Représentation de collections de documents textuels : application à la caractéristique thématique

Abdenour Mokrane

▶ **To cite this version:**

Abdenour Mokrane. Représentation de collections de documents textuels : application à la caractéristique thématique. Interface homme-machine [cs.HC]. Université Montpellier II - Sciences et Techniques du Languedoc, 2006. Français. NNT : . tel-00401651

HAL Id: tel-00401651
https://theses.hal.science/tel-00401651

Submitted on 3 Jul 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

T H E S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline : INFORMATIQUE
Formation Doctorale : INFORMATIQUE
Ecole Doctorale : I2S

présentée et soutenue publiquement

par

M. Abdenour MOKRANE

Le 17 Novembre 2006

Titre :

Représentation de collections de documents textuels : application à la caractérisation thématique

JURY

M. Bernard DOUSSET, Professeur, Univ. Toulouse III , Rapporteur
M. David William PEARSON, Professeur, Univ. St Etienne , Rapporteur
Mme. Danièle HERIN, Professeur, Univ. Montpellier II , Examinatrice
M. Pascal PONCELET, Professeur, Ecole des Mines d'Alès , Directeur de thèse
M. Gérard DRAY, Maître assistant, Ecole des Mines d'Alès , Co-encadrant

Dédicaces

A mes parents et ma femme,
A ma famille,
A tous mes amis,
…

Remerciements

Je tiens à exprimer ma profonde gratitude à Pascal Poncelet, professeur à l'Ecole des Mines d'Alès, directeur adjoint du LGI2P, qui a su diriger ces travaux de thèse, je le remercie vivement pour la qualité de son encadrement, ses orientations et tous ses précieux conseils. Tout au long de ces années de thèse, j'ai pu apprécier tant son dynamisme que sa rigueur scientifique. Je lui exprime toute ma reconnaissance.

Je tiens à exprimer également toute ma reconnaissance à Gérard Dray, maître assistant à l'école des Mines d'Alès pour son soutien constant tout au long de ces travaux de thèse. Je le remercie vivement pour la qualité de ses apports et son encadrement. Je tiens à lui exprimer également ma profonde gratitude.

Je remercie vivement Monsieur Bernard Dousset, professeur à l'Université Toulouse III et Monsieur David William Pearson, professeur à l'Université de Saint Etienne, qui m'ont fait l'honneur d'être rapporteurs de cette thèse. Je les remercie pour leurs remarques et conseils pour l'amélioration du manuscrit.

Je suis très honoré que Madame Danièle Hérin, professeur à l'Université Montpellier II, ait accepté d'examiner ce travail et de faire partie de mon jury. Je lui exprime mes vifs remerciements.

Je remercie également l'ensemble des chercheurs et collègues du LGI2P, notamment Michel Plantié et Rachid Arezki pour leurs collaborations, sans oublier Sylvie Cruvellier, Françoise Armand et le personnel de l'EMA qui m'ont permis de mener, dans de bonnes conditions, mes travaux de thèse.

Je voudrais exprimer également mes vifs remerciements à mes collègues de l'Antenne Universitaire de Blois, pour leur accueil chaleureux en tant qu'ATER, ce qui m'a permis de terminer la rédaction de ce manuscrit dans les meilleures conditions, particulièrement Arnaud Giacometti et Jean Yves Antoine, professeurs à l'Université de Tours, pour leurs nombreux conseils et soutien.

Ma famille à qui je dois bien plus que des remerciements. Je pense en particulier à mes parents qui m'ont permis de faire des études persévérées et m'ont appris la valeur du savoir. Je pense également à ma femme qui m'a soutenu et supporté plus que de raisons des horaires de travail pour le moins extravagants.

Enfin, toute ma reconnaissance et mes remerciements pour tous ceux qui m'ont soutenu et participé de près ou de loin à ces travaux.

Table des matières

Liste des figures

Chapitre I – Introduction

L'explosion du nombre d'informations accessibles et de documents disponibles rend les utilisateurs (entreprises, organismes ou individus) submergés. En effet, ces utilisateurs ne sont plus capables d'analyser ou d'appréhender ces informations dans leur globalité, notamment sur Internet où les informations sont le plus souvent sous formes textuelles [Zan 2005, Clo&al 2006]. Le problème aujourd'hui n'est plus d'accéder aux informations mais de caractériser ces dernières et déterminer l'information utile. Bien entendu, en fonction des besoins de l'utilisateur cette information pourra être utilisée de différentes manières : filtrage de documents, classification de documents, etc.

Il est évident cependant que l'un des éléments clés pour répondre à ces applications est d'être capable d'offrir rapidement l'information enfouie dans les documents. Cette dernière, par exemple, pourra permettre d'affecter un document dans la classe appropriée mais surtout elle offre à l'utilisateur un élément déterminant pour l'aider à prendre ses décisions.

De nombreux travaux de recherche s'intéressent depuis plusieurs années à la mise en œuvre de modèles et systèmes capables d'analyser les contenus textuels, de les organiser et de les représenter automatiquement [Poi 2003, Iha&al 2004, Naz 2004, Dou&Kar 2005, Clo&al 2006]. Il s'agit là d'un des objectifs principaux des approches d'extraction de connaissances à partir de données textuelles [Poi 2003, Zan 2005]. Traditionnellement ce processus est composé de trois phases principales (une présentation détaillée du processus est proposée dans le chapitre II) :

- La première phase a pour but de prétraiter les documents et de les représenter.

- La seconde phase consiste à extraire, à partir de ces représentations, des connaissances en appliquant des techniques de fouille de données.

- La troisième phase se focalise sur les résultats obtenus et cherche à valider la connaissance acquise en fonction des connaissances préalables du domaine.

Une condition sine qua non de la réussite d'un processus d'extraction est de réussir la première étape qui occupe généralement 80% du temps. En effet, cette étape nécessite beaucoup de travail et l'un des éléments clé est sans doute la représentation des documents. Obtenir cette représentation est une tâche difficile car nous sommes confrontés au dilemme suivant : Il faut d'une part être le plus

exhaustif possible (la connaissance acquise sera extraite de cette représentation) mais cette dernière doit également être la plus concise possible pour pouvoir être utilisable par les techniques de fouille de données de la seconde étape.

