permet d'influer sur le degré d'aplatissement des ellipsoïdes : pour une valeur nulle de ce paramètre l'algorithme résultant est celui des K-moyennes sphériques, pour de plus grandes valeurs l'importance des descripteurs les moins pertinents est inhibée. Nous avons par ailleurs proposé une nouvelle heuristique pour la méthode du gap qui tient explicitement compte de la sensibilité de l'algorithme des K-moyennes à l'étape d'initialisation aléatoire. Nous avons alors montré l'efficacité d'une procédure de sélection automatique pour l'ajustement du paramètre s.

Nous avons conduit plusieurs expériences à la fois sur des jeux de données synthétiques et réelles. Dans les cas où le nombre de documents est largement inférieur au nombre de descripteurs, les résultats obtenus montrent l'intérêt de notre proposition. Nous observons également que l'algorithme que nous proposons favorise des centroïdes plus parcimonieux, pour lesquels de nombreuses composantes sont nulles. Par ailleurs, pour des jeux de données de taille plus grande, les expériences réalisées montrent que comme attendu, la procédure de sélection automatique réduit les ellipsoïdes à l'hypersphère unité. L'algorithme proposé réalise alors un partitionnement pleinement dimensionnel des données.

Notre proposition est principalement motivée par le problème du clustering de sources dynamiques qui publient de nouveaux documents en continu : à chaque pas de temps, les descripteurs nouvellement observés enrichissent l'espace d'entrée, les vecteurs de représentation des sources deviennent ainsi rapidement très larges et très creux. Au chapitre suivant, nous mettons en application l'algorithme des K-moyennes ellipsoïdales dans un tel cadre.

Chapitre 9

Mise en œuvre expérimentale : analyse dynamique de la presse française

Nous considérons un problème de partitionnement de sources dynamiques qui publie des documents sur Internet à intervalles fréquents. Pour ce faire nous étudions les publications de la presse française sur une période de cinq mois commençant au 1er août 2012. Les documents que nous exploitons sont des résumés d'articles publiés sur Internet et recueillis auprès des principaux éditeurs d'information en France, au travers de leurs fils de syndication [1]. Un certain nombre d'évènements ont marqué la période de notre étude : pour certains, comme la cérémonie de clôture des jeux olympiques d'été à Londres ou la réélection du président Barack Obama, il est attendu qu'ils constituent des phénomènes d'emballement soudain auprès des sources étudiées ; d'autres comme les évènements rythmant le débat politique français ont un caractère plus épisodique et sont enclins à connaître une attention plus diffuse. Comme nous le discutons au long de ce chapitre, un certain nombre de choix de paramétrage influence directement le dynamisme inhérent aux données ; suite au partitionnement, ces mêmes paramètres conditionnent alors la nature des communauté identifiées. Ces communautés, dynamiques par nature, se recomposent continuellement selon les évènements marquant l'actualité, aussi considérons-nous ses *threads d'information* qui sont associés aux périodes de remarquable stabilité sémantique.

Cette mise en œuvre expérimentale présente ainsi deux objectifs : le premier est de proposer une évaluation du modèle dynamique que nous étudions pour l'analyse de flux de données ainsi que des méthodes proposées pour traiter spécifiquement de sources d'information publiant sur Internet. Le second est une étude des publications de la presse française sur Internet.

Le corpus expérimental est présenté à la section 9.1 : les documents collectés quotidiennement sont dans un premier temps pré-filtrés puis représentés d'après les termes qui les composent. Dans un second temps, les sources responsables de l'information sont identifiées à partir des urls associées à chacun des documents. Dans ce cadre nous employons la méthode incrémentale d'identification de sources présentée au chapitre 7 : à chaque nouvelle date, les documents produits sont susceptibles de donner naissance à de nouvelles sources qui spécialisent celles identifiées par le passé. En vue du problème de clustering de sources dynamiques, nous observons à la section 9.2, qu'à tout instant, le nombre total de sources ainsi identifiées est largement inférieur au nombre total de documents produits

1. Nous avons utilisé une liste de fils rss, auxquels sont fréquemment déposées de nouvelles publications.

et donc au nombre total de termes pour les décrire. Nous employons ainsi l'algorithme de clustering de documents présenté au chapitre 8 qui répond à la problématique du fléau de la dimension en effectuant une pondération de descripteurs. A la section 9.3 nous nous concentrons sur deux faits importants ayant marqué l'année 2012 : les jeux olympiques d'été de Londres ainsi que les élections présidentielles américaines. Enfin les conclusions de ce chapitre sont présentées à la section 9.4.

9.1 Constitution d'un corpus de sources dynamiques

A partir des données collectées durant 115 jours (un peu moins de quatre mois), du 1er août au 23 novembre 2012, nous extrayons un corpus de sources dynamiques sur lequel seront formées les communautés de sources. Dans un premier temps, les documents recueillis sont pré-filtrés puis représentés d'après les termes qui les composent. Dans un second temps, un ensemble de sources, responsables de l'information recueillie, est identifié puis représenté dans l'espace d'entrée des documents. Sur la période d'étude, nous avons ainsi constitué un corpus de sources dynamiques qui se déplacent, dans l'espace d'entrée, au gré des publications qu'elles émettent au cours du temps.

9.1.1 Modélisation de l'information

Sur l'ensemble de la période, un peu plus de 130 000 documents ont été publiés aussi bien sur des *blogs* d'information qu'au travers des principaux médias français. Ces documents constituent des résumés d'articles et sont recueillis quotidiennement sur les fils de syndication des principaux éditeurs d'information en France. Ces derniers étant continuellement mis à jour, nous obtenons au travers de ces fils un aperçu quotidien de l'actualité française.

Sur la partie haute de la figure 9.1 est représenté le nombre de documents observés à chaque nouvelle date, les documents comportant moins de six termes ayant été préalablement écartés. Nous observons que les jours ouvrés [2] présentent un nombre comparable de publications, autrement dit nous constatons une couverture homogène de l'information sur l'ensemble de la période. Une hausse remarquable est néanmoins visible entre le mois d'août et le mois de septembre, que nous attribuons à la fin des vacances scolaires françaises. Sur la partie basse de la figure, est représenté, à chaque date, le nombre cumulé de documents recueillis depuis le 1er août 2012. Nous remarquons que cette quantité suit une croissance quasi-linéaire : les analyses faites dans la suite ne sont donc pas assujetties à de brusques changements portant sur la fréquence des publications.

Comme décrit au chapitre 6, à chaque nouvelle date, un espace de représentation est constitué : les documents nouvellement publiés sont dans un premier temps filtrés grammaticalement puis représentés selon leurs unigrammes [3]. A l'instant t, l'espace de description $\mathcal{X}(t)$ est formé comme l'union de l'espace de description $\mathcal{X}(t-1)$ et de l'ensemble des nouveaux descripteurs ainsi obtenus. Sur la figure 9.2 nous avons représenté l'évolution de la taille de cet espace, on constate que cette quantité semble suivre, de manière classique, une croissance logarithmique.

Les documents nouvellement publiés sont ainsi décrits à tout instant dans cet espace, le schéma de pondération adopté est alors le classique *tf/idf* présenté à la section 1.1.1.1,

2. Trois jours de semaine subissent une baisse importante de publications : le 15 août et le 1er novembre correspondent à des jours fériés en France ; la baisse observée au 10 octobre est causée par une défaillance du système de collecte ce jour-là.

3. Seuls les noms propres et les noms communs sont retenus. L'étiquetage grammatical est effectué par *TreeTager*, une liste de mots vides est également employée (voir section 1.2.1.1, p. 20).

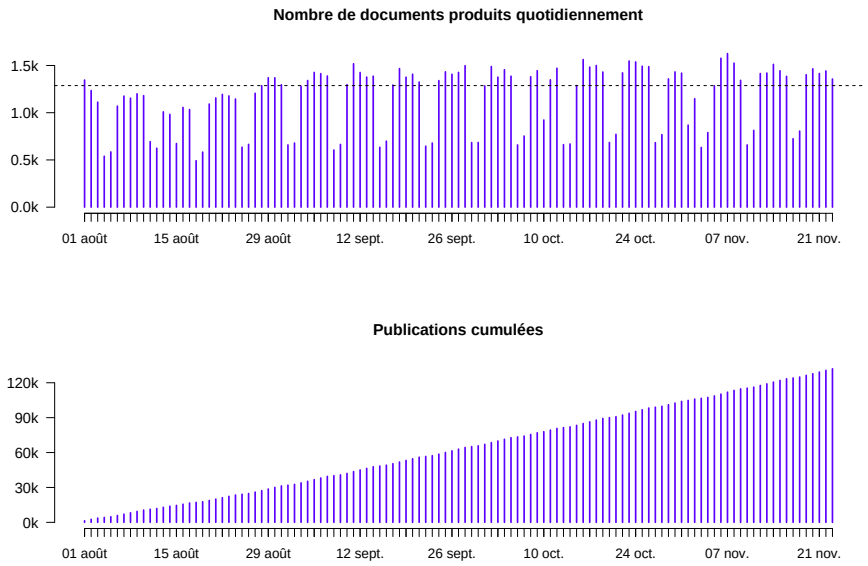


FIGURE 9.1 – Nombre de documents produits quotidiennement (haut), publications cumulées sur la période d'étude (bas).

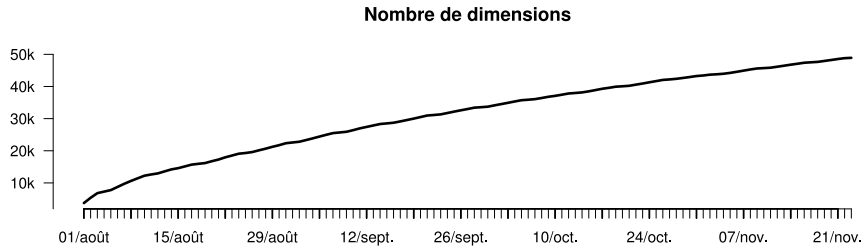


FIGURE 9.2 – Nombre de termes composant l'espace de description $\mathcal{X}(t)$ en fonction du temps.

p. 8 : pour chaque terme le tf (*term frequency*) correspond au nombre total de documents correspondants, rapporté au nombre total de documents nouvellement publiés. A chaque nouvelle date, il n'est ainsi pas nécessaire d'ajuster la représentation faite des documents observés par le passé.

Enfin, les vecteurs de description obtenus sont projetés sur l'hypersphère unité de sorte que chacune des composantes décrive une direction de l'espace d'entrée.

9.1.2 Sources d'information dynamiques

Les documents collectés auprès des fils de syndication ont été publiés sur un total de 230 domaines. Parmi ceux-ci, près de 73% sont des blogs d'information et seulement 27% constituent des média d'information. Néanmoins, ces derniers représentent à eux seuls la quasi totalité de l'information produite sur l'ensemble de la période. Sur la partie haute de la figure 9.3 nous avons représenté, pour chacun des domaines, son volume de publication par rapport à son intervalle de publication, mesuré comme le nombre de jours écoulés entre la première et la dernière publication. Nous constatons ainsi qu'une majorité des

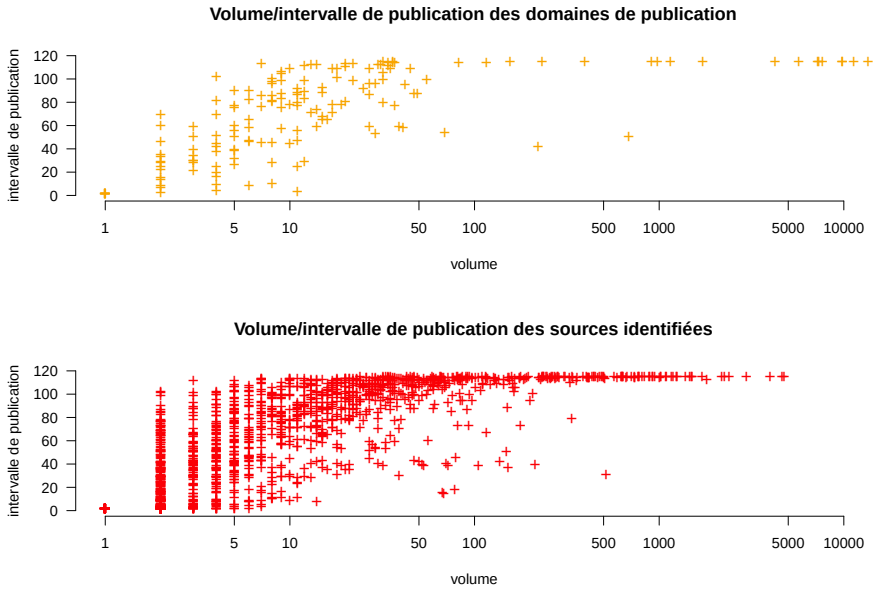


FIGURE 9.3 – Intervalles de publication des sources mesurés en nombre de jours entre la première et la dernière publication, rapportés aux volumes de publication. Les sources sont identifiées comme les domaines de publication (haut) et de manière à maximiser leur homogénéité (bas).

domaines publie environ entre 2 et 50 documents sur un intervalle moyen autour de 80 jours (un peu plus de deux mois et demi). Cette majorité est essentiellement composée des blogs d'information qui semblent naturellement exhiber une ligne éditoriale homogène. Sur le graphique, nous observons également qu'une grande majorité des documents collectés ont été déposés sur Internet au travers de quelques domaines seulement, ces derniers correspondent aux principaux médias d'information français. A la différence des blogs, ces domaines publient de nombreux documents susceptibles d'aborder des thématiques très variées et d'une manière plus générale, susceptibles de suivre des lignes éditoriales très différentes.

Dans le but de raffiner les sources en terme de l'information qu'elles génèrent, nous décomposons chacun des domaines en un sous-ensemble de sources, plus homogènes. Pour ce faire, nous exploitons les outils proposés au chapitre 7 et nous employons ainsi la méthode incrémentale d'identification de sources présentée à la section 7.4, p. 127. A mesure que les urls associées aux nouvelles publications sont observées, de nouvelles sources sont identifiées ; ces dernières réalisent, à toute nouvelle date, une spécialisation plus homogène des sources préalablement identifiées. Nous obtenons ainsi, à chaque date, un ensemble de sources homogènes, responsables de l'information étudiée. Pour chacune des sources identifiées au dernier jour de la période, nous examinons à nouveau le rapport entre le volume et l'intervalle de publication, représenté sur la partie basse de la figure 9.3. Outre le nombre désormais plus important de sources, nous constatons une meilleure répartition de l'information sur ces dernières. Néanmoins, nous relevons toujours une légère concentration du volume entre deux publications et une cinquantaine pour un intervalle moyen autour de 80 jours : tandis que pour une grande majorité des blogs il n'existe pas de sources plus homogènes que les domaines de publication, les domaines associés aux médias sont segmentés en sous-ensembles de sources qui suivent une ligne éditoriale plus homogène.

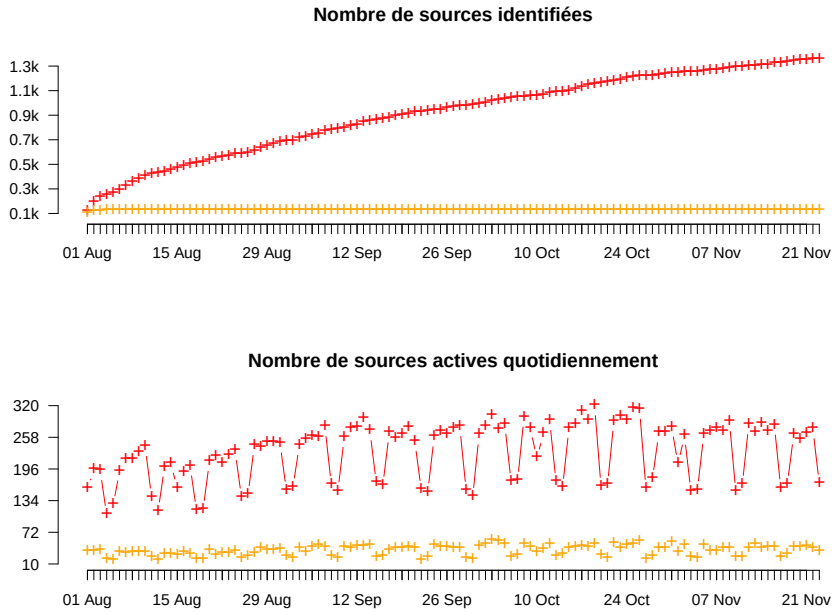


FIGURE 9.4 – Nombre de domaines observés (jaune) et de sources homogènes identifiées (rouge) au cours du temps (haut). Nombre de domaines (jaune) et de sources (rouge) qui publient de nouveaux documents au cours du temps (bas).