Pour extraire la connaissance de ces représentations, il existe à l'heure actuelle de très nombreux algorithmes de fouille de données adaptés à différents domaines d'applications (nous reviendrons sur ces techniques au cours du mémoire). Ainsi, à l'issue de ces deux étapes, l'utilisateur final se retrouve confronté avec la connaissance enfouie. Avant de pouvoir prendre une décision, il doit vérifier d'une part que la connaissance est réellement représentative de sa collection de documents et surtout il doit repérer la connaissance qui lui sera réellement utile. Par exemple, savoir que dans 95% des documents qui abordent le langage Java, le terme objet est fortement corrélé à Java est une connaissance mais pas réellement utile. Pour l'aider, il est là aussi indispensable de lui proposer des outils adaptés.

1. Principales contributions

Dans le cadre de ce mémoire, nous nous intéressons à des collections de documents qui abordent des thématiques différentes. Notre problématique s'inscrit complètement dans le cadre d'un processus d'extraction de connaissances à partir de documents textuels car nous souhaitons mettre à la disposition de l'utilisateur des outils pour lui permettre d'extraire automatiquement des informations sur les différentes thématiques abordées. Nous souhaitons également mettre à sa disposition des mécanismes lui offrant la possibilité de « résumer » les contenus. Avec de tels outils il devient alors possible de répondre à des applications aussi diverses que :

- Des cartes de connaissances offrant la possibilité d'accéder au contenu général des documents. Elles offrent également la possibilité de naviguer entre les différentes connaissances.

- Avoir une information générale sur le contenu. Très souvent, dans un premier temps, les utilisateurs ne cherchent pas quelque chose de précis et préfère avoir un aperçu général du contenu des documents. Une bonne méthode pour aider l'utilisateur est de lui offrir un aperçu général des documents de la même manière qu'il le ferait avec une table des matières.

- Classification de documents. En connaissant les différents groupes de thématiques existants dans la collection, il est possible par exemple de

classer automatiquement des documents mais également de les filtrer en fonction de l'intérêt de l'utilisateur.

Au cours de ce mémoire nous nous intéresserons à ce type d'applications. Ainsi, nous considérons les problèmes suivants : étant donné une collection de documents multithématiques, nous souhaitons trouver un modèle de représentation adapté. Via ce modèle notre problème est d'extraire de la manière la plus automatique possible les différentes thématiques et d'offrir des mécanismes pour naviguer dans les connaissances apprises.

1.1. Modéliser l'information pour extraire la connaissance

Nous verrons, au cours de ce mémoire, que nous proposons un nouveau modèle de représentation de collections de documents. Les travaux antérieurs ont montré qu'une approche basée sur la cooccurrence de termes était efficace pour aider à résumer le contenu des documents [Zan 2005, Iha&al 2004]. Cependant, elle souffre de certaines lacunes liées principalement aux choix de l'utilisateur sur les fréquences d'occurrences de termes ou sur le contexte dans lequel ces termes apparaissent. Nous verrons en particulier que via la fréquence de nombreuses informations sur les relations entre termes qui pourraient être utiles sont éliminées dès le début du processus. Pour résoudre ces problèmes, nous proposons un nouveau critère appelé « partage de contextes ». Intuitivement l'idée de ce critère est la suivante « l'ami de mon ami est mon ami ». Alors que le critère de cooccurrence a tendance à ne regrouper que des termes qui apparaissent ensemble dans un même contexte. Le partage de contextes offre la possibilité de regrouper des termes appartenant à des contextes différents mais qui possèdent des termes communs.

Sur ce principe, nous proposons une approche de modélisation de documents qui débute à l'issue de la phase de prétraitements et qui offre en sortie à la fois les termes en cooccurrences et ceux qui partagent les mêmes contextes.

1.2. Extraire et visualiser les connaissances des thématiques

Nous proposons également, à partir de ces représentations de documents, d'extraire et de regrouper automatiquement les thématiques des différents documents. Les poids des thématiques dans les collections de documents ne sont pas forcément équilibrés, i.e. le nombre de documents portant sur chacune des

thématiques n'est pas forcément le même, nous verrons que l'approche proposée n'est pas conditionnée par les proportions de documents portant sur chacune des thématiques. Nous proposons également à l'utilisateur de visualiser les connaissances extraites via des cartes dynamiques d'informations. Nous montrerons comment ces dernières, couplées à l'identification de groupes de thématiques, aident l'utilisateur dans ses tâches de consultation documentaire.

1.3. Le système IC-DOC

Pour valider nos propositions, les différents algorithmes proposés ont été intégrés dans un système appelé *IC-DOC* (*Information Characterization from Document Collections*) dont l'objectif est de proposer un environnement d'extraction de connaissances pour des données textuelles issues de thématiques différentes. Nous montrerons comment ce système a été étendu pour extraire des groupes de thématiques et faciliter la navigation dans les connaissances.

2. Organisation du mémoire

Dans le chapitre II, nous revenons plus en détail sur les problématiques étudiées dans le cadre de notre travail. Notre problématique s'inscrivant dans un contexte d'extraction de connaissances à partir de documents, nous présenterons le processus général. Nous nous focaliserons sur trois aspects de ce processus : les étapes de prétraitements, la modélisation des documents et la phase d'extraction. Même si notre objectif principal n'est pas de proposer une nouvelle approche de prétraitements de documents, il est indispensable de connaître les différentes approches existantes de manière à effectuer les choix les plus pertinents lors du développement d'un système d'extraction. Le second point concerne plus particulièrement la problématique abordée dans cette thèse. Aussi nous décrirons les principales approches existantes. Enfin, au cours de la phase d'extraction, nous proposerons un rapide survol des techniques de fouille de données et montrerons comment elles peuvent être utilisées pour traiter des données textuelles. Nous concluons ce chapitre par une discussion au cours de laquelle nous reviendrons sur les limites des approches traditionnelles de représentation des documents.

Le chapitre III présente notre nouveau modèle de représentation de collections de documents. Nous exposons notre proposition basée à la fois sur les notions d'association de termes et sur la notion de partage de contextes entre les différents termes d'une collection de documents. Au cours de ce chapitre, nous présenterons les deux principales phases de notre approche. Enfin, comme dans le chapitre

précédent nous concluons via une discussion sur les avantages de notre approche par rapport aux approches classiques proposées dans le chapitre II.

Le modèle que nous avons défini est intégré dans un système d'extraction de connaissances appelé *IC-DOC* que nous présenterons dans le chapitre IV. Après avoir présenté l'architecture générale du système et de ces principaux composants, nous montrerons deux domaines d'application et d'utilisation d'*IC-DOC*. Le premier domaine concerne l'identification de groupes de clusters de thématiques différentes. Avant de présenter les expérimentations menées nous proposons un aperçu des approches de clustering. La seconde application offre à l'utilisateur une cartographie des connaissances extraites de l'ensemble de documents.

Enfin, le chapitre V conclut ce mémoire en revenant sur les principales propositions et en présentant un certain nombre de perspectives associées.

Le mémoire comporte également une annexe présentant des extraits de documents utilisés pour illustrer les différents concepts ou définitions.

Chapitre II – Problématique et Etat de l'Art

Pour pouvoir faire face aux grandes quantités de données textuelles disponibles, il est indispensable d'offrir à l'utilisateur final de nouvelles approches qui vont lui permettre d'appréhender, de la manière la plus aisée possible, les éléments significatifs contenus dans ces documents. Les travaux sur l'analyse de documents textuels ont été largement étudiés par des communautés comme par exemple celle du traitement automatique du langage naturel. Cependant, récemment de nouveaux travaux connus sous le nom d'extraction de connaissances à partir de données textuelles sont apparus pour offrir à l'utilisateur, à partir d'un ensemble de documents, les connaissances qui pourront être actionnables. Pour l'utilisateur final, le résultat escompté est d'avoir, par exemple, une classification automatique des documents, (e.g. *comment gérer automatiquement les nombreux mails qui arrivent tous les jours ?*), d'obtenir un résumé ou un aperçu des documents (e.g. *quels sont les éléments importants qui sont contenus dans les textes ? est-il possible en quelques lignes ou via un schéma d'obtenir les principales connaissances associées à des textes ?*) ou encore de regrouper ensemble et de manière automatique les informations qui parlent des mêmes thématiques (e.g. *comment regrouper d'un côté les informations qui abordent le problème de l'économie du football et de l'autre la création d'entreprises sachant que l'utilisateur s'intéresse régulièrement aux deux thématiques ?*). L'objectif du processus d'extraction de connaissances est justement d'essayer de répondre à ces différents problèmes. Notre travail s'inscrit dans ce contexte.

Le chapitre est organisé de la manière suivante. Dans la section 1, nous revenons plus en détail sur la problématique étudiée. Dans la section 2, nous présentons les éléments significatifs du processus d'extraction de connaissances à partir de textes. Le traitement de données de type textuel nécessite d'effectuer des prétraitements sur les documents de manière à pouvoir les manipuler ou les analyser. Même si notre problématique n'est pas liée à cette étape nous présentons les principales approches dans la mesure où elles sont utilisées dans le système *IC-DOC* que nous présenterons dans le chapitre IV. De la même manière, nous présenterons quelques unes des approches d'analyse qui sont particulièrement adaptées à notre contexte. Notre proposition étant principalement accès sur la définition d'un nouveau modèle de représentation des collections de documents manipulés, nous insisterons sur les principales approches existantes à l'heure actuelle. Enfin dans la section 6, nous concluons ce chapitre par une discussion sur les différents modèles de représentation ainsi que sur leurs limites.

1. Problématique

Nous avons vu, dans le chapitre précédent, que pour pouvoir aider l'utilisateur final à traiter, de la manière la plus automatique possible, les grandes quantités de documents textuels disponibles aujourd'hui, il devient indispensable de proposer de nouvelles approches qui lui offrent un « aperçu » ou une « caractérisation » des différents contenus textuels. Bien entendu cet « aperçu » peut être décrit de différentes manières. Par exemple, en lui proposant de regrouper ensemble tous les documents qui abordent les mêmes thématiques, nous offrons, aux travers des différents groupes obtenus, une première étape pour faciliter la compréhension des textes manipulés. Une autre manière de résumer les documents est de « cartographier » les connaissances contenues dans les documents, i.e. d'extraire des informations représentatives du contenu, pour par exemple regrouper des éléments (termes, associations, etc.) sur chacune des thématiques abordées par les documents afin de les représenter de manière intelligible. Au travers de ces deux exemples, nous voyons que l'un des problèmes cruciaux à résoudre est : *quelle connaissance conserver ?* En effet, une condition sine qua non pour permettre de regrouper des documents est d'être capable d'extraire une connaissance commune entre les différents documents. De la même manière, pour extraire et cartographier les éléments importants caractérisant une collection de documents, il est indispensable de se préoccuper de ce que l'on souhaite représenter et de la nature des relations textuelles enfouies dans les documents. S'il y a trop peu d'éléments, il est clair que la représentation sera inutile car clairement pas suffisamment représentative. Inversement, si un trop grand nombre d'informations est proposé à l'utilisateur, il ne sera pas à même de les appréhender et se retrouvera confronté à son problème initial mais dans un autre contexte : *comment trouver ce qui est important dans ce qui est proposé ?*

Le premier problème que nous abordons dans ce mémoire concerne les connaissances à retenir. En d'autres termes, notre problème peut être décrit de la manière suivante :

Soit un ensemble de documents $D = \{D_1, D_2, \ldots, D_n\}$ où chaque document D_i est lui-même composé d'une liste de termes ordonnés $T = \{T_1, T_2, \ldots T_m\}$. Soit M un modèle de représentation de l'ensemble des documents. La problématique de la représentation des documents consiste à rechercher le modèle M tel que :

- M possède suffisamment d'informations pour représenter le contenu des collections de documents. En d'autres termes, il faut garantir que M conserve les éléments les plus représentatifs des différents contenus

textuels, par exemple ceux du contenu global ou ceux d'une thématique spécifique.

- *M* doit cependant être suffisamment réduit pour pouvoir appliquer des traitements et des algorithmes de fouille de données.

De manière à illustrer cette problématique, considérons les trois documents suivants : *D1*= « *Java est un langage de programmation objet.*», *D2*= « *Java est le plus utilisé dans les Ecoles d'Ingénieurs*.», *D3*= « *La programmation objet fait référence au langage Java.* », *D4* « « *Le football professionnel a pris aujourd'hui deux décisions importantes.* ». Une première analyse rapide de ces documents montre que certains termes (« un, de, le, les, la, au ») ne sont pas utiles et ne peuvent pas servir à représenter ces documents. En poursuivant ces analyses, nous pouvons transformer nos documents en : « *Java langage programmation objet.*», « *Java utiliser Ecole Ingénieur*.», «*programmation objet référence langage Java.* ». A partir de ces documents transformés, notre problème consiste donc à trouver le modèle *M* qui soit suffisamment représentatif pour exprimer, par exemple, qu'il existe des rapports entre « java et langage » ou « java et objet ».