Sur la partie haute du graphique 9.4 est représenté, en fonction du temps, le nombre de sources homogènes identifiées (en rouge) ainsi que le nombre de domaines observés (en jaune). Tandis que la quasi-totalité des domaines est observée après quatre jours, le nombre de sources homogènes croît en fonction du temps : à mesure que de nouvelles urls sont observées, l'algorithme d'identification dispose de nouveaux tokens susceptibles de spécialiser d'anciennes sources. Sur la partie basse de la figure est représenté le nombre de sources actives à chaque date (en rouge) et le nombre de domaines actifs (en jaune) : celles qui publient de nouveaux documents. Nous observons que ces dernières représentent, à chaque date, une faible proportion du nombre total de sources : tandis que certaines ne publient que rarement, d'autres ne publient plus au bout de quelque temps. De même, comme sur la figure 9.1, nous constatons une dépendance hebdomadaire du nombre de sources actives. Ainsi, en vue du problème de clustering sous-jacent est-il important de tenir compte de l'âge de publication des sources : l'influence d'une source qui ne se déplace plus dans l'espace d'entrée est alors diminuée en fonction du temps.

9.2 Partitionnement de sources dynamiques

Le corpus constitué $X(t)$ est composé des coordonnées des vecteurs de publication des sources dans l'espace de représentation $\mathcal{X}(t)$. Comme décrit au chapitre 6, ces dernières sont obtenues par agrégation de l'information produite par chacune. Dans cette section, nous exploitons ce corpus dans le cadre d'une tâche de clustering de sources dynamiques telle que présentée au chapitre 6. Pour ce problème, les sources évoluent dans un espace de représentation textuel au gré des documents qu'elles publient.

Nous souhaitons identifier un partitionnement du corpus qui reflète à tout moment

les derniers changements observés. Dans un premier temps nous discutons le choix de l'algorithme des K-moyennes ellipsoïdales pour le problème considéré. Comme présenté au chapitre 8, ce dernier exploite les particularités d'un espace de description textuel et repose sur une mesure de similarité non uniforme.

Sur l'ensemble de la période, les partitions sont par ailleurs soumises à un fort dynamisme : d'une part les communautés se recomposent régulièrement dans le temps, d'autre part les thématiques qui leur sont associées évoluent. Dans un second temps, nous proposons d'observer ce dynamisme en examinant d'une part les transitions au sein des communautés, les mouvements des communautés dans l'espace de description. Nous étudions également le rôle des paramètres employés sur le dynamisme.

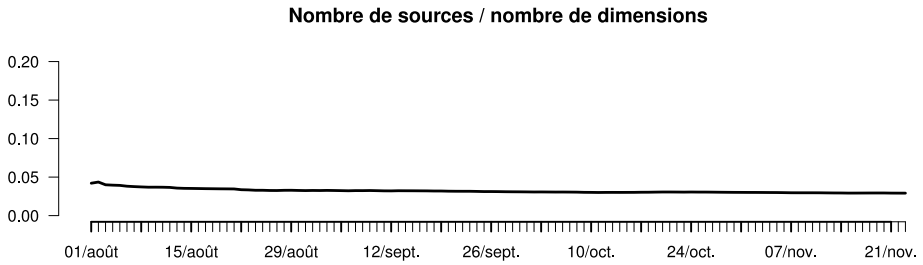
9.2.1 Algorithme de partitionnement : K-moyennes ellipsoïdales

Pour le clustering de données textuelles, les K-moyennes ellipsoïdales effectuent un partitionnement des données sur des ellipsoïdes tout en exploitant les particularités d'un espace de représentation bas niveau (voir chapitre 1). Les ellipsoïdes associés localement à chaque cluster sont identifiées automatiquement, leur orientation dans l'espace d'entrée indiquent alors les dimensions les plus pertinentes au sens de l'homogénéité de la partition constituée. Leur aplatissement est ajustable par l'intermédiaire du paramètre de parcimonie s qui influe sur la quantité d'information émanant d'un cluster : pour une faible valeur (proche de 0), les sous-espaces de projection sont poussés vers des formes sphériques pour lesquelles l'information peut se répartir de manière homogène sur toutes les dimensions. A mesure que le paramètre de parcimonie tend vers 1, les sous-espaces dégénèrent généralement en des ellipsoïdes présentant un aplatissement pour lequel quelques dimensions suffisent à capturer l'intégralité de l'information.

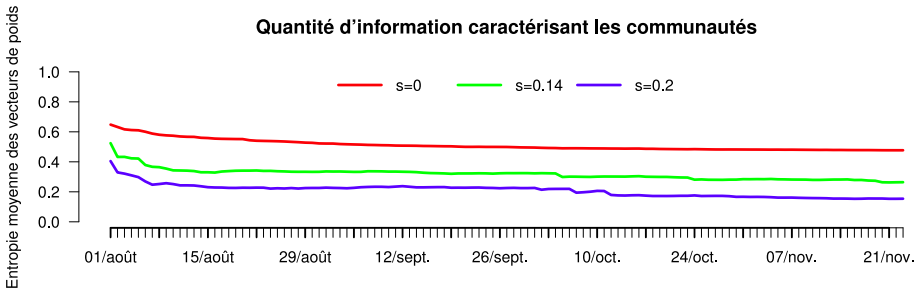
Sur la figure 9.5(a) est représenté le ratio entre le nombre de sources et le nombre de dimensions de $\mathcal{X}(t)$. Comme au chapitre 8, les faibles valeurs observées semblent indiquer la nécessité de contraindre les ellipsoïdes à un paramètre de parcimonie non nul. Néanmoins, cette quantité ne donne qu'une indication approximative de la sensibilité du problème au dilemme du fléau de la dimension, dans un contexte dynamique le vieillissement imposé aux données joue en effet un rôle non négligeable.

Dans le cadre de cette mise en œuvre expérimentale, nous exploitons plutôt le paramètre de parcimonie comme un moyen de contrôle sur la quantité d'information émanant des clusters. Une valeur proche de 0 encourage l'identification de communautés expliquées autour d'une grande majorité des termes composant l'espace d'entrée. Pour une valeur strictement positive, au sein des communautés, l'information est condensée autour de seulement quelques termes spécifiques. Sur la figure 9.5(b) nous avons représenté la quantité d'information moyenne émanant des communautés pour différentes valeurs de s. La figure montre, en fonction du temps, la moyenne des mesures d'entropie associées à chacun des $K = 10$ vecteurs de poids qui caractérisent les sous-espaces de projection. Pour la communauté π_k, il s'agit de l'entropie du vecteur de poids $\boldsymbol{\lambda}_k$ dont les composantes somment à 1. Il faut noter que pour le problème de clustering ces coefficients sont toujours rapportés à la puissance s. Les mesures d'entropie que nous étudions ici donnent néanmoins un aperçu de la répartition de l'information sur les sous-espaces de projection. A toute date nous constatons comme attendu que les communautés se forment autour de thématiques plus spécifiques à mesure que s augmente.

Entre deux dates, les nouvelles publications donnent lieu à de nouvelles sources et engendrent des déplacements pour les sources préalablement identifiées. Les sources transitent ainsi entre les communautés au gré de leurs déplacements, ce qui se manifeste à la



(a) Ratio entre le nombre de sources et le nombre de dimensions pour les décrire en fonction du temps.



(b) Quantité moyenne d'information retenue quotidiennement au sein de partitions formées autour de $K = 10$ communautés sans vieillissement.

FIGURE 9.5 – Description du problème de clustering et distribution de l'information pour différentes valeurs du paramètre de parcimonie.

fois par le déplacement des centroïdes dans l'espace d'entrée, mais aussi par l'instabilité des communautés au fil du temps. Dans la suite nous proposons d'étudier ces deux types de dérives en observant leur sensibilité au nombre K de communautés recherchées, à la durée $t_{1/2}$ de la demi-vie caractéristique du vieillissement imposé ainsi qu'au paramètre s des K-moyennes ellipsoïdales qui influe sur la quantité d'information émanant des communautés.

9.2.2 Transitions au sein des communautés

Les sources se déplacent dans l'espace de description au gré des documents qu'elles publient, elles transitent ainsi de communautés en communautés selon leurs changements d'intérêt. Sur l'ensemble de la période, ces transitions caractérisent un dynamisme pour le partitionnement réalisé à toute date, nous proposons ici d'observer l'influence des paramètres sur ce dernier.

Paramètres étudiés Les déplacements des sources dans l'espace de description étant directement responsables des transitions effectuées, nous étudions l'influence de leurs historiques de publication sur le dynamisme engendré. Pour ce faire nous examinons les effets d'un vieillissement exponentiel imposant aux données une demi-vie de $t_{1/2} \in \{\infty, 7, 3, 1\}$ jour(s) (voir section 6.3.2.1, p. 110), où ∞ indique qu'aucun vieillissement n'est appliqué.

De plus, toute communauté étant projetée sur un sous-espace correspondant, nous étudions également le rôle du paramètre de parcimonie sur ce dynamisme. Pour ce faire nous comparons le cas sphérique ($s = 0$) pour lequel l'espace originel est conservé au cas ellipsoïdal $s \in \{0.14, 0.20\}$ qui concentre les sous-espaces de projection dans des régions spécifiques.

(a) $K = 5$

	sp.(0)	**ell.(0.14)**	**ell.(0.20)**
∞	0.02	0.03	0.04
7	0.03	0.06	0.08
3	0.06	0.09	0.13
1	0.15	0.20	0.25

(b) $K = 8$

	sp.(0)	**ell.(0.14)**	**ell.(0.20)**
∞	0.02	0.03	0.05
7	0.04	0.07	0.08
3	0.08	0.11	0.15
1	0.20	0.25	0.29

(c) $K = 10$

	sp.(0)	**ell.(0.14)**	**ell.(0.20)**
∞	0.02	0.03	0.05
7	0.04	0.07	0.09
3	0.08	0.12	0.15
1	0.24	0.27	0.30

(d) $K = 15$

	sp.(0)	**ell.(0.14)**	**ell.(0.20)**
∞	0.02	0.03	0.04
7	0.04	0.07	0.10
3	0.09	0.13	0.16
1	0.26	0.29	0.33

TABLE 9.1 – Taux moyen de transitions par jour et par communauté, mesuré par Γ (défini dans l'éq. 9.1) et exprimé en fonction de la demi-vie de l'information (lignes) $t_{1/2} \in \{\infty, 7, 3, 1\}$ et du taux de parcimonie imposé (colonnes) $s \in \{0, 0.14, 0.2\}$.

Enfin, le nombre K de communautés recherchées conditionne le partitionnement réalisé à tout moment, nous observons l'évolution du nombre de transitions pour des partitions composées de $K \in \{5, 8, 10, 15\}$ communautés.

Taux moyen de transitions Sur l'ensemble de la période, T partitionnements sont réalisés depuis la date origine t_0. Nous proposons de mesurer les transitions effectuées en moyenne par jour et par communauté à partir du taux de transition $1 - \tau(k, t) \in [0, 1]$ donné dans l'équation (6.9), p. 113. Parmi la population de la communauté π_k au temps $t - 1$, ce taux mesure la proportion d'individus qui ont transité vers une communauté différente au temps t. Pour une série de partitions $[\Pi_1 \dots \Pi_T]$ nous mesurons ainsi un *taux moyen de transitions* de la manière suivante :

$$\Gamma(\Pi_1, \dots, \Pi_T) = \frac{1}{TK} \sum_{t=t_0}^{T} \sum_{k=1}^{K} 1 - \tau(k, t) \in [0, 1] \tag{9.1}$$

Ce taux est maximal et vaut 1 lorsqu'à toute nouvelle date, les communautés subissent toutes un renouvellement complet de population. Il est nul quand le partitionnement ne change jamais entre deux dates.

Résultats et discussion Les résultats sont reportés dans le tableau 9.1.

Quand les données ne sont pas vieillies ($t_{1/2} = \infty$), nous constatons que quel que soit le nombre K de communautés recherchées et quelle que soit la valeur du paramètre de parcimonie, les partitions sont peu dynamiques sur l'ensemble de la période. En effet, à mesure que l'historique de publication des sources grandit, leurs déplacements diminuent et leurs transitions se font plus rares. Dans cette configuration, les sources sont quasi stationnaires et le problème de partitionnement correspondant s'approche d'une tâche de clustering traditionnel. Pour un vieillissement non nul (à partir de la deuxième ligne), nous observons un dynamisme plus important : l'historique des sources perd de son importance et ces dernières effectuent de plus grands déplacements dans l'espace de représentation.

Nous constatons également une dépendance du dynamisme à la valeur du paramètre de parcimonie. Dans le cas ellipsoïdal, à mesure que s augmente, la quantité d'information

émanant des communautés diminue. Les sources sont alors projetées sur des sous espaces où seules les dimensions les plus pertinentes sont retenues. L'inertie des clusters se voit ainsi réduite et les transitions effectuées par les sources entre communautés favorisées.

Enfin, nous observons une légère dépendance du taux moyen de transitions au nombre de communautés recherchées. A mesure que le nombre K de régions qui partitionnent l'espace de description augmente, il est plus facile pour les sources de transiter entre les communautés et de plus grands taux sont ainsi observés.

9.2.3 Déplacement des communautés

Les transitions effectuées par les sources au sein de la partition sont directement liées à leurs déplacements dans l'espace de représentation. Nous avons observé ci-dessus ces transitions en moyenne par jour et par communauté et nous avons montré qu'elles varient de manière non négligeable selon la demi-vie $t_{1/2}$ imposée au données, selon le taux de parcimonie s qui contraint les ellipsoïdes, et dans une moindre mesure selon le nombre K de communautés recherchées. Du fait de ces déplacements, les communautés dérivent elles-mêmes dans l'espace de représentation, nous examinons maintenant les changements de thématiques correspondants.

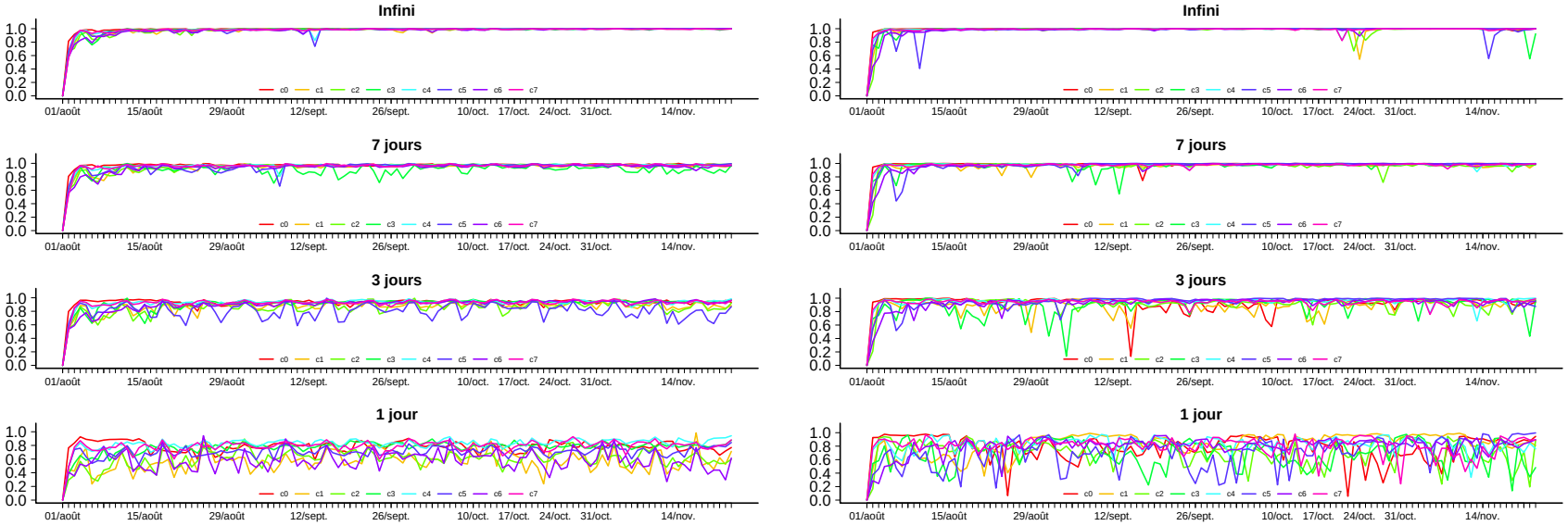
Influence des paramètres Comme présenté au chapitre 6, la thématique associée à une communauté est caractérisée par la direction que donne son représentant dans l'espace de description. Nous proposons ainsi d'observer l'influence des paramètres sur les changements de thématiques au travers des déplacements effectués par les centroïdes représentant les communautés dans l'espace de description.

Il faut remarquer que la trajectoire d'une communauté est contrainte par la nature des sous-espaces de projection : en vue de leur comparaison, les déplacements correspondants ne peuvent être, comme précédemment, synthétisés sur l'ensemble de la période d'étude. Pour les mêmes paramètres qu'employé précédemment nous proposons alors d'étudier deux cas extrêmes pour un partitionnement intermédiaire : nous fixons le nombre de communautés à $K = 8$ et nous comparons les mouvements des communautés obtenues dans le cas sphérique ($s = 0$) à ceux réalisés dans le cas ellipsoïdal ($s = 0.2$). Nous étudions également l'effet du vieillissement sur ces déplacements : comme précédemment nous imposons aux données une demi-vie de $t_{1/2} \in \{\infty, 7, 3, 1\}$ où ∞ indique qu'aucun vieillissement n'est appliqué.

Alignement dans l'espace de description Tel que présenté au chapitre 6, nous proposons d'observer les mouvements des communautés au travers de leur alignement dans l'espace d'entrée. En notant $\boldsymbol{d}_k(t)$ la direction donnée par les M termes les plus importants de la thématique associée à la communauté π_k au temps t, et définie dans l'équation (6.7), p. 113, nous mesurons l'alignement donné dans l'équation (6.8), p. 113 et rappelé ici :

$$\cos\ \theta(k,t) = \boldsymbol{d}_k(t-1)^\intercal \boldsymbol{d}_k(t) \in [0,1]$$

il est total et vaut 1 quand la direction donnée à l'état courant dans π_k n'évolue pas depuis l'état précédent, et il est nul quand ces deux directions sont orthogonales. Dans la suite nous qualifions de stationnaire une communauté dont l'alignement présenté est total.



(a) $s = 0$, cas sphérique

(b) $s = 0.2$, cas ellipsoïdal

FIGURE 9.6 – Déplacements des communautés pour $K = 8$ et $s = 0.2$ mesurés par leur alignement (défini dans l'éq (6.8), p. 113).

Résultats et discussions La figure 9.6 représente l'alignement en fonction du temps de la direction donnée par chacune des communautés sur leurs $M = 100$ termes les plus caractéristiques.

Nous constatons à nouveau une influence non négligeable du vieillissement imposé sur les données : quand $t_{1/2} = \infty$, l'importance de l'historique de publication n'est pas ajusté et à quelques rares exceptions, les communautés atteignent rapidement un régime stationnaire. Comme rappelé précédemment dans ce cadre les données évoluent peu dans l'espace de représentation et le problème de partitionnement correspondant s'approche d'une tâche de clustering traditionnel. A mesure que le vieillissement augmente, les sources effectuent de plus grands déplacements ce qui se traduit par des changements de thématiques plus fréquents.

Lorsque le paramètre de parcimonie est nul, les sous-espaces de projection sont des sphères sur lesquelles les communautés se déplacent librement dans toutes les directions. Or cet espace étant composé d'un très grand nombre de termes, à l'état courant un changement de direction important nécessite un fort dynamisme de la part de sa population. Comme le montre la figure 9.6, dans le cas sphérique les communautés subissent ainsi un mouvement continu et progressif. Pour une demi-vie de 3 jours par exemple, la communauté c_5 ne se stabilise jamais complètement, sa thématique évolue continuellement sans pour autant changer radicalement. Au contraire, on constate de plus amples changements sur la figure pour un vieillissement paramétré par une demi-vie d'une journée.

Pour un paramètre de parcimonie non nul, les sous-espaces de projection sont des ellipsoïdes sur lesquels les déplacements des communautés sont contraints. L'ellipsoïde associé à une communauté définit en effet une région de l'espace de description dans laquelle les sources d'une communauté sont encouragées à se déplacer : dans cet espace réduit, les mouvements d'une communauté sont limités et la direction que donne sa thématique présente un alignement plus grand. Ainsi, dans le cas ellipsoïdal le mouvement des communautés peut être décrit comme une série de décrochages : une communauté qui s'aligne dans une région y demeure stationnaire jusqu'à ce que sa population l'incite à en déterminer une nouvelle. Comme représenté sur la figure 9.6, l'alignement des communautés dans le temps forme des motifs en « dents de scie » de plus en plus marqués en fonction du vieillissement employé.

En période de stabilité, l'alignement d'une communauté est alors plus important que dans le cas sphérique. Son mouvement est par ailleurs plus ample lors d'un décrochage. Ainsi, à mesure que le dynamisme augmente, les communautés correspondantes s'adaptent mieux aux mouvements qui opèrent dans l'espace de représentation puisqu'elles se forment autour de thématiques plus spécifiques.

9.3 Résultats expérimentaux

Nous discutons plus en détail les résultats obtenus en fixant les paramètres aux valeurs suivantes : nous identifions un ensemble de $K = 8$ communautés, projetées sur des ellipsoïdes contraints à un paramètre de parcimonie de $s = 0.2$. De plus, le vieillissement employé impose aux données une demi-vie de $t_{1/2} = 3$ jours.

Dans un premier temps nous discutons des communautés identifiées sur l'ensemble de la période : un fort dynamisme régit à toute date le partitionnement, de sorte que les regroupements évoluent grandement entre intervalles de temps consécutifs.

Dans un second temps nous proposons d'identifier les threads d'information associés : tandis que certaines communautés sont gouvernées par une thématique générale sur l'ensemble de la période, d'autres connaissent des changements complets de direction se-

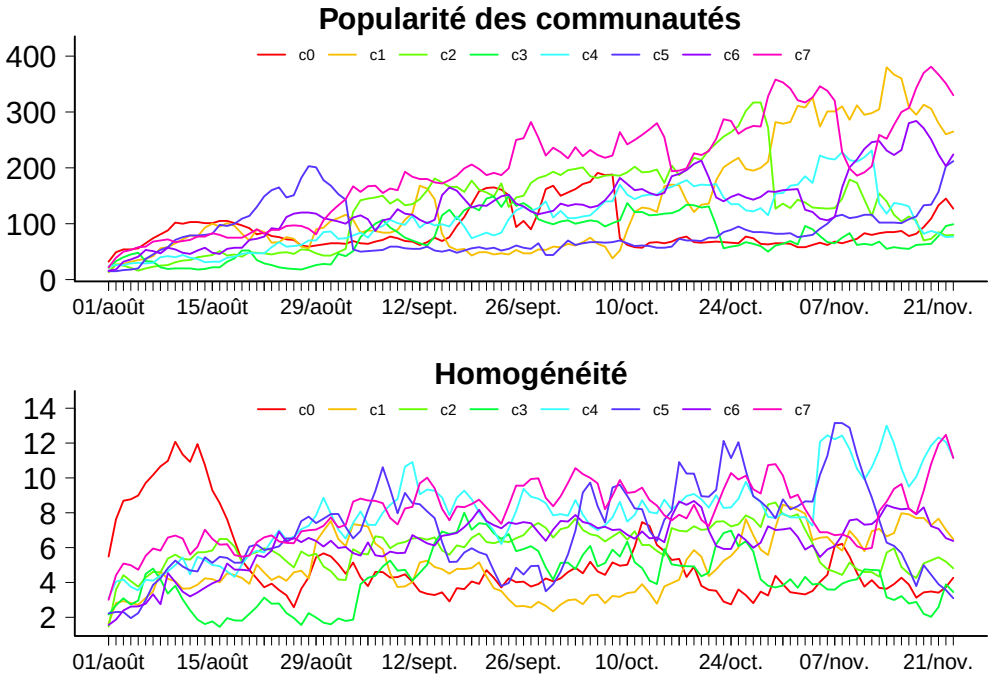


FIGURE 9.7 – Popularité quotidienne des communautés mesurée en nombre de sources (haut) et homogénéité mesurée par F_{ellkm} (bas) pour $K = 8$, $s = 0.2$ et $t_{1/2} = 3$ jours.

lon leurs renouvellements de population, nous souhaitons identifier pour ces dernières les périodes durant lesquelles elles forment un regroupement de remarquable stabilité sémantique.

Nous prêtons enfin notre attention à deux faits importants de l'année 2012 : les jeux olympiques d'été de Londres ainsi que la campagne présidentielle américaine qui s'achève en fin d'année. Chacun jouit d'une portée internationale : pour chacun nous discutons des threads identifiés dans la presse française ainsi que des évolutions de ces derniers.

9.3.1 Résultats globaux

Pour le jeu de paramètres employé, nous utilisons deux mesures de qualité pour évaluer les communautés formées à tout instant : nous mesurons d'une part le nombre de sources qui composent une communauté, appelé sa popularité, nous évaluons d'autre part son score d'homogénéité donné par la fonction objectif des K-moyennes ellipsoïdales F_{ellkm}. Nous rappelons que pour cet algorithme, les données présentées tiennent compte d'un historique de publication, les partitions obtenues à chaque date représentent ainsi un optimum local pour une mesure d'homogénéité historique. Les résultats sont présentés sur la figure 9.7.

Popularité des communautés Le graphique du haut représente la popularité de chacune des $K = 8$ communautés identifiées en fonction du temps pour toute la période d'étude : la somme quotidienne de ces scores fournit l'évolution du nombre de sources traitées. Tandis que certaines communautés regroupent globalement un grand nombre de sources, d'autres connaissent un succès moins important. Toute communauté étant formée autour d'un sujet fédérateur, pour les sources étudiées certaines thématiques semblent plus populaires que d'autres.

Nous constatons également que les communautés varient elles-mêmes dans le temps. Sur l'ensemble de la période la popularité d'une communauté connaît des changements plus ou moins importants et progressifs. Ces changements semblent respecter le motif général suivant : une montée progressive de la popularité est suivie d'une chute soudaine. Ainsi, tandis que les sources tardent à s'intéresser à un nouveau sujet, elles semblent convenir de son abandon d'un commun accord. Nous associons ces périodes à des phénomènes d'emballement autour de quelques sujets d'actualité précis.

Homogénéité des communautés Le graphique du bas représente l'homogénéité, historique, associée à chacune des communautés en fonction du temps : la somme quotidienne de ces scores donne l'évolution de la valeur objectif des K-moyennes ellipsoïdales. Sur l'ensemble de la période, nous constatons que les partitions identifiées quotidiennement sont composées de regroupements plus ou moins denses de sources. Tandis qu'il existe une association naturelle entre le nombre de sources observées (qui grandit au cours du temps) et cette densité, certaines périodes sont néanmoins marquées par un partitionnement de meilleure qualité. C'est notamment le cas au début du mois de septembre, lors de la rentrée scolaire française.

L'homogénéité varie également selon les communautés : certaines constituent globalement des regroupements compacts de sources, d'autres ont plus de peine à rassembler leur population autour d'une thématique précise.

Enfin, comme précédemment une même communauté obtient des résultats variables dans le temps. Nous observons plus précisément des périodes durant lesquelles les communautés atteignent des pics d'homogénéité. Pour une même communauté, ces périodes constituent des phénomènes d'emballement d'un autre genre durant lesquelles sa population forme un regroupement solidaire autour d'une thématique précise.

9.3.2 Identification des threads d'information

Comme observé précédemment, une même communauté subit de nombreux changements dans le temps. D'une part sa popularité évolue en fonction des sources qu'elle absorbe et de celles qui l'abandonnent, d'autre part, son homogénéité reflète à tout moment la qualité de la thématique qui la gouverne. Ainsi selon les déplacements effectués par les sources, des transitions opèrent à tout moment au sein de la partition et entraînent aussi bien des changements de thématiques que des renouvellements de population pour les communautés correspondantes. Selon l'ampleur de ces déplacements, une même communauté est alors susceptible de changer radicalement entre deux dates données.

Dans le but de caractériser finement dans le temps les regroupements qu'une communauté constitue, nous réalisons ici un découpage temporel de ces dernières. Nous proposons en particulier d'identifier les threads d'information définis à la section 6.4.3, p. 114. Pour une communauté donnée, nous identifions ainsi toute plage temporelle durant laquelle elle constitue un regroupement de sources cohérent. Ces plages représentent alors des périodes suffisamment longues, pendant lesquelles une communauté est gouvernée par une thématique remarquablement stable. Nous imposons une durée minimale de $\delta_t^{\min} = 6$ jours à ces périodes.

9.3.2.1 Cas général

La figure 9.8 décrit le découpage temporel des communautés, à chacun des threads d'information est donné un identifiant correspondant à son ordre d'apparition. Dans le tableau 9.2 cet identifiant est repris de manière à décrire la sémantique correspondante :

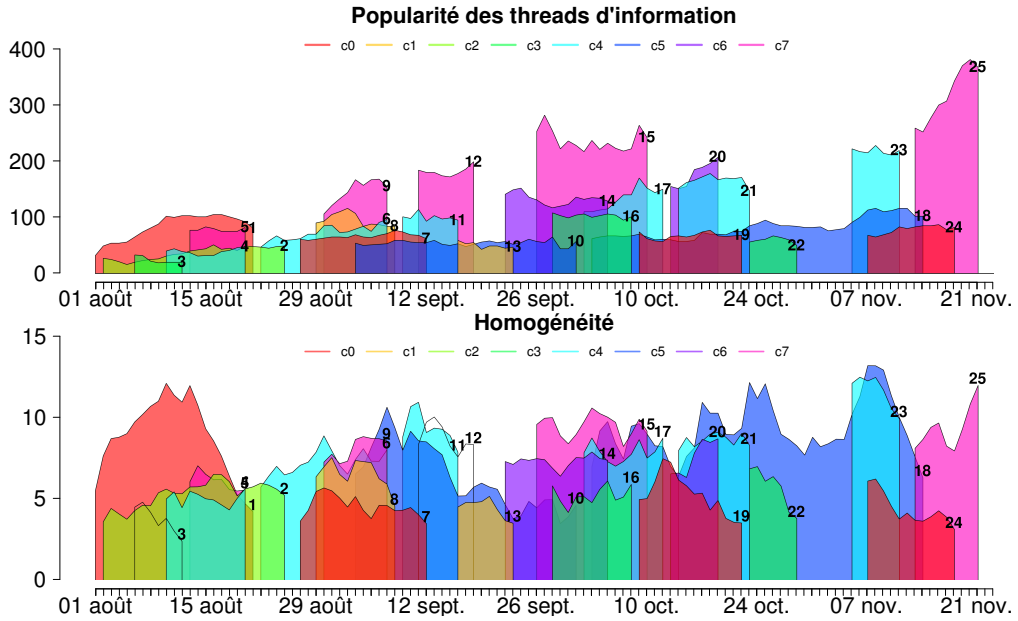


FIGURE 9.8 – Popularité des threads détectés pour $\delta_t^{\min} = 6$ jours. La popularité est mesurée en nombre de sources participant à un thread au temps t.

pour ce faire sont représentés les 8 termes les plus caractéristiques de la communauté durant la durée d'existence du thread identifié. Ces termes correspondent aux 8 termes les plus fréquents durant la période d'existence du thread parmi ceux décrits quotidiennement par $\boldsymbol{d}_k(t)$.

Analyse globale Sur la figure nous constatons une représentativité variable des communautés, comme remarqué précédemment certaines ne sont que rarement associées à une thématique précise et produisent alors moins de threads. Sur le graphique du bas nous remarquons que les threads capturent correctement les pics en homogénéité que forment les communautés correspondantes. Au contraire, les périodes les plus populaires pour une communauté sont moins souvent associées à un thread correspondant. Il est effectivement plus difficile de rassembler une grande population autour d'une thématique précise. Sur le graphique du haut nous relevons enfin que les périodes marquées par un fort dynamisme sont celles durant lesquelles de nombreux threads sont formés en parallèle.

Analyses individuelles Nous observons sur la figure un découpage inégal des communautés. Tandis que certaines ne se stabilisent que de manière ponctuelle, d'autres forment plusieurs threads bien répartis sur l'ensemble de la période, et d'autres encore sont gouvernées par une thématique remarquablement stable durant de longs intervalles de temps. Les communautés évoluent ainsi de manière spécifique, leurs comportements doivent être étudiés de manière individuelle.