Cependant l'obtention d'un modèle de représentation des documents laisse en suspens la question suivante : *quid des thématiques des documents ?* En effet, quelque soit le modèle de représentation trouvé, il est indispensable de montrer que celui-ci peut effectivement aider l'utilisateur final. La seconde problématique consiste donc, à partir du modèle *M*, à montrer que nous sommes à même soit de regrouper ensemble les informations (éléments cohérents extraits des textes et représentatifs du contenu) qui abordent les mêmes thématiques soit de visualiser la connaissance associée au contenu. Ainsi dans notre exemple précédent, nous devons être à même de montrer que la collection aborde à la fois la thématique de « programmation en java » et de « l'économie du football ».

Enfin, si nous sommes capables de proposer un modèle *M* et si nous pouvons extraire des connaissances représentatives des différentes thématiques, le dernier problème est d'offrir à l'utilisateur une plateforme intégrant tous ces aspects. Bien entendu, nous souhaitons que cette plateforme soit la plus automatique possible.

2. Le processus d'Extraction de Connaissances

Motivés par des problèmes d'Aide à la Décision, les chercheurs de différentes communautés (Intelligence Artificielle, Statistiques, Bases de Données, Interface Homme Machine) se sont intéressés à la conception et au développement d'une

nouvelle génération d'outils permettant d'extraire automatiquement de la connaissance de grandes bases de données. Ces outils, techniques et approches sont le sujet d'un thème de recherche connu sous le nom d'Extraction de Connaissances dans les Bases de données (*Knowledge Discovery in Databases*). Ce dernier est défini comme un processus non trivial qui consiste à identifier, dans les données, des schémas ou modèles nouveaux, valides par rapport aux connaissances du domaine, potentiellement utiles et surtout compréhensibles et utilisables [Fay&al 1996]. Ce processus, décrit dans la Figure 1, comprend globalement trois phases :

- **Préparation des données :** L'objectif de cette phase consiste à sélectionner uniquement les données potentiellement utiles dans la base. L'ensemble des données est ensuite soumis à des prétraitements, afin de les transformer et de gérer des données manquantes ou invalides. L'étape suivante dans cette phase consiste à formater ces données pour les rendre compréhensibles aux algorithmes de fouille de données (opérations de transformation et réduction).

- **Extraction :** En appliquant des techniques de fouille de données, l'objectif est de mettre en évidence des caractéristiques ou des modèles contenus implicitement dans les données.

- **Interprétation des résultats :** Le but de cette dernière phase est d'interpréter la connaissance extraite lors de l'étape précédente, pour la rendre lisible et compréhensible par l'utilisateur et permettre ainsi de l'intégrer dans le processus de décision.

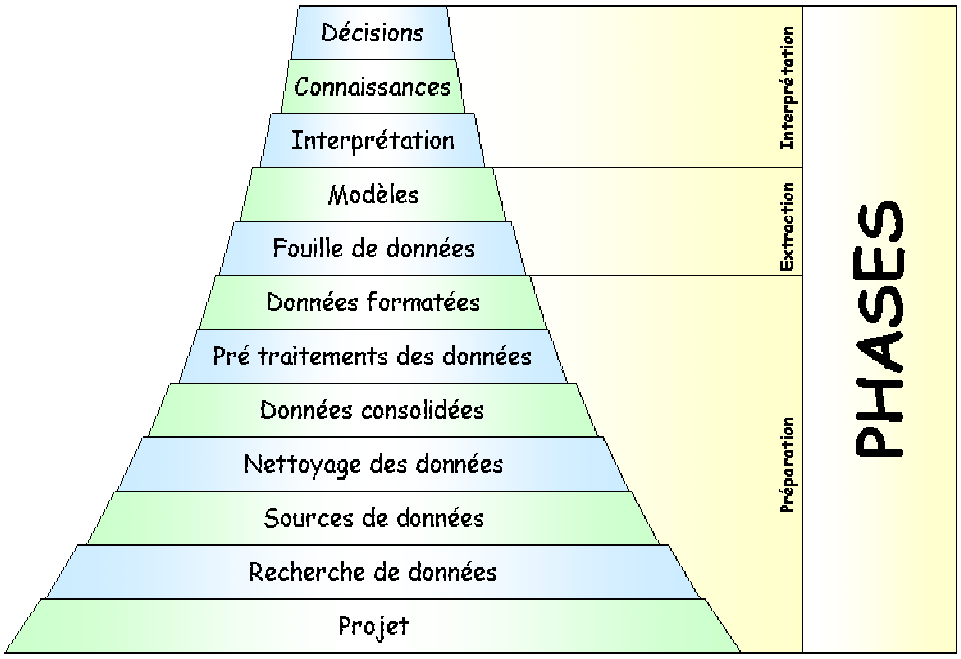


Figure 1. Les différentes phases du processus d'Extraction

Ce processus est utilisé couramment dans des applications de compagnies d'assurance, compagnies bancaires (crédit, prédiction du marché, détection de fraudes), marketing (comportement des consommateurs, « mailing » personnalisé), recherche médicale (aide au diagnostic, au traitement, surveillance de population sensible), réseaux de communication (détection de situations alarmantes, prédiction d'incidents), analyse de données spatiales, etc.

Ce processus général est bien sûr adapté aux différents types de données manipulées. Ainsi dans le cas des données textuelles, le terme extraction de connaissances à partir de bases de données textuelles (*Knowledge Discovery in Textual Databases*) ou Fouille de Textes (*Text Mining*) est apparu au milieu des années quatre vingt dix [Fel&Dag 1995, Fel&al 1998, Hea 1999] avec pour objectif « de trouver des relations intéressantes impossibles ou difficiles à détecter par une analyse séquentielle de l'information » [Kod 2000].

Le processus d'extraction à partir de textes est illustré dans la Figure 2 et est assez similaire au précédent. Nous y retrouvons les phases traditionnelles :

- **Phase I :** Prétraitements et transformations des textes. Cette phase fait appel à des techniques de prétraitements des textes : nettoyage des textes,

suppression des mots peu informatifs, et/ou normalisation. En outre, étant donné qu'il n'est pas possible de traiter les documents dans leur globalité, la transformation offre une représentation réduite des différents documents.

- **Phase II : S**élection et réduction des données. Cette phase est généralement regroupée avec la première et a pour but soit de réduire réellement le volume des contenus textuels, soit de minimiser l'espace de recherche (e.g. dans le cas d'une représentation des documents basée sur les vecteurs par exemple, C.f. Section 4).

- **Phase III :** Utilisation d'algorithmes de fouille de données.

- **Phase IV :** Analyse, interprétation et validation des résultats.