En examinant les sémantiques fournies dans le tableau 9.2, nous constatons que pour certaines communautés les threads constituent des déclinaisons d'une même thématique générale, c'est en particulier le cas pour les communautés $c_4, c_5, c_6, \text{et} c_7$. A ces communautés nous associons un étiquetage manuel, fourni à la dernière ligne du tableau. Pour les autres en revanche, les threads identifient des changements complets de directions, les communautés correspondantes n'exhibent pas de sémantique particulière sur l'ensemble

c0	c1	c2	c3	c4	c5	c6	c7
1	8	2	3	4	10	14	5
jo	rentrée	syrie	curiosity	hollande	obama	champion	année
londre	ministre	alep	robot	syrie	romney	pari	euro
france	france	article	planète	président	barack	ligue	france
jeu	pari	onu	nasa	article	etats-unis	match	prix
champion	peillon	personne	vie	ump	mitt	france	million
médaille	retour	rapport	rover	ministre	président	tennis	ministre
équipe	jour	rebelle	arrivée	rapport	candidat	face	trimestre
finale	série	combat	article	sarkozy	campagne	finale	banque
7	13		16	6	18	20	9
jeu	image		france	ministre	obama	match	prix
londre	photo		monde	hollande	rapport	monde	groupe
france	france		tour	rapport	romney	pari	ministre
para	kate		cyclisme	article	article	france	euro
sport	william		champion	france	candidat	tennis	président
record	prince		formule	gouvernement	barack	équipe	baisse
soir	closer		président	rentrée	campagne	affaire	france
médaille	chanteur		schumacher	ump	mitt	armstrong	gouvernement
19			22	11			12
prix			france	hollande			france
nobel			tour	ump			hollande
année			armstrong	france			euro
quinté			cyclisme	homme			groupe
littérature			image	président			ministre
monde			match	article			psa
france			lance	gouvernement			etats-unis
livre			dopage	rapport			iphone
24				17			15
prix				france			france
goncourt				rapport			pari
roman				article			prix
essai				gouvernement			euro
livre				traité			groupe
écrivain				pari			ministre
f1				ump			million
auteur				hollande			année
				21			25
				hollande			euro
				pari			france
				président			pari
				débat			prix
				france			groupe
				ministre			million
				mort			entreprise
				rapport			pays
				23			
				pari			
				rapport			
				compétitivité			
				france			
				ministre			
				euro			
				gouvernement			
				hollande			
-	-	-	-	**politique**	**élection us**	**sport**	**économie**

TABLE 9.2 – Sémantique associée aux threads identifiés sur les $K = 8$ communautés pour $\delta_t^{\min} = 6$ jours.

de la période.

On peut observer le thread 1 qui est produit par la communauté c_0 à laquelle aucune sémantique globale n'est accordée, sa thématique porte sur les jeux olympiques d'été de Londres. Il constitue le thread le plus populaire et le plus homogène durant sa période d'existence, il se démarque des autres threads de manière remarquable. Les threads 10 et 18 sont tous deux produits par la communauté c_5 qui couvre les élections américaines aux Etats-Unis. Ceux-ci constituent les threads les plus longs et bien qu'ils ne soient pas les plus populaires, ils atteignent tous deux plusieurs pics remarquables en homogénéité. Dans la suite nous étudions plus en détail ces deux communautés ainsi que leurs threads correspondants.

9.3.2.2 Jeux olympiques d'été à Londres

Description des évènements Les jeux olympiques d'été se sont déroulés à Londres : ils ont débutés le 27 juillet par la cérémonie d'ouverture et se sont terminés le 12 août (au

soir) par la cérémonie de clôture. Les jeux se sont poursuivis par les jeux paralympiques dont les cérémonies d'ouverture et de clôture se sont déroulées respectivement le 29 août et le 9 septembre.

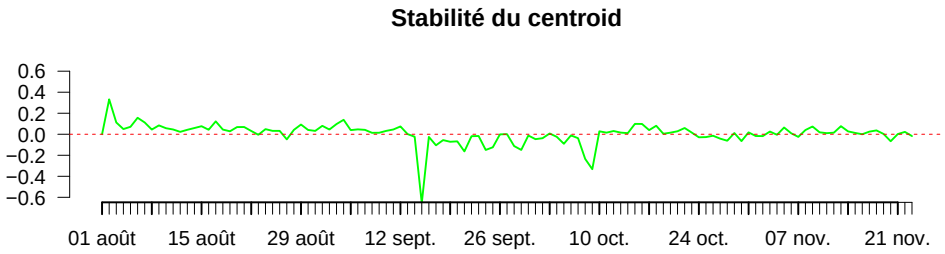
Threads détectés Dans la partition, la communauté c_0 rassemble une population qui couvre l'évènement des jeux olympiques durant l'été 2012. Cette dernière produit en effet les threads 1 et 7, représentés sur la figure 9.8 et dont la sémantique est fournie dans le tableau 9.2. La cérémonie d'ouverture des jeux se situe en dehors de la période d'étude et le thread 1 connaît une popularité croissante jusque mi-août, peu après quoi il disparaît. Sur le graphique du bas, nous constatons de même une baisse importante et soudaine de l'homogénéité du thread peu après cette date. La communauté ne produit pas de nouveau thread avant le 27 août, deux jours avant l'ouverture des jeux para-olympiques. Le thread 7 connaît alors une popularité ainsi qu'une homogénéité comparable sur l'ensemble de son existence, avant de se terminer le 12 septembre un peu après la date de clôture des épreuve paralympiques.

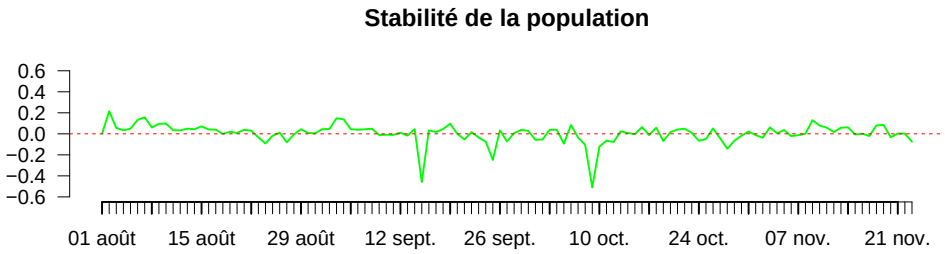
Sur le graphique du haut nous constatons que les épreuves olympiques constituent l'évènement le plus populaire durant la durée de leur existence, de même, sur le graphique du bas nous constatons que cet évènement atteint un pic remarquable en homogénéité. Les épreuves olympiques dominent ainsi l'actualité française durant l'été 2012. Les épreuves paralympiques jouissent, elles, d'une moins grande popularité ainsi que d'une moins grande homogénéité : ces épreuves sont effectivement plus récentes et se déroulent de plus durant la rentrée scolaire, comme évoqué précédemment, période marquée d'un plus grand dynamisme.

Evolutions temporelles Sur le graphique du haut de la figure 9.9(a), nous avons représenté pour la communauté c_0 les écarts quotidiens à la moyenne de son alignement $\cos\theta_0(t) - \cos\bar{\theta}(t)$ où $\cos\theta_0(t)$ décrit l'alignement de c_0 sur l'intervalle $[t-1, t]$ et $\cos\bar{\theta}(t)$ est l'alignement moyen durant ce même intervalle. Sur le graphique du bas de cette même figure, nous avons représenté les écarts quotidiens à la moyenne de la stabilité de sa population $\tau_0(t) - \bar{\tau}(t)$ où $\tau_0(t)$ est la stabilité de sa population sur l'intervalle $[t-1, t]$ et $\bar{\tau}(t)$ est la stabilité moyenne sur ce même intervalle. Pour chacun, une valeur positive indique une remarquable stabilité par rapport au reste des communautés : en particulier, les threads d'information correspondent aux périodes durant lesquelles l'écart en alignement est positif durant une durée minimale de $\delta_{\min} = 6$ jours.

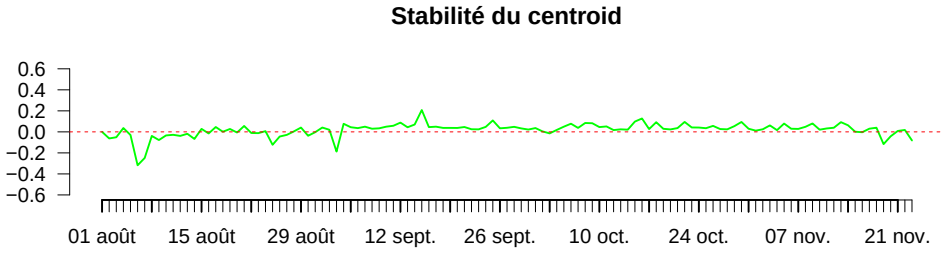
Lors de la période des jeux olympiques, nous observons que cette communauté regroupe une population remarquablement fidèle, qui publie de plus une information remarquablement similaire. A la suite de cette période, nous observons un changement de thématique soudain qui marque l'arrêt du thread 1 : les sources couvrant les jeux olympiques semblent alors se désintéresser totalement et dérivent vers de nouveaux sujets. Avec l'arrivée des jeux paralympiques la communauté, qui commençait à se dissoudre, se recompose [4] à nouveau autour d'un sujet fédérateur. Cette période de stabilité est alors suffisamment longue pour que l'évènement des paralympiques soit détecté comme le thread 7.

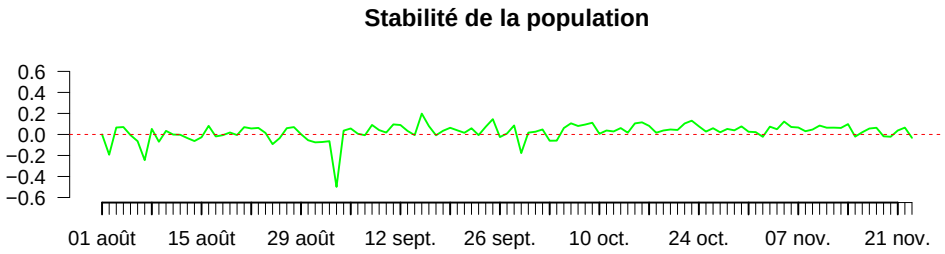
4. Bien qu'il soit visible que la communauté se recompose (i.e. que les sources lui restent plus fidèles), il serait intéressant de mesurer l'auto-corrélation de sa population pour voir en quelle mesure il s'agit des mêmes sources qui couvraient les épreuves olympiques.





(a) Jeux olympiques





(b) Elections US

FIGURE 9.9 – Ecarts quotidiens à la moyenne de la stabilité du représentant et écarts quotidiens à la moyenne de la stabilité de la population pour les communautés respectivement associées aux jeux olympiques et aux élections américaines.

9.3.2.3 Elections présidentielles américaines

Description des évènements Durant la fin de l'année 2012 se sont déroulés, les élections présidentielles américaines. Les derniers mois avant le 6 novembre 2012, date du scrutin, ont été marqués par la campagne électorale menée par les partis démocrate et républicain. Cette dernière a pris un tournant à la suite de la nomination officielle du candidat républicain et du candidat démocrate, durant la convention nationale organisée du 28 au 30 août par le parti républicain et du 4 au 6 septembre par le parti démocrate. A partir du 3 octobre, commence une nouvelle phase durant laquelle a été organisée une série de débats entre les deux candidats. Dans la nuit du 6 au 7 novembre 2012, la victoire du candidat démocrate marque la fin des élections.

Threads détectés Nous identifions deux threads associés à l'ensemble de ces évènements, il s'agit des threads 10 et 18. Le premier débute tout début septembre, un peu après l'annonce des candidats républicain et démocrate (voir figure 9.8), le second démarre le 3 octobre, date de début des débats.

En dépit de la portée internationale de ces élections, sur le graphique du haut de la figure 9.8 nous observons avec surprise que les évènements associés n'engendrent jamais d'emballement de popularité. Par opposition aux jeux olympiques, la France ne joue effectivement pas de rôle direct dans les élections américaines et il semblerait que dans les médias français, ce soient les affaires internes au pays qui reçoivent le plus d'attention (voir les threads $12, 15, 23$ et 25 par exemple).

Néanmoins, sur le graphique du bas de la figure 9.8 nous constatons que les threads 10 et 18 constituent les évènements les plus durables et les plus homogènes sur l'ensemble de la période, il semblerait que ces élections soient couvertes par un ensemble de sources spécialisées et solidaires de leur communauté. Nous relevons par ailleurs cinq pics remarquables d'homogénéité lors de la phase finale des élections. Une analyse détaillée montre que le premier et l'avant-dernier font respectivement suite à l'ouverture et à la fermeture des débats. Entre ces deux, paraissent deux annonces importantes : le 9 octobre le candidat républicain effectue une percée dans les sondages, huit jours plus tard, le candidat démocrate effectue un retour remarqué. Dans la nuit du 6 au 7 novembre 2012, ce dernier remporte les élections et le pic en homogénéité associé est le plus grand observé sur l'ensemble de la période.

Evolutions temporelles Sur la figure 9.9(b), nous observons que durant ses premières semaines d'existence, la communauté subit un fort dynamisme, avant de se stabiliser autour du 29 août lors de la nomination des candidats. Après cette date, plus de 80% des sources qui la rejoignent lui restent fidèles d'un jour à l'autre. De même, la thématique qui gouverne cette communauté ne subit que peu de changement de direction d'un jour à l'autre.

9.4 Conclusions

Nous avons réalisé une mise en œuvre expérimentale des méthodes proposées pour le partitionnement incrémental de sources qui publient des documents sur Internet à intervalles fréquents.

Dans ce but, nous avons constitué un corpus de sources dynamiques représentant aussi bien des médias que des blogs d'information français. Les documents publiés ont fait l'objet d'une collecte de cinq mois auprès des fils de syndication des principaux médias d'information français. Pour le problème considéré, nous avons constaté que les domaines

de publication ne constituent pas des sources suffisamment homogènes : nous avons alors exploité l'algorithme d'identification de sources incrémentale présenté au chapitre 6 pour décomposer ces derniers en des ensembles de sources plus homogènes.

Nous avons motivé le choix de l'algorithme des K-moyennes ellipsoïdales pour le partitionnement des sources représentés dans un espace de description textuel. Nous avons d'une part constaté que le nombre de sources est toujours largement inférieur au nombre de descripteurs pour les décrire. Nous avons d'autre part remarqué que de nombreuses régions de l'espace de description constituent du bruit dans un contexte dynamique. Nous avons alors montré l'intérêt du paramètre de parcimonie qui ajuste la quantité d'information émanant des communautés formées à tout moment.

Nous avons de plus observé le dynamisme inhérent au problème considéré au travers de deux mesures de qualité : la première consiste à examiner l'alignement des communautés dans l'espace de représentation, la seconde repose sur les transitions effectuées par les sources entre les communautés. Nous avons alors exploité ces deux mesures afin de réaliser une analyse préliminaire de l'influence des paramètres sur le dynamisme des partitions. Tandis que la demi-vie imposée aux données conditionne la nature du problème, le nombre de communautés recherchées joue un rôle moins essentiel. Nous avons de plus montré que le paramètre de parcimonie influe sur le nature des communautés identifiées. Contrairement au cas sphérique, dans le cas ellipsoïdal, les communautés s'alignent dans des régions spécifiques de l'espace de représentation, y restent stationnaires avant de se déplacer vers une nouvelle région. Elles adoptent alors un comportement en « dents de scie », plus à même de décrire les derniers changements observés en présence d'un fort dynamisme. Afin d'ajuster automatiquement ce comportement selon le dynamisme régissant les partitions, il serait intéressant d'étudier les outils de sélection automatique présentés au chapitre 6 pour le choix du paramètre s.

Nous avons enfin exploité le corpus constitué pour étudier les publications de la presse française. Nous avons ainsi extrait les threads d'information associés aux communautés obtenues d'après un jeu de paramètres fixé. Tandis que certaines communautés demeurent cohérentes sur de longs intervalles de temps, d'autres ne se stabilisent que de manière ponctuelle, et d'autres encore s'alignent à différents instants, répartis uniformément sur la période d'étude. Nous avons alors étudié les threads correspondants aux jeux olympiques de Londres ainsi que ceux associés aux élections américaines. Les résultats que nous obtenons sont compatibles avec le déroulement de ces évènements, ils fournissent de plus de précieuses informations sur leur couverture médiatique par les médias et les blogs d'information français.

Conclusions et perspectives

Contributions

Dans cette thèse nous nous sommes intéressé aux problématiques liées à la représentation et à l'apprentissage à partir de textes, à la fois pour des informations émotionnelles et pour des informations dynamiques. Nous organisons nos contributions selon trois axes : le premier porte sur la représentation des données textuelles, le second sur les méthodes et les algorithmes d'apprentissage, enfin le dernier concerne le traitement du dynamisme en apprentissage.

Représentation des textes

Nous avons proposé et étudié la construction de représentations adaptées à l'étude des émotions dans les textes, à ce titre nous avons également considéré le problème de la description des émotions en vue de leur analyse automatique dans les textes. Nos contributions se divisent en trois axes selon que la représentation des documents repose sur des descripteurs bas niveau, sémantiques, ou sur la combinaison de ces deux types de représentation.

Descripteurs bas niveau pour les informations émotionnelles Au chapitre 3 nous avons étudié la constitution d'un espace de représentation bas niveau pour une tâche de discrimination des émotions. A cet effet nous avons considéré l'utilisation de p-grammes pour des ordres $p \geq 1$ mot(s). Nous avons montré que pris de manière isolée, les unigrammes représentent des descripteurs génériques qui exhibent une bonne couverture des documents, tandis que les bigrammes capturent des constructions plus rares mais plus précises et permettent notamment de modéliser une information plus complexe tenant naturellement compte des marqueurs de négation ou des marqueurs d'intensité linguistique par exemple. Sur les données étudiées, nous avons également montré que les trigrammes modélisent généralement une information trop fine et ne permettent pas de décrire correctement les émotions considérées, à l'exception de quelques unes pour lesquelles des constructions plus riches semblent caractéristiques.

Nos résultats rejoignent ainsi ceux de l'état de l'art sur le fait que la pertinence des descripteurs demeure spécifique aux émotions considérées : la taille du fossé émotionnel entre les informations bas niveau et les concepts cibles varie selon les problèmes. Nous avons montré qu'un vocabulaire simple permet d'en caractériser certaines comme l'*amour*, pour d'autres en revanche il est nécessaire de disposer d'une représentation plus complexe.

Nous avons alors proposé d'exploiter des mélanges de p-grammes pour décrire les documents de manière plus exhaustive, en considérant une stratégie de fusion anticipée qui consiste à concaténer les descripteurs associés à différents ordres. A ce titre, nous avons proposé d'éliminer les descripteurs les moins pertinents pour chacune des émotions considérées

en examinant leurs gains en information respectifs. En effet, pour décrire des concepts émotionnels le vocabulaire bas niveau est peu filtré, or la rareté des marqueurs d'émotions fait obstacle aux méthodes d'apprentissage régularisé. Le filtre que nous proposons permet d'éliminer l'influence des p-grammes vides, qui ne portent pas de sémantique particulière pour les émotions considérées, que ces derniers soient fréquents ou non. L'intérêt du filtre proposé a été observé au chapitre 3 sur un corpus de données réelles étiquetées selon une catégorisation des émotions.

Sur ces mêmes données, nous avons montré que la fusion des unigrammes et des bigrammes permet une légère amélioration des performances en moyenne : il apparaît que chacune des représentations d'origine doit décrire une information pertinente et différente pour que leur fusion améliore la qualité de la représentation faite des textes. En analysant les descripteurs les plus discriminants pour les émotions les plus fréquentes du corpus, nous avons montré que dans le cadre de leur fusion, les unigrammes constituent des descripteurs génériques mais peu discriminants, ils semblent améliorer le rappel des classifieurs ; tandis que les bigrammes produisent des descripteurs plus spécifiques mais moins nombreux, ils semblent améliorer leur précision. Nous avons de plus montré que les bigrammes les plus discriminants sont notamment composés de mots simples peu discriminants.

Descripteurs sémantiques pour les informations émotionnelles Nous avons proposé d'organiser les principaux modèles de représentation issus des travaux en psychologie et en linguistique pour décrire les émotions selon trois composantes de gradualité : la *composition*, qui vise à combiner des états basiques en des états transitoires et permet de modéliser des émotions complexes exprimant un état imprécis, l'*intensité*, qui permet de différencier les états platoniques des états passionnés, et l'*héritage*, qui permet de préciser la sémantique d'une émotion en la déclinant. Ces trois composantes peuvent être mises en œuvre lors d'un enrichissement sémantique des documents, ils permettent notamment de décrire finement les émotions en vue des nombreuses imprécisions et subtilités du langage.

Nous avons proposé une modélisation des émotions adaptée à leur analyse automatique dans les textes. Le modèle que nous proposons est motivé par les travaux en psychologie et en linguistique. Il repose sur une catégorisation floue des émotions : un état affectif est décrit par un vecteur d'appartenance à des états basiques. Comme rappelé au chapitre 5, ce modèle permet de décrire des états transitoires comme des combinaisons d'états primaires ainsi que des états imprécis qui connotent des mélanges d'états primaires. Une notion d'intensité est de plus définie pour préciser la charge émotionnelle des états. La richesse du modèle proposé est adaptable, nous avons en particulier décrit trois tâches classiques de l'état de l'art comme des instanciations de ce dernier.

Nous avons étudié l'utilisation du modèle proposé pour un enrichissement sémantique des documents, dans ce cadre nous avons proposé un ensemble de spécifications pour la constitution de ressources linguistiques associant aux mots d'un vocabulaire générique, l'un des états affectifs décrit par le modèle. Nos recommandations visent à adapter la modélisation faite des émotions à la granularité des mots et des groupes de mots annotés par des outils linguistiques. Ainsi, en plus des caractéristiques proposées par le modèle, nous proposons l'emploi d'un marqueur d'ambiguïté pour traiter de la polysémie du langage, ainsi que d'un marqueur de négation visant notamment à déléguer le traitement des négations non résolues à l'apprentissage mis en œuvre.

En supposant que soit disponible un lexique sémantique reposant sur la modélisation proposée pour décrire les émotions ainsi qu'une fonction pour extraire des annotations sémantiques d'un document, nous avons proposé un ensemble de descripteurs sémantiques tirant parti de la richesse du modèle proposé. Ces descripteurs sont formés comme des

agrégats spécifiques aux caractéristiques proposées par le modèle ainsi qu'aux concepts étudiés. Ils sont de plus obtenus comme le résultat d'un processus de synthèse pertinent qui tient compte des hypothèses qu'il est possible de faire sur les mécanismes en jeu lors de l'expression écrite des émotions. En particulier, nous avons proposé de modéliser l'accumulation des caractéristiques au travers d'opérateurs d'agrégation présentant des propriétés de renforcement, nous avons de plus proposé de tenir compte d'informations de position et de dispersion des annotations sémantiques dans les documents.

Représentation bas niveau enrichie de descripteurs sémantiques pour des informations émotionnelles Nous avons proposé la construction d'un espace de représentation enrichi mettant en œuvre une fusion anticipée de descripteurs sémantiques et de descripteurs bas niveau. Pour ce faire nous avons extrait les descripteurs sémantiques proposés au chapitre 5 à partir des annotations linguistiques fournies sur le corpus étiqueté dans le cadre du projet DoXa, et nous avons comparé les performances d'une frontière de décision linéaire pour chacune des représentations prises de manière isolée ainsi que pour des combinaisons de ces dernières. Nous avons observé que pour les données étudiées les descripteurs sémantiques offrent une meilleure représentation. Nous avons de plus constaté que les frontières de décision induites bénéficient d'une plus grande robustesse lorsque combinée à des descripteurs bas niveau, à condition que les vecteurs de représentation soient normalisés dans leurs espaces d'origine.

Méthodes et algorithmes d'apprentissage

Nous avons considéré deux paradigmes d'apprentissage, supervisé et non supervisé, respectivement pour la discrimination de concepts émotionnels et pour le clustering de sources dynamiques. Pour chacun, les données sont représentées dans un espace textuel, presque vide et de très grande dimensionalité. Nous avons également considéré une tâche de caractérisation, assimilée à un apprentissage non supervisé dans un espace de représentation non textuel.

Caractérisation fine de la charge émotionnelle d'un texte Au chapitre 4, nous avons proposé une méthode pour caractériser finement la charge émotionnelle portée dans les textes. Notre approche consiste à mettre en œuvre un enrichissement sémantique des documents à partir d'un lexique associant aux mots d'un vocabulaire générique des coordonnées dans un espace sémantique multidimensionnel. Un document est alors représenté comme un nuage de points dans cet espace, sa charge émotionnelle est caractérisée par les propriétés géométriques et statistiques de cet ensemble. Nous avons considéré deux cadres d'étude : dans le premier nous discriminons les états affectifs de deux textes chargés émotionnellement, dans le second nous appliquons la méthode proposée aux dialogues d'un film afin d'observer l'évolution temporelle de sa charge émotionnelle. Les résultats obtenus ont motivé le modèle de représentation proposé pour décrire les émotions dans les textes, ils ont également montré l'intérêt d'un enrichissement sémantique associé à une description fine et graduelle des émotions.

Apprentissage supervisé Nous avons proposé, au chapitre 3, une approche pour la discrimination de concepts émotionnels dans les textes à partir de descripteurs bas niveau. La stratégie d'apprentissage consiste à mettre en œuvre un ensemble de classifieurs linéaires spécifiques à chacune des émotions considérées dans un paradigme « un contre tous » : un classifieur initial discrimine les documents neutres de ceux chargés émotionnellement, puis

à opposer successivement chaque émotion à toutes les autres pour chaque sous-problème binaire de manière à identifier le vocabulaire qui distingue chacune des émotions de toutes les autres.

Une implémentation de cette approche a fait l'objet d'une participation à la compétition I2B2 (*track2*) (Pestian et al., 2012), le système proposé s'est bien classé parmi les approches dépourvues d'enrichissements sémantiques. Comme rappelé précédemment, le système proposé se distingue sur trois points : il met en œuvre un système de décision en deux étapes pour discriminer les documents neutres de ceux chargés émotionnellement, il emploie une stratégie de fusion anticipée pour décrire les documents comme des mélanges de p-grammes pour des ordres croissants, et il effectue un filtrage des descripteurs spécifique à chacune des émotions considérées afin d'éliminer l'influence des moins discriminants.

Nous avons également considéré des tâches de classification dans le cadre du projet DoXa : nous avons proposé une mise en œuvre expérimentale du modèle proposé sur des données nouvelles, en collaboration étroite avec les partenaires linguistes du projet. Sur le corpus, nous avons réalisé de nombreuses expérimentations visant à étudier la pertinence des algorithmes d'apprentissage mis en œuvre. De nos analyses (non reportées dans cette thèse) il ressort que les frontières de décisions linéaires offrent un meilleur cadre que les frontières non linéaires, respectivement obtenues par un arbre de décisions et une mesure de similarité gaussienne.

Apprentissage non supervisé Au chapitre 8, nous avons proposé un algorithme de partitionnement pour des représentations textuelles qui repose sur les K-moyennes sphériques et qui effectue une pondération des descripteurs. Nous avons en particulier proposé une transformation qui change l'espace d'entrée en un ellipsoïde sur lequel les dimensions de l'espace originel sont dilatées ou contractées selon leur capacité à exposer une structure de clusters dans les données. Nous avons dérivé une solution analytique pour la nouvelle fonction objectif correspondante et nous avons proposé l'algorithme *ellkm* dont la complexité est similaire à celle des K-moyennes traditionnelles. Nous avons aussi proposé une procédure d'ajustement automatique pour le paramètre de parcimonie qui contrôle le degré d'aplatissement des ellipsoïdes. Sur des données synthétiques et sur des données réelles, nous avons montré l'intérêt de l'algorithme *ellkm* : dans le cas où le nombre de documents est largement inférieur au nombre de descripteurs pour les décrire, différentes mesures de qualités ont montré la pertinence de la méthode proposée. Dans un cadre plus classique pour lequel plus de documents sont disponibles, la procédure de sélection automatique conserve l'espace de représentation originel et les résultats obtenus sont équivalents à ceux de l'état de l'art.

Sur Internet, l'ensemble des documents publiés sur un domaine constitue souvent une information très hétérogène. Au chapitre 7, nous avons proposé de décomposer un domaine en un sous-ensemble de sources homogènes. Dans un premier temps, nous avons défini une source comme un regroupement d'urls représentées comme des ensembles de tokens. Dans un second temps nous avons proposé de réaliser un partitionnement hiérarchique des urls associées aux documents publiés. Le dendogramme obtenu contient alors un ensemble de sources homogènes sur le domaine. Nous avons étudié des dendogrammes de deux types : pour le premier les partitions respectent un critère de cohérence, pour le second elles constituent des regroupements compacts d'urls. Nous avons aussi considéré deux modes pour sa constitution : d'une part, un mode de traitement par lots nécessite que l'ensemble des urls disponibles soient traitées en une fois, dans la seconde approche le dendogramme est formé de manière incrémentale, à mesure que de nouvelles urls sont présentées. Nous avons alors proposé deux méthodes d'identification de sources sur Internet, qui représentent

deux extrêmes pour ces deux axes, respectivement de construction de partitions cohérentes par lots et d'identification incrémentale de partitions compactes. Nous avons enfin réalisé une étude comparative sur des données réelles qui montre l'intérêt des systèmes proposés.

Traitement du dynamisme

Au chapitre 6, nous avons proposé d'étudier une nouvelle tâche pour le partitionnement de données dynamiques dans le cadre de représentations textuelles. Un ensemble de sources dynamiques publie des documents à intervalles de temps fréquents. Afin de caractériser la sémantique associée aux communautés identifiées nous proposons de partitionner les sources dans l'espace de description des documents qu'elles publient. Nous avons alors proposé de synthétiser la totalité de l'information produite par une source depuis sa création au travers de son vecteur de publication. Après normalisation, ce dernier indique à tout moment une direction de l'espace de description qui tient compte à la fois des documents publiés à l'état courant et de l'historique de publication ajusté par un vieillissement paramétré. Pour ce faire, l'ensemble des traitements proposés ne requièrent qu'une utilisation linéaire de la mémoire avec le nombre de sources considéré, et autorisent de plus un mode de fonctionnement incrémental.

Au chapitre 9, nous avons réalisé une mise en œuvre expérimentale visant à évaluer l'ensemble des méthodes proposées dans la partie II d'une part et à étudier les publications de la presse française sur Internet d'autre part. Nous avons ainsi collecté, durant cinq mois, les publications quotidiennes des blogs et des principaux médias d'information en France : au total nous avons réuni un peu plus de 130 000 documents publiés sur 230 domaines.

A partir du corpus de sources dynamiques constitué en mettant en œuvre les propositions faites aux chapitres 6 et 7 sur les données collectées, nous avons réalisé une étude préliminaire sur le rôle des paramètres pour le dynamisme des partitions. Pour ce faire nous avons proposé d'observer ce dynamisme à la fois au niveau de l'information produite en continu par les sources et au niveau des transitions effectuées par ces dernières entre les communautés. De nos analyses, il ressort que le vieillissement imposé aux données est la première cause de dynamisme.

Le nombre de communautés recherchées présente, lui, une influence non négligeable, et le taux de parcimonie employé pour contrôler la quantité d'information émanant des communautés permet, en cas de fort dynamisme, d'identifier des communautés plus cohérentes en limitant leurs mouvements à des régions denses de l'espace de description.

Nous avons de plus étudié la couverture médiatique des jeux olympique d'été de Londres ainsi que les élections américaines par la presse française. Pour ce faire nous avons extrait les threads d'information associés aux communautés obtenues d'après un jeu de paramètres fixé. Nous avons alors montré que les threads identifiés correspondent à l'agencement de ces évènements.

Perspectives

Nous présentons ici les perspectives de ces travaux de thèse en considérant successivement celles qui concernent les informations émotionnelles et celles qui portent sur les informations dynamiques.

Informations émotionnelles

Nous avons exposé les difficultés liées à l'étiquetage de corpus selon des concepts émotionnels, une première perspective concerne la constitution d'un corpus étiqueté en

émotions qui permettrait d'évaluer plus en profondeur la pertinence des propositions faites dans le modèle proposé pour décrire les émotions dans les textes. De plus pour le modèle proposé, nous avons constaté les difficultés liées à la constitution manuelle d'un lexique affectif : la richesse du modèle proposé permet de décrire finement les émotions, sa complexité peut s'avérer un obstacle dans un cadre d'utilisation manuel. Une seconde perspective vise à une expansion automatique d'enrichissements sémantiques basés sur le modèle. Enfin, pour cette première partie, ces travaux de thèse ont été guidés par la difficulté du passage d'une information bas niveau à une information émotionnelle, une dernière perspective ouvre la voie vers de nouveaux procédés de combinaisons.