Dans les sections suivantes nous présentons plus en détail les principales approches de prétraitements dans la mesure où elles seront utilisées dans le système que nous proposons dans le chapitre IV. Nous nous focaliserons ensuite sur les principales approches, proches de notre problématique, de représentation de documents afin de mieux en étudier leurs limites. Enfin nous proposerons un aperçu des principales approches de fouille de données particulièrement utilisées aujourd'hui pour analyser des collections de documents.

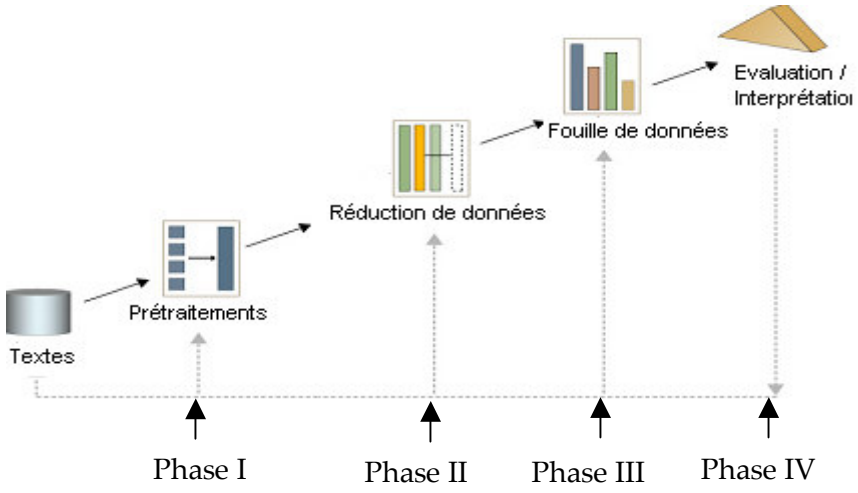


Figure 2. Le processus d'extraction de connaissances à partir de documents textuels

3. Le prétraitement des documents

Les données textuelles sont une forme particulière de données complexes. Elles ne sont pas délimitées, structurées et étiquetées sémantiquement de façon explicite. En conséquence ces données nécessitent un traitement préalable. De manière générale, l'objectif de ce prétraitement est de minimiser l'espace de recherche. En effet, même si les capacités des ordinateurs évoluent constamment, il n'est malheureusement pas possible de traiter les documents dans leur intégralité. En fonction des différentes thématiques de recherche, cet objectif est mené de différentes manières. Par exemple, les approches issues du traitement automatique du langage naturel offriront des techniques pour obtenir les différents composants d'une phrase, pour désambiguïser, pour associer du sens aux termes manipulés, etc. D'un autre côté les approches plus statistiques se focaliseront sur des techniques plus globales en recherchant par exemple les préfixes de mots communs.

Il est important de noter que quelque soit la technique utilisée, elle est souvent couplée à une approche plus classique basée sur la suppression de mots ou termes appartenant à une liste (e.g. bad words, liste noire, etc.). Cette approche offre en outre l'avantage d'éliminer des mots qui ne seraient pas utiles dans le cadre d'une analyse donnée. Par exemple, l'auxiliaire être et avoir sont souvent utilisés et ne caractérisent pas le contenu des documents, donc peuvent être éliminés du traitement. [1]

Dans la suite de cette section, nous nous intéressons aux approches basées sur un étiquetage morphosyntaxique. Cependant cette étape n'est pas suffisante et elle est généralement suivie d'une étape de « lemmatisation ».

3.1. Approche morphosyntaxique et lemmatisation

L'étiquetage morphosyntaxique correspond à la préparation des textes pour la phase de modélisation du contenu. Il comprend une analyse morphologique et une analyse syntaxique [Cha 1984, Bri 1992]. Notons que ces deux analyses sont précédées par certains prétraitements (traitement des ponctuations, majuscules, codages et formats). Une analyse morphologique peut être considérée comme un automate qui traite isolément chaque forme d'un texte en lui associant des traits informationnels (ou propriétés) [Fay&al 1991]. L'analyse syntaxique permet de segmenter les textes en propositions. Chaque proposition est formée de couples

[1] Ceci est généralement utilisé dans les approches statistiques. Toutefois dans certains contextes ces verbes doivent être conservés car ils sont porteurs de sens.

(entrée lexicale, catégorie). Les seules ambiguïtés qui demeurent sont internes à une catégorie. Les résultats de l'analyse des propositions sont des arborescences de structures syntaxiques attestées par la langue [Hab&al 1997]. Au final, l'étiquetage morphosyntaxique associe à chaque mot d'une phrase sa catégorie morphologique (genre, nombre) et syntaxique (nom, adjectif, verbe, etc.).

Exemple 1 :

> Soit le document *D1*= « *Java est un langage de programmation objet.*», une analyse morphosyntaxique de ce document donne :
>
> « java – nom commun féminin singulier », « est – verbe indicatif présent 3ème personne du singulier », « un – article indéfini masculin singulier », « langage – nom commun masculin singulier », « de – préposition », « programmation – nom commun féminin singulier », « objet – nom commun masculin singulier », « . – ponctuation forte (fin de phrase) ».

Plusieurs étiqueteurs, ou « taggers », existent à l'heure actuelle, pour le français ou l'anglais notamment, et atteignent des performances qui dépassent les 90% de correction (i.e. le quotient du nombre de mots correctement étiquetés sur le nombre total de mots étiquetés). Un état de l'art sur l'analyse morphosyntaxique est disponible dans [Par&Raj 2000].

Dans la suite de cette section, nous présentons la lemmatisation qui est généralement couplée à l'analyse morphosyntaxique.

Dans le langage naturel, il existe une grande redondance dans les marques morphologiques. Par exemple, dans la morphologie des adjectifs, rien ne correspond en oral aux marques écrites du féminin. En effet, la règle écrite de passage du masculin au féminin consiste à ajouter un « e ». Par exemple, dans « enfoui/enfouie », le « e » ne se prononce pas à l'oral. De la même manière, le pluriel des syntagmes nominaux, « le scientifique européen » s'écrit « les scientifiques européens » au pluriel. Il y a plus d'économie des marques en oral qu'en écrit. Pour remédier au problème des redondances dans les marques morphologiques, il serait intéressant de regrouper, par exemple, les formes singulier/pluriel sous une forme unique.

La lemmatisation permet le regroupement des formes morphologiques d'une même unité linguistique en une seule unité appelé *lemme*. Elle réduit ainsi des mots en entités premières, appelées lemmes ou formes canoniques. Par exemple « journal » est la forme canonique de « journal » et « journaux ». Le lemme est

l'infinitif pour les verbes, la forme masculine singulière pour les noms, etc. La lemmatisation permet de réfléchir en fonction du sens des mots en faisant abstraction de leurs formes. Elle permet d'analyser le contenu d'une collection de documents, sans avoir à rentrer l'ensemble de ses variantes pour chacun des mots contenus dans les documents.

Exemple 2 :

> Considérons le document de l'Exemple 1, les différents lemmes associés sont les suivants :
>
> « java », « être », « un », « langage », « de », « programmation », « objet ».

Il existe également une autre catégorie d'algorithmes ou techniques proches de la lemmatisation. Il s'agit des algorithmes de « stemming » (appelé déssuffixation en français). Ces algorithmes ne sont pas couplés généralement à l'analyse morphosyntaxique. Ils associent plusieurs mots ayant le même radical, i.e. enlever les suffixes des mots pour ne conserver que la partie racine, en s'aidant de règles et de listes d'exceptions. Les algorithmes de « stemming » les plus connus sont ceux de Lovins [Lov 1968] et de Porter [Por 1980]. Par exemple, le stem des mots « manger » et « mangeable » est « mang ».

Exemple 3 :

> Les stems associés au document de l'Exemple 1, sont les suivants : « jav », « est », « un », « langag », « de », « programm », « objet ».

4. La représentation des documents

Il existe dans la littérature de nombreux modèles de représentation de documents textuels. Nous pouvons citer, par exemple, l'utilisation de vecteurs dont les composantes représentent des termes [Sal 1973, Sal 1989], des matrices de distribution de termes ou de relations entre termes [Bes&al 2001, Lan&Lit 1991]. Dans le cadre de notre problématique, nous nous intéressons aux modèles utilisant les mots ou les termes comme unité de représentation. En effet, ces derniers ont la réputation d'être simples et efficaces. Ils existent cependant certains travaux qui utilisent la phrase (ou un segment de texte) comme unité de représentation [Sch&al 1995, Tze&Har 1993, Fuh&Buc 1991]. L'argumentation en faveur de ces modèles réside dans le fait que la phrase est plus informative que le mot. Cette dernière offre également l'avantage de conserver les positions des

mots. Cependant, les expériences menées ont montré que les résultats n'étaient pas meilleurs [Car&al 2001] et que surtout ces approches sont très difficiles à mettre en œuvre. En conséquence, en utilisant les phrases comme des unités de représentation, les connaissances syntaxiques sont conservées mais les connaissances statistiques sont dégradées à cause du trop grand nombre de combinaisons possibles [Car&al 2001, Lew 1992a].

L'objectif de la suite de cette section est d'étudier les différents modèles de représentation de textes largement utilisés dans la littérature. Pour chacun des modèles nous analysons les critères pris en considération pour la représentation des différents termes.

4.1. Représentations vectorielles

Les modèles vectoriels sont largement utilisés pour la représentation de textes. Le modèle vectoriel standard et le modèle *Latent Semantic Indexing* sont les plus utilisés et implémentés. Nous détaillons ces derniers dans la suite de cette section.

Le modèle standard MS « sacs de mots »

Dans le cadre du modèle vectoriel standard (*MS*), les textes sont considérés comme des « sacs de mots » [Sal 1971a, Sal 1971b, Sal&McG 1983]. L'idée principale est de transformer les différents documents d'une base documentaire en vecteurs où chacun des éléments d'un vecteur de texte représente des unités textuelles ou tout simplement des mots appelés aussi « termes d'indexation ». Plusieurs travaux utilisent les mots comme termes d'indexation [Dum&al 1998, Aas&Eik 1999, Apt&al 1994, Lew 1992b]. Un mot est considéré comme étant une suite de caractères encadrés par des caractères de ponctuation ou appartenant à un dictionnaire spécifique. Des outils, basés sur des approches linguistiques, statistiques ou mixtes sont utilisés pour l'identification des différentes unités textuelles.

Dans le modèle vectoriel standard, les composantes d'un vecteur représentant un texte sont fonction de l'occurrence des mots dans le texte. Ce modèle a été initialement introduit par *Gérard Salton* [Sal&Les 1965, Sal 1971a] dans l'objectif d'implémenter un système de recherche d'informations. L'implémentation la plus connu de ce modèle est le système de recherche documentaire *SMART*. L'évolution de ce système est décrite dans [Sal 1991]. Dans ce modèle les composantes des vecteurs représentent des termes considérés comme les plus discriminants. Dans le cadre du modèle vectoriel standard, ces termes sont

sélectionnés en fonction de leurs fréquences d'apparition dans les documents et en fonction du nombre de documents contenant ces termes.

Soit *C* une collection de documents textuels, D_i un document de *C*, soit *t* le nombre de termes d'indexation et $T = \{T_1,\ldots, T_j,\ldots,T_t\}$ l'ensemble de ces derniers. Dans le modèle vectoriel standard, le document D_i est représenté par un vecteur V_i. La collection de textes peut être ainsi représentée par une matrice dont les colonnes représentent les termes d'indexation et les lignes représentent les documents de cette collection.

$$V_i = (W_1,..., W_j,...W_t),$$
où W_j est le poids d'un terme T_j dans le document D_i et $j = 1..t$.

Le poids donné à un terme d'indexation dans un document est calculé en fonction de la fréquence d'occurrence du terme *TF* (*Term Frequency*) dans le document et du nombre de documents contenant le terme *IDF* (*Inverse Document Frequency*). Les calculs des poids et les pondérations accordés à un document *D* ont fait l'objet de nombreuses études [Sin 1997, Lee 1995, Buc&al 1992, Sal&Buc 1988].