Construction d'un corpus émotionnel Nous avons exposé les difficultés liées à l'étiquetage de corpus selon des concepts émotionnels pour les corpus exploités dans ces travaux : le corpus I2B2 présente un grand déséquilibre entre les classes et le corpus DoXa contient peu de documents comparé à la difficulté de l'apprentissage des concepts considérés. Nous souhaiterions évaluer l'intérêt des propositions faites au chapitre 5 sur un nouveau corpus. Une piste consisterait à étudier l'applicabilité des méthodes proposées sur le corpus I2B2 sur lequel nous disposons par ailleurs des performances obtenues par de nombreuses méthodes très variées pour la discrimination des émotions dans les textes (Pestian et al., 2012). Une autre voie consisterait aussi a constituer un nouveau corpus étiqueté en émotions de manière similaire à Pak et Paroubek (2010). Ce corpus peut alors faire l'objet d'une participation à l'organisation d'une compétition pour la classification des émotions dans les textes.

Expansion automatique de ressources sémantiques Bien que la richesse du modèle permette de décrire finement les états émotionnels et ainsi de constituer une information plus riche en vue d'une analyse automatique dans les textes, nous avons constaté que sa complexité peut constituer un obstacle dans un cadre d'instanciation manuel. Nous souhaitons étudier l'adaptation des méthodes automatiques d'expansion de ressources sémantiques présentées à la section 2.3.2, p. 47 pour la constitution d'un lexique affectif qui associe aux entrées du vocabulaire l'un des états affectifs décrits par le modèle. En particulier, la base *wordnet* permet de propager des connaissances initiales à un vocabulaire plus large, des méthodes de partitionnement appliquées à des corpus de données génériques peuvent être employées pour étendre davantage ce vocabulaire. L'étude de l'applicabilité de ces méthodes pour le modèle proposé nécessite un travail de synthèse plus approfondi, à l'issue duquel il est possible qu'une simplification du modèle soit également nécessaire.

Combinaisons de descripteurs hétérogènes pour décrire les documents Nous avons successivement montré l'intérêt des représentations bas niveau et des représentations sémantiques pour décrire les documents en vue d'une analyse des émotions dans les textes. Au chapitre 5 nous avons de plus proposé d'exploiter conjointement ces deux types de représentation pour discriminer les émotions dans les documents. Nous souhaitons poursuivre cette voie et construire cet espace en prenant mieux en compte la nature spécifique des descripteurs d'origine.

En particulier, le vocabulaire bas niveau capture une information proche des données étudiées, les descripteurs correspondants présentent une bonne couverture des textes et des combinaisons de ces derniers permettent naturellement de tenir compte de constructions plus complexes, nécessaires pour les concepts émotionnels.

De plus, les descripteurs sémantiques modélisent une information riche et haut niveau qui comble en partie le fossé entre le bas niveau et l'émotionnel. Ces descripteurs décrivent

des états affectifs au niveau des mots et des groupes de mots, dans le modèle proposés ils leur associent notamment un degré d'appartenance aux états primaires ainsi qu'une intensité. Pour répondre aux nombreuses ambiguïtés et subtilités du langage, les descripteurs sémantiques, couplés aux descripteurs bas niveau, guident l'apprentissage pour des concepts émotionnels et permettent de pallier la rareté des marqueurs d'émotions dans les corpus.

Néanmoins, la combinaisons de ces deux modes de représentation représente un défi, les descripteurs correspondants sont en effet très hétérogènes, tant en nombre qu'en nature. Tandis que les espaces de représentation bas niveau sont généralement presque vides en très grande dimension, les espaces de représentation sémantiques sont bien moins larges et n'exhibent pas la même parcimonie.

De plus, contrairement aux descripteurs bas niveau, les descripteurs sémantiques caractérisent les émotions sur des échelles numériques, par exemple autour d'une valeur spécifique d'intensité ou autour d'une valeur donnée d'appartenance aux états primaires. Or comme rappelé au chapitre 1, la mesure de comparaison employée pour réaliser un apprentissage dépend grandement de la nature des données étudiées et le produit scalaire, adapté aux descripteurs bas niveau, ne s'active que pour des valeurs élevées de ses arguments. Pour les descripteurs sémantiques, il est alors nécessaire d'employer une mesure permettant de caractériser des concentrations de données autour de certaines valeurs comme par exemple le noyau gaussien. De plus, nous avons observé au chapitre 3 que selon les émotions considérées les meilleures stratégies de représentation diffèrent, nous souhaitons pouvoir pondérer les différentes représentations faites des documents selon leur pouvoir de discrimination pour les concepts considérés.

Dans ce contexte, une perspective vise à mettre en œuvre une stratégie de fusion intermédiaire pour constituer un espace de représentation final à la fois fidèle aux données étudiées et proche des concepts considérés. Les méthodes d'apprentissage par noyaux multiples rappelées à la section 1.3.4.3, p. 32 et présentées plus en détail en annexe A, p. 197, offrent alors un cadre d'étude naturel et prometteur.

Informations dynamiques

Les perspectives associées à nos travaux sur les informations dynamiques peuvent être organisées autour de quatre axes.

Extensions pour l'identification de sources sur Internet Nous avons proposé deux méthodes pour l'identification de sources sur Internet à partir du dendogramme des urls publiées sur un domaine de publication. Pour l'élaboration de ce dendogramme, la première effectue un traitement par lots et identifie un partitionnement cohérent, la seconde réalise un traitement incrémental et identifie un partitionnement compact. Les données employées pour évaluer ces deux méthodes ne permettant néanmoins pas de mettre en évidence leurs différences en termes de performances. Nous soupçonnons que les urls considérées ne reposent pas sur des constructions suffisamment « exotiques », pour lesquelles les différences entre les deux méthodes peuvent être appréciées. Ainsi, nous souhaitons étudier un jeu de données sur lequel nous pourrions comparer ces deux méthodes et étudier alors les cas intermédiaires qui sont respectivement une méthode d'identification incrémentale qui produit un partitionnement cohérent et une méthode de partitionnement compact par lots.

D'un point de vue formel, il serait de plus intéressant d'étudier dans quelle mesure tout partitionnement cohérent des urls peut être caractérisé par une relation d'ordre sur l'ensemble des tokens. De même nous souhaitons étudier la complémentarité entre les

critères définis pour la compacité et la cohérence des partitions d'urls.

Profils de publications pour des sources dynamiques Nous souhaiterions aller plus loin dans les analyses faites sur les communautés de sources identifiées sur les publications de la presse française. Nous considérons deux pistes de développement, la première consiste à étudier l'auto-corrélation entre les populations d'une même communauté découpée en différents threads afin de mieux décrire les mouvements de population. La seconde consiste à établir des profils de publication pour les sources en les décrivant d'après l'ensemble des threads auxquels elles ont participé sur la période. Nous pourrions ainsi caractériser différents comportements types, représentant par exemple les sources à l'origine des thématiques les plus populaires, celles qui suivent toujours le mouvement général, ou encore celles qui restent toujours fidèles aux mêmes thématiques. De même, comme proposé par Leskovec et al. (2009) ces analyses pourraient être remontées à un niveau plus global afin d'établir des profils de publication pour les média d'information d'une part et pour les blogs d'autre part.

Extensions de *ellkm* Nous envisageons trois extensions de l'algorithme des K-moyennes ellipsoïdales, nous présentons dans un premier temps des perspectives de travaux préalables visant à mieux comprendre les propriétés de l'algorithme proposé.

La procédure d'ajustement automatique pour le paramètre de parcimonie des K-moyennes ellipsoïdales nécessite le partitionnement d'un certain nombre de variantes aléatoires des données étudiées. Bien que nous ayons montré l'intérêt de la procédure proposée, nous souhaitons étudier d'autres pistes pour le choix du paramètre qui ne feraient pas intervenir de données externes. Une piste consiste par exemple à déterminer une expression analytique du paramètre à partir de la fonction objectif des K-moyennes ellipsoïdales.

De plus nous avons réalisé, sur des données synthétiques, une étude comparative de l'algorithme proposé à celui des K-moyennes sphériques en générant des données qui exhibent les caractéristiques des représentations textuelles. Banerjee et al. (2005) montrent cependant que les données textuelles suivent une loi de von Mises-Fisher qui décrit la distribution de vecteurs agencés sur une hypersphère. Comme nous souhaitons étudier et utiliser cette hypothèse dans le but de mieux contrôler les corpus générés et ainsi de mieux saisir les différences entre le cas ellipsoïdal et le cas sphérique.

Une première extension des K-moyennes ellipsoïdales consiste à étudier un partitionnement flou des données afin de décrire chacun des documents par un vecteur de degrés d'appartenance aux clusters identifiés. Cette extension réaliserait une généralisation ellipsoïdale des c-moyennes floues (Pal & Bezdek, 1995), plus robuste au bruit dans les données que les K-moyennes traditionnelles. De plus, en vue des analyses dynamiques, les degrés d'appartenance aux clusters identifiés apportent une information plus riche pour décrire les transitions effectuées par les sources entre les communautés.

Une seconde extension repose sur l'observation que l'algorithme des K-moyennes ellipsoïdales réalise une pondération et non une sélection des descripteurs : en ce sens, l'importance des dimensions non pertinentes est réduite à une valeur négligeable par rapport aux données mais elle n'est pas nécessairement annulée. Nous souhaitons établir les conditions dans lesquelles un descripteur est considéré comme non pertinent pour un cluster et qui permettraient ainsi d'effectuer une sélection des descripteurs. Dans cette même voie, nous souhaitons explorer une approche différente, consistant à définir un ensemble de contraintes parcimonieuses. Pour ce faire, une approche consisterait à examiner l'emploi d'une norme $l1$ lors de la définition des centroïdes, cette dernière présente en effet d'intéressantes propriétés de parcimonie.

Enfin, une autre extension, probabiliste, consiste à étudier les conséquences de l'hypothèse ellipsoïdale pour l'algorithme probabiliste proposé par Banerjee et al. (2005) comme généralisation des K-moyennes sphériques (Bouberima et al., 2010). Cette extension permettrait en particulier de mieux comprendre les conditions nécessaires à la mise en œuvre d'une réduction des dimensions pour le partitionnement de données textuelles. Pour ce faire, peuvent notamment être exploitées des données synthétiques générées d'après une distribution de von Mises-Fisher.

Dynamisme et émotions

Une autre perspective vise à combiner les deux thématiques traitées dans cette thèse, ce qui peut être envisagé de différentes manières et offre nombre de directions de recherche. En effet, ces deux problématiques s'entremêlent dans leurs cadres d'application respectifs : aussi les émotions et leurs moyens d'expression sont continuellement changeants et dynamiques, aussi leur étude sur les réseaux de communication nécessite des outils adaptés au dynamisme de l'information. Sur ces mêmes réseaux, les sujets fédérateurs sont souvent soumis à la partialité des sources (pour des thématiques politiques ou sportives par exemple), l'analyse des communautés formées autour de thématiques actuelles profiterait de méthodes adaptées à l'étude de la subjectivité, des opinions (Lansdall-Welfare et al., 2012) ou de manière plus générale des émotions dans les textes.

Une piste consiste par exemple à intégrer à la notion de thread d'information celle de subjectivité avec la prise en compte de descripteurs émotionnels. Il serait alors pertinent d'étudier dans quelle mesure il est possible d'intégrer dans la représentation faite des sources, les modèles émotionnels entraînés sur des données externes. Pour une même thématique, deux communautés peuvent ainsi correspondre à deux points de vue qui divergent dans les médias : il serait alors intéressant d'observer dans quelle mesure les threads d'information extraits tiennent compte des opinions et des émotions exprimées (Mitrović et al., 2011). Cet axe de recherche soulèvent néanmoins de nombreuses questions portant sur la fusion d'informations hétérogènes (Gönen & Alpaydin, 2011), l'apprentissage émotionnel multi-domaines (apprendre sur un corpus pour en étiqueter un nouveau) (Glorot et al., 2011), ainsi que sur la validité temporelle des marqueurs émotionnels (Dodds et al., 2011).

Bibliographie

Ackermanny, M. R., Lammersenz, C. & Märtensy, M. (2012). Streamkm++ : A clustering algorithm for data streams. *Journal of Experimental Algorithmics*, *17*.

Adelson-Velskii, M. & Landis, E. M. (1963). *An algorithm for the organization of information* (Technical Report).

Aggarwal, C., Han, J., Wang, J. & Yu, P. S. (2004). A framework for projected clustering of high dimensional data streams. *Int. Conf. on Very Large Data Bases, VLDB*.

Aggarwal, C., Wolf, J., Yu, P., Procopiuc, C. & Park, J. S. (1999). Fast algorithms for projected clustering. *ACM SIGMOD Record*, *28*, 61–72.

Aggarwal, C. C., Han, J., Wang, J. & Yu, P. S. (2003). A framework for clustering evolving data streams. *29th Int. Conf. on Very large data bases*.

Aggarwal, C. C. & Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data*. Springer US.

Bach, F. R., Lanckriet, G. R. G. & Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the smo algorithm. *Int. Conf. on Machine Learning*.

Balahur, A. & Montoyo, A. (2008). Applying a culture dependent emotion triggers database for text valence and emotion classification. *Procesamiento del lenguaje natural ISSN 1135-5948*, *40*, 107–114.

Banerjee, A., Dhillon, I. S., Ghosh, J. & Sra, S. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, *6*, 1345–1382.

Barrett, L. F. & Russell, J. A. (1999). The structure of current affect : Controversies and emerging consensus. *Current Directions in Psychological Science*, *8*, 967–984.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D. & Subrahmanian, V. (2007). Sentiment analysis : Adjectives and adverbs are better than adjectives alone. *Proc. of Int. Conf. on Weblogs and Social Media*.

Beringer, J. & Hüllermeier, E. (2006). Online clustering of parallel data streams. *Data & Knowledge Engineering*, *58*, 180–204.

Berners-Lee, T. (2005). *Request for comments : 3986* (Technical Report). Network Working Group.

Bestgen, Y., Fairon, C. & Kevers, L. (2004). Un baromètre affectif effectif. *Actes des journées internationales d'analyse statistique des données textuelles* (pp. 182–191).

Bifet, A. & Kirkby, R. (2009). *Data stream mining a practicle approach*. University of waikato.

Black, P. E. (2004). *Dictionary of algorithms and data structures*. National Institute of Standards and Technology.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blum, A. & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Conf. on Computational Learning Theory*.

Blum, A. L. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*, 245—-271.

Bottou, L. & Bengio, Y. (1995). Convergence properties of the k-means algorithms. *Advances in Neural Information Processing Systems*.

Bouberima, W. P., Nadif, M. & Bencheikh, Y. K. (2010). Assessing the number of clusters from a mixture of von mises-fisher. *World Congress on Engineering*.

Boucouvalas, A. C. (2003). Real time text-to-emotion engine for expressive internet communications. In *Being there : Concepts, effects and measurement of user presence in synthetic environments*. Ios Press.

Cambria, E., Speer, R., Havasi, C. & Hussain, A. (2010). Senticnet : A publicly available semantic resource for opinion mining. *Commonsense Knowledge : AAAI Fall Symposium*.

Cao, F., Ester, M., Qian, W. & Zhou, A. (2006). Density-based clustering over an evolving data stream with noise. *SIAM Conf. on Data Mining*.

Chesley, P. (2006). Using verbs and adjectives to automatically classify blog sentiment. *Proc. of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*.

Cortes, C. & Vapnik, V. (95). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cowie, R. & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Commun.*, *40*, 5–32.

Cowie, R., Douglas-Cowie, E. & Romano, A. (1999). Changing emotional tone in dialogue and its prosodic correlates. *Proc. of ETRW on Dialogue and Prosody*.

Crespo, F. & Weber, R. (2005). A methodology for dynamic data mining based on fuzzy clustering. *Fuzzy Sets and Systems*, *150*, 267–284.

Cristianini, N., Eliseeff, A. & Shawe-Taylor, J. (2002). On optimizing kernel alignement. *Neural Information Processing Systems (NIPS) Conf.*.

Cui, H., Mittal, V. & Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. *21st National Conf. on Artificial Intelligence*.

Damez, M., Lesot, M.-J. & d'Allonnes, A. R. (2012). Dynamic credit-card fraud profiling. In V. Torra, Y. Narukawa, B. López and M. Villaret (Eds.), *Modeling decisions for artificial intelligence*, vol. 7647 of *Lecture Notes in Computer Science*, 234–245. Springer Berlin Heidelberg.

Das, S. R. & Chen, M. Y. (2007). Yahoo! for amazon : Sentiment extraction from small talk on the web. *Management Science*, *53*, 1375–1388.

Delavallade, T. (2007). *Evaluation des risques de crise, appliquée à la détection des conflits armés intra-étatiques*. Thèse de doctorat, Université Pierre et Marie Curie.

Detyniecki, M. (2002). *Mathematical aggregation operators and their application to video querying*. Thèse de doctorat, Université Pierre et Marie Curie (UPMC).

Dhillon, I. & Modha, D. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*, 143–175.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network : Hedonometrics and twitter. *PloS one*, *6*, e26752.

Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M. & Papadopoulos, D. (2007). Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery*, *14*, 63–97.

Dray, G., Plantié, M., Harb, A., Poncelet, P., Roche, M. & Trousset, F. (2009). Opinion mining from blogs. *IJCISIM'09 : Int. Journal of Computer Information Systems and Industrial Management Applications*, *1*, 205–213.

Duthil, B., Trousset, F. & Dray, G. (2012). Vers une caractérisation automatique de critères pour l'opinion-mining. *Les Cahiers du numérique*, *7*, 41–62.

Dzogang, F., Lesot, M.-J., Rifqi, M. & Bouchon-Meunier, B. (2010a). Analysis of texts' emotional content in a multidimensional space. *Int. Conf. on Kansei Engineering and Emotion Research, KEER* (pp. 877–886).

Dzogang, F., Lesot, M.-J., Rifqi, M. & Bouchon-Meunier, B. (2010b). Expressions of graduality for sentiments analysis - a survey. *Int. Conf. on Fuzzy Systems, FUZZ-IEEE* (pp. 1–7).

Dzogang, F., Lesot, M.-J., Rifqi, M. & Bouchon-Meunier, B. (2012). Early fusion of low level features for emotion mining. *Biomedical Informatics Insights (suppl. 1)*, *5*, 129–136.

Dzogang, F., Marsala, C., Lesot, M.-J. & Rifqi, M. (2012b). An ellipsoidal k-means for document clustering. *Int. Conf. on data mining, ICDM* (pp. 221–230).

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Ekman, P. (1999). *Basic emotions*, chapter 3, 45–60. John Wiley.

Esuli, A. & Fabrizio, S. (2006). Sentiwordnet : A publicly available lexical resource for opinion mining. *5th Conf. on Language Resources and Evaluation (LREC'06).*

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. (2008). Liblinear : A library for large linear classification. *Journal of Machine Learning*, *9*, 1871–1874.

Fitrianie, S. & Rothkrantz, L. J. (2008). An automated online crisis dispatcher. *Int. Journal of Emergency Management*, *5*, 123–144.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B. & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science : a journal of the American Psychological Society*, *18*, 1050–7.

Friedman, J. H. & Meulman, J. J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society : Series B*, *66*, 815–849.

Gaertler, M., Görke, R., Wagner, D. & Wagner, S. (2006). *How to cluster evolving graphs* (Technical Report). Faculty of Informatics, Universitä Karlsruhe.

Glorot, X., Bordes, A. & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification : A deep learning approach.

Go, A., Bhayani, R. & Huang, L. (2009). *Twitter sentiment classification using distant supervision* (Technical Report). Stanford.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Gönen, M. & Alpaydin, E. (2008). Localized multiple kernel learning. *25th international conference on Machine learning, ICML.*

Gönen, M. & Alpaydin, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, *12*, 2211–2268.

Han, J., Pei, J. & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Int. Conf. on Management of Data*.

Hanczar, B. & Nadif, M. (2011). Using the bagging approach for biclustering of gene expression data. *Neurocomputing*, *74*, 1595–1605.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning*. Springer New York Inc.

Hatzivassiloglou, V. & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. *18th Conf. on Computational Linguistics* (pp. 299–305).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177–196.

Hüllermeier, E. (2011). Fuzzy sets in machine learning and data mining. *Applied Soft Computing*, *11*, 1493–1505.

Igel, C., Glasmachers, T., Mersch, B., Pfeifer, N. & Meinicke, P. (2007). Gradient-based optimization of kernel-target alignment for sequence kernels applied to bacterial gene start detection. *Transactions on Computational Biology and Bioinformatics, IEEE/ACM*, *4*, 216–226.

Jain, A. K. (2010). Data clustering : 50 years beyond k-means. *Pattern Recognition Letters*, *31*, 651–666.

Jing, L., Ng, M. K. & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, *19*, 1026–1041.

Johnson-Laird, P. N. & Oatley, K. (1989). The language of emotions : An analysis of a semantic field. *Cognition and Emotion*, *3*, 81–123.

Kalogeratos, A. & Likas, A. (2011). Document clustering using synthetic cluster prototypes. *Data & Knowledge Engineering*, *70*, 284–306.

Kawadia, V. & Sreenivasan, S. (2012). Online detection of temporal communities in evolving networks by estrangement confinemen. In *Bulletin of the american physical society*.

Ketchen, D. J. & Shook, C. L. (1996). The application of cluster analysis in strategic management research : an analysis and critique. *Strategic management journal*, *17*, 441–458.

Kittler, J., Hatef, M., Duin, R. P. & Matas, J. (1998). On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *20*, 226–239.

Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.-R. & Zien, A. (2009). Efficient and accurate lp-norm multiple kernel learning. *Advances in Neural Information Processing Systems, NIPS*, *22*, 997–1005.

Koht-arsa, K. & Sanguanpong, S. (2001). In-memory url compression. *National Computer Science and Engineering Conf.*.

Kranen, P., Assent, I., Baldauf, C. & Seidl, T. (2011). The clustree : indexing micro-clusters for anytime stream mining. *Knowledge and Information Systems*, *29*, 249–272.

Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E. & Jordan, M. I. (2004a). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, *5*, 27–72.

Lanckriet, G. R. G., Bie, T. D., Cristianini, N., Jordan, M. I. & Noble, W. S. (2004b). A statistical framework for genomic data fusion. *Bioinformatics*, *20*, 2626–2635.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Lansdall-Welfare, T., Lampos, V. & Cristianini, N. (2012). Effects of the recession on public mood in the uk. *UK. Mining Social Network Dynamics (MSND) session on Social Media Applications in News and Entertainment (SMANE) at WWW*.

Lei, T., Cai, R., Yang, J.-M., Ke, Y., Fan, X. & Zhang, L. (2010). A pattern tree-based approach to learning url normalization rules. *Int. Conf. on World Wide Web*.

Leleu, S. (1987). Un atlas sémantique de concepts d'émotions : normes et validation. Master's thesis, Université catholique de Louvain, faculté de psychologie et des sciences de l'éducation.

Leskovec, J., Backstrom, L. & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *ACM Int. Conf. on Knowledge Discovery and Data Mining* (pp. 497–506).

Lesot, M.-J., Rifqi, M. & Benhadda, H. (2009). Similarity measures for binary and numerical data : a survey. *Int. Journal of Knowledge Engineering and Soft Data Paradigms*, *1*, 63–84.

Leung, C. K.-S. & Khan, Q. I. (2006). Dstree : A tree structure for the mining of frequent sets from data streams. *Int. Conf. on Data Mining*.

Liao, T. W. (2005). Clustering of time series data - a survey. *Pattern Recognition*, *38*, 1857–1874.

Liu, T., Liu, S., Chen, Z. & Ma, W.-Y. (2003). An evaluation on feature selection for text clustering. *Int. Conf. on Machine Learning*.

Liu, Z., Yu, J. X., Ke, Y., Lin, X. & Chen, L. (2008). Spotting significant changing subgraphs in evolving graphs. *Int. Conf. on Data Mining*.

Martin, J. C., Niewiadomski, R., Devillers, L., Buisine, S. & Pelachaud, C. (2006). Multimodal complex emotions : gesture expressivity and blended facial expressions. *Int. Journal of Humanoid Robotics (IJHR)*, *3*, 269–291.

Mathieu, Y. Y. (2006). A computational semantic lexicon of french verbs of emotion. In *Computing attitude and affect in text : Theory and applications*. Springer Netherlands.

Mei, Q., Ling, X., Wondra, M., Su, H. & Su, H. (2007). Topic sentiment mixture : modeling facets and opinions in weblogs. *16th Int. Conf. on World Wide Web*.

Mejova, Y. & Srinivasan, P. (2011). Exploring feature definition and selection for sentiment classifiers. *Int. AAAI Conf. on Weblogs and Social Media*.

Melville, P., Wojciech, G. & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. *KDD '09 : 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (pp. 1275–1284). New York, NY, USA : ACM.

Michel, B. S., Nikoloudakis, K., Reiher, P. & Zhang, L. (2000). Url forwarding and compression in adaptive web caching. *Joint Conf. of the IEEE Computer and Communications Societies*.

Miller, G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, *38*, 39–41.

Mishne, G. (2005). Experiments with mood classification in blog posts. *1st Workshop on Stylistic Analysis Of Text For Information Access (Style 2005)*.

Mitrović, M., Paltoglou, G. & Tadić, B. (2011). Quantitative analysis of bloggers' collective behavior powered by emotions. *Journal of Statistical Mechanics : Theory and Experiment*, *2011*, P02005.

Modha, D. & Spangler, S. (2003). Feature weighting in k-means clustering. *Machine Learning*, *52*, 217–237.

Mohammad, S., Dunne, C. & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. *Conf. on Empirical Methods in Natural Language Processing :*.

Moriyama, T. & Ozawa, S. (2001). Measurement of human vocal emotion using fuzzy control. *Systems and Computers in Japan*, *32*, 59–68.

Mucha1, P. J., Richardson, T., Macon, K., Porter, M. A. & Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*, 876–878.

Mullen, T. & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. *Proc. of EMNLP*.

Muthukrishnan, S. (2005). *Data streams : algorithms and applications*. Now Publishers Inc.

Neviarouskaya, A., Prendinger, H. & Ishizuka, M. (2007). Textual affect sensing for sociable and expressive online communication. In *Affective computing and intelligent interaction*, vol. 4738/2007, 218–229. Springer Berlin.

Ng, V., Dasgupta, S. & Arifin, S. M. N. (2006). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. *COLING/ACL (COLING-ACL '06)*.

Nielsen, F. Å. (2011). A new anew : Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv :1103.2903*.

Núnez, M., Fidalgo, R. & Morales, R. (2007). Learning in environments with unknow dynamics : Towards more robust concept learners. *Journal of Machine Learning Research*, 2595–2628.

Ortony, A. & Turner, T. J. (1990). What's basic about basic emotions ? *Psychological Review*, *97(3)*, 315–331.

Ovesdotter-Alm, C., Roth, D. & Sproat, R. (2005). Emotions from text : machine learning for text-based emotion prediction. *Conf. on Human Language Technology and Empirical Methods in Natural Language Processing.*.

Pak, A. & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Seventh Int. Conf. on Language Resources and Evaluation*.

Pal, N. R. & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. *Fuzzy Systems, IEEE Transactions on*, *3*, 370–379.

Palla, G., Barabási, A.-L. & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, *446*, 664–667.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, *2*, 1–135.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up ? : sentiment classification using machine learning techniques. *ACL-02 Conf. on Empirical methods in natural language processing (EMNLP '02)* (pp. 79–86). Morristown, NJ, USA : Association for Computational Linguistics.

Park, M. Y. & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society : Series B*, *69*, 659—677.

Paroubek, P., Pak, A. & Mostefa, D. (2010). Annotations for opinion mining evaluation in the industrial context of the doxa project. *Int. Conf. on Language Resources and Evaluation (LREC'10)*. Valletta, Malta : European Language Resources Association (ELRA).

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., Wiebe, J., Cohen, K., Brew, C., Hurdle, J., Uzuner, O. & South, B. (2012). Sentiment analysis of suicide notes : A shared task.

Petersen, M. K. & Butkus, A. (2008). Modeling emotional context from latent semantics. *1st Int. Conf. on Designing interactive user experiences for TV and video (UXTV '08)* (pp. 63–66). New York, NY, USA : ACM.

Picard, R. W., Vyzas, E. & Healeys, J. (2001). Toward machine emotional intelligence : Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 1175–1191.

Piolat, A. & Bannour, R. (2009). An example of text analysis software (emotaix-tropes) use : The influence of anxiety on expressive writing. *Current psychology letters*, *25*.

Platt, J. C. (1998). *Sequential minimal optimization : A fast algorithm for training support vector machines* (Technical Report). Advances in kernel methods- Support vector learning.

Plutchik, R. (1990). *The emotions*. University Press of America.

Prabowo, R. & Thelwall, M. (2009). Sentiment analysis : A combined approach. *Journal of Informetrics*, *3*, 143–157.

Qiu, S. (2009). A framework for multiple kernel support vector regression and its applications to sirna efficacy prediction. *Transactions on Computational Biology and Bioinformatics, IEEE/ACM*, *6*, 190–199.

Rafrafi, A., Guigue, V. & Gallinari, P. (2012). Coping with the document frequency bias in sentiment classification. *Int. AAAI Conf. on Weblogs and Social Media*.

Rakotomamonjy, A., Bach, F., Canu, S. & Grandvalet, Y. (2008). Simplemkl. *Journal of Machine Learning Research*, *9*, 2491–2521.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Read, J. (2004). Recognising affect in text using pointwise-mutual information ann. Master's thesis, University of Sussex.

Rosvall, M. & Bergstrom, C. T. (2010). Mapping change in large networks. *PLoS ONE*, *5*, e8694.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.

Russell, J. A. & Mehrabian, A. A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*, 273–294.

Salway, A. & Graham, M. (2003). Extracting information about emotions in films. *11th ACM Int. Conf. on Multimedia* (pp. 299–302). New York, NY, USA : ACM.

Scherer, K. R. (1981). Speech and emotional states. *Speech evaluation in psychiatry*, 189–22.

Scherer, K. R. (2005). What are emotions ? and how can they be measured ? *Social Science Information*, *44*, 695–729.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision tree. *Int. Conf. on New Methods in Language*.

Schmidt, M., Fung, G. & Rosaless, R. (2009). *Optimization methods for l1-regularization* (Technical Report). University of British Columbia.

Sculley, D. (2010). Web-scale k-means clustering. *Int. Conf. on World wide web, WWW* (pp. 1177–1178).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*.

Singh, H. L. P. (2004). Conceptnet : A practical commonsense reasoning tool-kit. *BT Technology Journal*, *22*, 211–226.

Singh, P. (2002). The open mind common sense project.

Sonnenburg, S., Rätsch, G., Schäfer, C. & Schölkopf, B. (2006). Large scale multiple kernel learning. *The Journal of Machine Learning Research*, *7*, 1531–1565.

Stone, P. J., Bales, R. F., Namenwirth, J. Z. & Ogilvie, D. M. (1962). The general inquirer : a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*, 484–498.

Strapparava, C. & Mihalcea, R. (2007). Semeval-2007 task 14 : affective text. *Proc.of the 4th Int. Workshop on Semantic Evaluations*.

Strapparava, C. & Mihalcea, R. (2008). Learning to identify emotions in text. *ACM Symposium on Applied computing*.

Strapparava, C. & Valitutti, A. (2004). Wordnet-affect : an affective extension of wordnet. *4th Int. Conf. on Language Resources and Evaluation*.

Strehl, A. & Ghosh, J. (2002). Cluster ensembles a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.

Strehl, A., Ghosh, J. & Mooney, R. (2000). Impact of similarity measures on web-page clustering. *In Workshop on Artificial Intelligence for Web Search*.

Subasic, P. & Huettner, A. (2000). Affect analysis of text using fuzzy semantic typing. *Proc. of Fuzzy Systems Conf., FUZZ-IEEE.* (pp. 647–652 vol.2).

Tanabe, H., Ho, T. B., Nguyen, C. H. & Kawasaki, S. (2008). Simple but effective methods for combining kernels in computational biology. *Int. Conf. on Research, Innovation and Vision for the Future*.

Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S. & Lee, Y.-K. (2008). Cp-tree : A tree structure for single-pass frequent pattern mining. In *Advances in knowledge discovery and data mining*. Springer Berlin Heidelberg.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*, 267–288.

Tibshirani, R., Walther, G. & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B*, *63*, 411–423.

Tsymbal, A. (2004). *The problem of concept drift : definitions and related work* (Technical Report). Department of Computer Science, Trinity College : Dublin, Ireland.

Turney, P. D. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. *40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wang, P. & Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. *Int. Conf.on Knowledge discovery and data mining*.

Wang, X. & Shen, H. (2009). Clustering high dimensional data streams with representative points. *Proceedinfs of the Sixth Int. Conf. on Fuzzy Systems and Knowledge Discovery*.

Weston, J., Elisseeff, A., Schölkopf, B. & Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, *3*, 1439–1461.