Le modèle LSI/PLSI

Le modèle *Latent Semantic Indexing* (*LSI*) découle du modèle vectoriel standard. Il tente de prendre en considération la structure sémantique des termes pour la représentation des documents. Dans ce modèle, les documents sont représentés dans un espace réduit de termes d'indexation [Dee&al 1990, Sch&al 1995, Lan&al 1998]. Les techniques *LSI* utilisent, dans un premier temps, une matrice *M* (documents × unités linguistiques), dans laquelle chaque élément W_{ij} est une pondération en fonction du nombre d'occurrences du terme T_j dans le document D_i. Soit *n* le nombre de documents de la collection et *t* le nombre des termes d'indexation. La matrice M peut être représenté comme suit.

$$M = \begin{pmatrix} d_1 \\ d_2 \\ \ldots \\ d_n \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & \ldots & W_{1t} \\ W_{21} & W_{22} & \ldots & W_{2t} \\ \ldots & \ldots & W_{ij} & \ldots \\ W_{n1} & W_{n2} & \ldots & W_{nt} \end{pmatrix}$$

Une décomposition en valeurs singulières (*SVD*) de la matrice *M* est ensuite effectuée. Après cette décomposition, seuls les *k* premiers vecteurs propres sont intégrés, les axes factoriels correspondant aux plus grandes valeurs propres sont conservés en application du théorème de *Eckart et Young* [Dee&al 1990]. Ce théorème montre qu'il s'agit dans ce cas des axes permettant un ajustement dans un espace de dimension réduite minimisant la perte d'information. Dans *LSI*, la valeur représentée dans la matrice sur laquelle est appliquée la décomposition est définie comme étant le produit du poids local du terme, i.e. poids du terme dans le document, et du poids global du terme, i.e. poids du terme dans la collection de documents. Cette valeur peut aussi correspondre à la fréquence d'un terme donné dans un document donné. Ces valeurs ne sont que des pondérations proches de *TF/IDF* [Dee&al 1990, Lan&al 1998].

Hoffmann a proposé un modèle probabiliste du *Latent Semantic Indexing (PLSI)*. Il considère l'hypothèse que les documents sont associés à un certain nombre de sens (*Latents*) et que les termes correspondent à l'expression de ces sens [Hof 1999]. De façon probabiliste, notant *W* l'ensemble des mots, *D* l'ensemble des documents, tel que : $W = \{W_1,\ldots,W_t\}$ et $D = \{D_1,\ldots,D_n\}$, la probabilité de la paire observée ($d \in D$, $w \in W$) est donnée par la formule suivante :

$$p(d, w) = p(w \mid d)\, p(d)$$

Cette probabilité est calculée en utilisant un algorithme de type *Expectation-Maximization (EM)* [Dem&al 1977]. Les dimensions de l'espace réduit du modèle *LSI* correspondent ici aux sens du modèle *PLSI*.

Représentations conceptuelles et basées thésaurus

Ils existent des approches utilisant les concepts pour la représentation des textes tel que *Word Category Map* ou des modèles issus de la méthode *LSI* [Koh&al 2000, Dee&al 1990, Lio&al 2004]. Ces méthodes dépendent de la distribution des probabilités des mots au sein du jeu d'apprentissage, en conséquence les données du jeu ont une influence sur la génération de l'espace des concepts. J. Chauché [Cha 1990] a proposé un nouveau modèle vectoriel de représentation de textes. Au lieu de définir un espace vectoriel dont chaque dimension représente un terme d'indexation, l'ensemble des termes est projeté sur un ensemble fini de concepts extraits d'un thesaurus. Cette représentation permet une factorisation des termes par regroupement de leurs champs sémantiques. Par exemple, deux synonymes

partageront un ensemble de mêmes concepts. L'auteur utilise, pour des documents français, un thésaurus composé de 873 concepts hiérarchisés en 4 niveaux. Rappelons qu'un thésaurus permet uniquement d'explorer, à partir d'un concept, les mots qui s'y rattachent et inversement.

Par exemple, le mot « *interview* » défini par les concepts 419 (question), 583 (compagnie) et 726 (communication), 749 (conversation) et 766 (presse) du thésaurus, sera représenté par un vecteur de dimension 873 dont toutes les composantes sont nulles sauf celles associées aux concepts 419, 583, 726, 749 et 766 qui seront identiques (C.f. Figure 3). Le thésaurus est donc défini comme un ensemble de couples de $L \times R^{873}$ avec L l'ensemble des lemmes du thésaurus. Les dimensions de l'espace vectoriel ne sont pas associées à des termes d'indexation mais à des concepts.

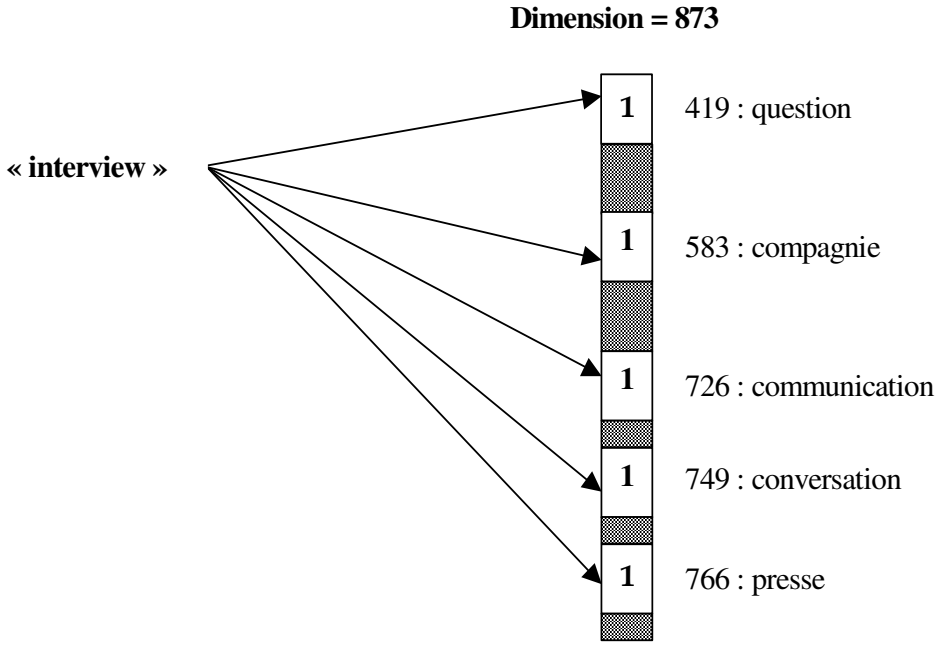


Figure 3. Exemple de représentation conceptuelle du mot « interview »

Après l'extraction de l'ensemble des lemmes d'un texte, une association est réalisée entre les lemmes et le vecteur qui leur est associé au sein du thésaurus. Ensuite, pour chaque texte un vecteur conceptuel est calculé en fonction de la moyenne normalisée des lemmes qu'il contient [Jai&al 2005].

Autres modèles vectoriels

Ils existent d'autres modèles vectoriels de représentation de textes, appelés modèles probabilistes. Des modèles particuliers, développés spécifiquement pour la recherche documentaire, utilisent des pondérations justifiées sur des bases probabilistes dépendantes des requêtes (une requête est formée d'un ensemble de mots clés pour effectuer une recherche documentaire). La représentation finale de chaque document est un vecteur dont chaque composante est une pondération associée à un terme.