Whitelaw, C., Garg, N. & Argamon, S. (2005). Using appraisal groups for sentiment analysis. *CIKM '05 14th ACM international conference on Information and knowledge management*.

Widmer, G. & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*, 69–101.

Wiebe, J. (2009). Bibliography of work in subjectivity and sentiment analysis. Internet URL [http ://www.cs.pitt.edu/ wiebe/subjectivityBib.html].

Wilbur, J. W. & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, *18*, 44–45.

Wilson, T., Wiebe, J. & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Conf. on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354).

Witten, D. M. & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*, 713–726.

Xu, Z., Jin, R., Yang, H., King, I. & Lyu, M. R. (2010). Simple and efficient multiple kernel learning by group lasso. *Int. Conf. on Machine Learning (ICML)*.

Yang, T., Chi, Y., Zhu, S., Gong, Y. & Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Mach*, *82*, 157–189.

Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. *annual Int. Conf. on Research and development in information retrieval*.

Yossef, Z. B., Keidar, I. & Schonfeld, U. (2009). Do not crawl in the dust : different urls with similar text. *Transactions on the Web*, *3*, 1–31.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.

Zhou, Y. & Liu, L. (2012). Clustering analysis in large graphs with rich attributes. *Data Mining : Foundations and Intelligent Paradigms*, *23*, 7–27.

Annexe A

Apprentissage par noyaux multiples

Parmi l'ensemble des opérateurs qui conservent la propriété de noyaux, certains sont paramétrés par un vecteur de poids $\boldsymbol{\lambda} \in \mathbb{R}_+^L$, nous considérons ici le cas linéaire. Les poids λ_l caractérisent alors l'importance associée à chacune des L représentations originelles, et ont pour effet de dilater ou de contracter les espaces de caractéristique induits par chacun des noyaux.

A.1 Approches heuristiques

Plusieurs heuristiques ont été proposées pour l'ajustement du vecteur de poids, Gönen et Alpaydin (2011) en proposent un état de l'art fourni. Nous regroupons ici les approches qui ajustent le vecteur de poids indépendemment de l'apprentissage réalisé pour construire f.

Pour une tâche de classification binaire où le vecteur $\boldsymbol{y} \in \{-1, 1\}^n$ représente les étiquettes associées à chacun des n documents, une approche consiste à définir le *noyau idéal* comme la matrice $\boldsymbol{y}^\intercal \boldsymbol{y}$ de dimension $n \times n$, pour laquelle une entrée à 1 correspond à deux documents qui partagent la même étiquette, et une entrée à -1 est associée à deux documents d'étiquettes différentes. Le noyau idéal étant construit comme la similarité entre documents dans l'espace des concepts cibles, il est immédiat que la matrice $\boldsymbol{y}^\intercal \boldsymbol{y}$ représente un noyau valide.

Avec le noyau idéal, il est alors possible d'évaluer la qualité d'un noyau κ, et donc celle du vecteur de poids $\boldsymbol{\lambda}$ associé à un ensemble de mesures de similarités. Pour ce faire, il convient de définir une mesure de similarité entre matrices noyaux, Cristianini et al. (2002) proposent de définir l'*alignement de noyaux* comme suit :

$$A(\kappa_1, \kappa_2) = \frac{\langle \kappa_1, \kappa_2 \rangle_F}{\sqrt{\langle \kappa_1, \kappa_1 \rangle_F \ \langle \kappa_2, \kappa_2 \rangle_F}}$$

où $\langle \kappa_1, \kappa_2 \rangle_F = \sqrt{\sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{\boldsymbol{z} \in \mathcal{X}} \kappa_1(\boldsymbol{x}^1, \boldsymbol{z}^1) \kappa_2(\boldsymbol{x}^2, \boldsymbol{z}^2)}$. Cristianini et al. (2002) montrent que la mesure d'alignement présente la propriété de *concentration* qui assure que l'alignement mesuré sur un corpus reste cohérent sur un corpus indépendant.

Appliqué au noyau idéal, $A(\kappa_1, \boldsymbol{y}^\intercal \boldsymbol{y})$ peut être vu comme une mesure de qualité pour le partitionnement induit par κ_1 sur les données (Cristianini et al., 2002). Ainsi, plus

l'alignement avec le noyau idéal est élevé, plus le noyau évalué est en adéquation avec les concepts à apprendre. En particulier, Qiu (2009) propose d'ajuster le vecteur de poids $\boldsymbol{\lambda}$ comme :

$$\lambda_l = \frac{A(\kappa_l, \boldsymbol{y}^\intercal \boldsymbol{y})}{\sum_{j=1}^{L} A(\kappa_j, \boldsymbol{y}^\intercal \boldsymbol{y})}$$

L'intérêt de cette méthode est que la construction de κ ne demande qu'un calcul simple, effectué en amont de l'algorithme d'apprentissage. Une autre approche consiste à identifier le vecteur de poids qui présente un alignement optimal (Lanckriet et al., 2004a; Igel et al., 2007), les méthodes proposées nécessitent néanmoins de coûteux traitements et leur implémentation est plus complexe.

Une autre approche repose sur l'évaluation indépendante de f de pour chacune des mesures κ_l, les performances individuelles ainsi obtenues sont alors vues comme des indices de qualité pour les L espaces de description d'origine. Dans un contexte supervisé, Tanabe et al. (2008) proposent par exemple d'ajuster les poids comme :

$$\lambda_l = \frac{\tau_l - \delta}{\sum_{j=1}^{L} \tau_j - \delta}$$

où τ_l représente une mesure de performance comme le taux de bonne classification de f associé à la mesure κ_l. Le paramètre δ permet de diminuer l'influence des espaces de description les moins prometteurs. Dans un contexte non supervisé, τ_l pourrait représenter une mesure de qualité des partitions produites par f sous κ_l. Cette approche présente cependant un inconvénient : comme pour la fusion tardive, L processus d'apprentissage doivent être réalisés de manière indépendante.

A.2 Approches simultanées

Contrairement aux approches heuristiques, les approches simultanées consistent à apprendre f en même temps que $\varphi_{\boldsymbol{\lambda}}$. L'apprentissage peut être réalisé en une passe : l'algorithme d'apprentissage mis en œuvre pour construire f produit également le vecteur de poids $\boldsymbol{\lambda}$, on parle alors de *méthodes directes*. L'apprentissage peut aussi être réalisé en deux passes itérées : tandis que f est fixé $\boldsymbol{\lambda}$ est mis à jour, puis $\boldsymbol{\lambda}$ est à son tour fixé tandis que f est mis à jour, jusqu'à convergence de l'algorithme. On parle alors de *méthodes enveloppantes*. Pour ces méthodes, l'apprentissage de f est communément désigné par *problème maître* tandis que celui du vecteur de poids par *problème esclave*. Les méthodes enveloppantes présentent un intérêt particulier étant donné que l'apprentissage de f pour un $\boldsymbol{\lambda}$ fixé peut bénéficier des développements passés et futurs d'algorithmes d'apprentissage classiques (Rakotomamonjy et al., 2008).

Parmi les méthodes simultanées, on peut également distinguer les méthodes globales qui, comme présenté ci-dessus, exploitent un unique vecteur de poids $\boldsymbol{\lambda}$, des méthodes locales qui spécifient ce dernier pour chacun des n documents. Ces dernières définissent ainsi une matrice de poids M_λ de dimensions $L \times n$. Pour les machines à vecteurs de support, Gönen et Alpaydin (2008) montrent expérimentalement qu'il est possible d'approcher la frontière de décision induite par un noyau non linéaire (les auteurs utilisent le noyau gaussien) par une frontière de décision induite par un noyau localement linéaire et dont l'interprétation est bien plus aisée. Dans la suite nous considérons spécifiquement les méthodes globales.

A.2.1 Méthodes directes pour les SVM

Lanckriet et al. (2004b) posent les fondations de l'apprentissage de machines à vecteurs de support pour noyaux multiples. A partir du problème de formulation des machines à vecteurs de support, les auteurs définissent le noyau $\kappa = \sum_{j=1}^{L} \lambda_l \kappa_l$ pour des λ_l tous positifs ou nuls (κ est donc un noyau valide). Les auteurs formulent alors le problème de recherche de la frontière de décision f et du noyau κ qui minimise la fonction objectif des machines à vecteurs de support comme un problème d'optimisation semi-défini positif, c'est-à-dire comme un problème d'optimisation convexe dont l'espace de recherche contient des matrices noyaux.

Bach et al. (2004) montrent que la formulation utilisée par Lanckriet et al. (2004b) est équivalente à l'emploi d'une norme mixte $l2/l1$, composée à la fois d'une norme $l2$ sur les espaces de représentation d'origine, comme c'est le cas dans la formulation originelle du problème des SVM, et d'une norme $l1$ sur l'espace des matrices noyaux. Cette dernière a pour effet de promouvoir une parcimonie au sein des κ_l et donc d'effectuer d'une sélection de noyaux. L'intérêt ici est que le problème des SVM ainsi formulé offre un cadre d'étude bien fondé pour le processus de sélection de noyaux, traditionnellement réalisé empiriquement par validation croisée.

Néanmoins, l'algorithme d'apprentissage proposé par Lanckriet et al. (2004b) demande beaucoup de calculs et devient impossible à mettre en œuvre pour de très grandes masses de données ou pour un très grand nombre d'espaces de représentation. Bach et al. (2004) observent que la complexité de l'algorithme provient en partie de l'optimisation explicite de la norme $l1$ (voir section 1.2.1, p. 20) et proposent une formulation du problème dite lissée. En décomposant le problème original, les auteurs obtiennent un problème pour lequel ils proposent un algorithme de recherche séquentiel, similaire à l'algorithme de résolution séquentiel du problème des SVM (Platt, 1998). Les auteurs montrent expérimentalement que l'algorithme qu'ils proposent réalise un apprentissage plus efficace, qui permet de traiter des jeux de données de tailles plus conséquentes.

A.2.2 Méthodes enveloppantes pour les SVM

Les méthodes en deux passes sont introduites par Sonnenburg et al. (2006) qui proposent une nouvelle formulation du problème originel de Lanckriet et al. (2004b) : un problème maître consiste en la recherche du meilleur hyper-plan séparateur étant donné un vecteur de poids fixé. Un problème esclave consiste en la recherche du vecteur de poids optimal étant donné la frontière de décision courante. Les auteurs montrent que le problème maître est la formulation classique du problème des SVM pour lequel les algorithmes classiques de résolution peuvent être employés. Ils montrent aussi que le problème esclave est un programme linéaire pour lequel des outils d'optimisation linéaire classiques peuvent être employés.

L'algorithme d'apprentissage qu'ils proposent présente donc de nombreux avantages : la dichotomie qui est faite entre l'apprentissage de f et celui du vecteur de poids rend possible l'exploitation de la puissance d'algorithmes de résolution des SVM. De plus, l'efficacité des outils d'optimisation linéaire pour la mise à jour des poids rend possible le traitement de jeux de données de tailles conséquentes. Ainsi, Sonnenburg et al. (2006) proposent le premier algorithme d'apprentissage de machines à vecteurs de support multi-noyaux qui passe à l'échelle. Rakotomamonjy et al. (2008) proposent une méthode enveloppante pour laquelle l'apprentissage est plus stable et exhibe une convergence plus rapide.

Kloft et al. (2009) organisent l'ensemble des méthodes proposées jusqu'alors pour l'apprentissage de machines à vecteurs de support multi-noyaux au sein d'un modèle généralisé.

Contrairement aux méthodes précédentes qui imposent des contraintes de parcimonie au niveau de la combinaison des κ_l, l'une des motivations de ce modèle est également l'emploi de normes non parcimonieuses. L'un des résultats remarquables du modèle est que la solution du problème esclave existe sous forme analytique. Il faut noter qu'en adoptant une démarche différente, Xu et al. (2010) obtient parallèlement les mêmes résultats que Kloft et al. (2009). Les vecteurs de poids optimaux pour le problème esclave sont mis à jour par la formule suivante :

$$\lambda_l = \frac{\|\boldsymbol{w}_l\|_2^{2/(p+1)}}{\left(\sum_{j=1}^{L} \|\boldsymbol{w}_j\|_2^{2p/(p+1)}\right)^{1/p}}$$

où p, spécifié par l'utilisateur, indexe la norme utilisée pour régulariser les κ_l et le vecteur $\boldsymbol{w}_l$ définit la frontière de décision apprise par l'algorithme des SVMs. Ici $\|\boldsymbol{w}_l\|_2$ est donné en fonction du vecteur de poids $\boldsymbol{\lambda}$:

$$\|\boldsymbol{w}_l\|_2 = \lambda_l^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \kappa_l(x_i^l, x_j^l)$$

où les α_i et les $\boldsymbol{y}_i$ sont respectivement les poids SVM (Cortes & Vapnik, 95) et les étiquettes associées aux données $\boldsymbol{x}_i$.

Annexe B

Deux textes chargés émotionnellement

La figure B.1 fournit le texte de la chanson « Hymne à la joie », la figure B.2 donne la traduction française de la chanson « *You are not alone* » de Mickael Jackson. Ces deux textes sont disponibles en ligne aux adresses respectives :

– `http://fr.wikipedia.org/wiki/Ode_a_la_joie` et
– `http://www.lacoccinelle.net/242720.html`.

Une étude sur la charge émotionnelle de ces deux textes est présentée à la section 4.3.1, p. 76 : les documents sont dans un premier temps enrichis sémantiquement, les états affectifs auxquels ils sont associés sont ensuite discriminés dans un espace continu pour décrire les émotions.

Hymne à la joie

Joie ! Belle étincelle divine
Fille de l'Élysée,
Nous entrons l'âme enivrée
Dans ton temple glorieux.
Tes charmes lient à nouveau
Ce que la mode en vain détruit ;
Tous les hommes deviennent frères
Là où tes douces ailes reposent.

Que celui qui a le bonheur
D'être l'ami d'un ami ;
Que celui qui a conquis une douce femme,
Partage son allégresse !
Oui, et aussi celui qui n'a qu'une âme
À nommer sienne sur la terre !
Et que celui qui n'a jamais connu cela s'éloigne
En pleurant de notre cercle !

Tous les êtres boivent la joie
Aux seins de la nature,
Tous les bons, tous les méchants,
Suivent ses traces de rose.
Elle nous donne les baisers et la vigne,
L'ami, fidèle dans la mort,
La volupté est donnée au ver,
Et le chérubin apparaît devant Dieu.

Heureux, alors que Ses soleils volent
Sur le glorieux système céleste,
Courez, frères, sur votre voie,
Joyeux, comme un héros vers la victoire.

Qu'ils s'enlacent, tous les êtres !
Ce baiser au monde entier !
Frères, au plus haut des cieux
Doit habiter un père aimé.
Tous les êtres se prosternent ?
Pressens-tu le créateur, Monde ?
Cherche-le au-dessus des cieux d'étoiles !
Au-dessus des étoiles il doit habiter.

Joie ! Belle étincelle des dieux
Fille de l'Élysée,
Soyez unis êtres par millions !
Qu'un seul baiser enlace l'univers !

FIGURE B.1 – Texte de la chanson « Hymne à la joie » (*Ode to joy*).

You Are Not Alone **(Tu n'es pas seul)**

Un autre jour se termine
Je suis encore tout seul
Comment cela se fait-il
Que tu ne sois pas ici avec moi
Tu ne dis jamais au revoir
Quelqu'un m'a dit pourquoi
Mais étais-tu obligée de partir
Et de laisser mon monde si froid

Chaque jour je m'assois et je me demande
Comment l'amour a pu disparaître
Quelquechose me murmure à l'oreille et me dit :

Tu n'es pas seul
Je suis là avec toi
Bien que tu sois loin
Je suis là pour y rester
Tu n'es pas seul
Je suis là avec toi
Bien que nous soyons loin
Tu es toujours dans mon coeur
Tu n'es pas seul

Seul, seul
Pourquoi, seul

Justement la nuit dernière
J'ai pensé que je t'entendais pleurer
Me demandant de venir
Et de te prendre dans mes bras
Je peux entendre tes prières
Ton fardeau je porterai
Mais d'abord j'ai besoin de ta main
Et alors l'éternité peut commencer

Chaque jour je m'assois et je me demande
Comment l'amour a pu disparaître
Quelquechose me murmure à l'oreille et me dit :

Murmure trois mots et j'accourrai
Et ma chère tu sais que je serai là
Je serai là

Tu n'es pas seul...

FIGURE B.2 – Traduction française de la chanson « *You are not alone* » de Mickael Jackson.