Les modèles probabilistes de recherche documentaire sont par exemple présentés dans [Spa&al 1998]. Ces modèles ont également l'avantage d'éviter le problème de la haute dimensionnalité de la requête. Nous pouvons citer, par exemple, le modèle probabiliste de représentation de *Robertson* et *Spark-Jones* développé pour la recherche documentaire [Rob&Spa 1976]. Il repose sur le principe d'ordre des probabilités (*Probability Ranking Principle*) [Rob 1977] pour présenter les documents retournés aux utilisateurs. Nous pouvons également citer, comme modèle probabiliste, *Okapi*, qui prend en considération la distribution des fréquences des termes dans les documents [Rob&al 1981, Rob&al 1994]. Ce modèle a été testé avec succès pour les campagnes *TREC*. L'hypothèse faite par les auteurs est qu'un terme est relié à un thème (i.e. un document parle ou ne parle pas de ce thème). Le document parlant du thème peut utiliser un certain nombre de fois le terme correspondant ou ne pas l'utiliser. Réciproquement, un document ne parlant pas du thème peut néanmoins utiliser le terme associé. Dans ce modèle la distribution des fréquences des termes dans les documents est un mélange de deux distributions : la première pour les documents parlant du thème et la seconde pour les autres documents. L'hypothèse retenue est que ces deux distributions suivent des lois de Poisson [Rob&Wal 1994].

4.2. Représentations basées sur les associations de termes

Nous avons présenté dans la section précédente divers modèles de représentations vectorielles de textes. Ces modèles exploitent essentiellement la structure explicite des documents (les mots). Certains d'entre eux tentent de prendre en considération les dépendances qui peuvent exister entre les mots par des modèles statistiques ou des transformations stochastiques en partant des informations sur l'occurrence des mots dans les documents (*TF*) ou en documents (*IDF*). L'origine de ces approches est issue de la *Sémantique Distributionnelle* (*SD*) qui est fondée sur l'hypothèse suivante : « La sémantique des éléments textuelles est déterminée par leurs distributions dans les textes ». En effet, elle suppose l'existence

d'une forte corrélation entre les caractéristiques distributionnelles observables des mots et leurs sens. Ainsi, la sémantique d'un mot est reliée à l'ensemble des contextes dans lesquels apparaît ce dernier [Har 1988, Har&al 1989]. Notons que la sémantique distributionnelle hérite de la théorie de Firth « *The basic assumption of the theory of analysis by levels is that any text can be regarded as a constituent of a context of situation, ……, You shall know a word by the company it keeps* » [Fir 1957] qui peut se résumer par « le sens d'un mot peut être donné par ses voisins ».

Les modèles qui prennent en compte les dépendances entre les mots n'effectuent pas une analyse préalable de la distribution des associations des termes dans les collections de documents. Ils utilisent des techniques statistiques, linguistiques ou mixtes. Ainsi, les approches linguistiques cherchent à dégager le lexique des documents puis étudient les différentes dépendances morphologiques et syntaxiques des mots dans les phrases [Chu&Han 1989, Paz 1999]. L'analyse des documents par des méthodes statistiques est plus récente et fait suite aux travaux d'Estoup et de Zipf [Leb 1998]. Elle se base sur l'application de méthodes quantitatives aux éléments linguistiques. Avant d'exposer les modèles existants dans la littérature, nous précisons la notion d'association de termes.

Association de termes

En raison des difficultés de représentation des connaissances textuelles par des modèles symboliques structurés tel que les « frames » et les modèles logiques, des méthodes plus adaptées aux applications exploitant des collections de documents sont nécessaires. Ces méthodes ont suscité un intérêt particulier ces dernières années. Nous nous intéressons particulièrement dans la suite de cette section aux modèles de représentation basés sur la notion d'association de termes.

D'une manière générale, dans la littérature, la cooccurrence (appelée également association ou co-citation) de termes est définie de la manière suivante :

Définition 1 (Cooccurrence de termes)

> Soient deux termes T_1 et T_2. Une cooccurrence entre les termes T_1 et T_2 est définie comme étant l'apparition commune des deux termes dans un même contexte. Elle correspond à deux termes qui apparaissent ensemble dans un même segment de texte.

En fonction des modèles un segment de texte peut être une fenêtre de mots, une phrase ou un paragraphe.

Le modèle DSIR

Le modèle de représentation de textes *Distributional Semantic for Information Retrieval (DSIR)* [Raj&al 2000, Bes&al 2001] est basé sur les cooccurrences des mots dans les collections de documents. Les contextes des unités linguistiques sont des éléments essentiels du modèle *DSIR* car ils constituent le support principal pour la dérivation des représentations des mots. Ces dernières sont obtenues à partir des fréquences de cooccurrences entre l'ensemble des mots d'une base documentaire et les termes d'indexation de cette dernière. Dans ce contexte, un mot (unité linguistique) U_i est représenté par un vecteur de poids associés aux fréquences de cooccurrences de ce mot avec l'ensemble des termes d'indexation T. Ce vecteur est appelé profil de cooccurrences de U_i par rapport à T. Notons que dans le modèle *DSIR* un contexte de cooccurrence correspond à une phrase.

Soit W_{ij} le poids associé à la fréquence de cooccurrence de U_i avec un terme d'indexation T_j et t = |T|. U_i est représenté par un vecteur VCO_i :

$$VCO_i = (W_{i1},\ldots,W_{ij}, \ldots,W_{it})$$

Ainsi, une collection de documents est représentée par une matrice M de cooccurrences (unités linguistiques × termes d'indexation).

Soit U l'ensemble des mots de la collection et n le nombre de ces mots. Soit T l'ensemble des termes d'indexation et t le nombre de ces derniers. La matrice M est définie comme suit.

$$M = \begin{pmatrix} U_1 \\ U_2 \\ \ldots \\ U_n \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & \ldots & W_{1t} \\ W_{21} & W_{22} & \ldots & W_{2t} \\ \ldots & \ldots & W_{ij} & \ldots \\ W_{n1} & W_{n2} & \ldots & W_{nt} \end{pmatrix}$$

Le modèle *DSIR* utilise une fonction de représentation d'un document à partir de la matrice de distribution des unités linguistiques. Dans ce cas, il s'agit de la matrice M. Dans un premier temps, un document est représenté par un vecteur V d'occurrence suivant le modèle vectoriel standard. Puis ce vecteur est transformé via la matrice M par une multiplication du vecteur V par M. Cette représentation

prend en considération la distribution des cocitations de mots et intègre des connaissances syntaxiques dans la phase de sélection des cooccurrences. En effet, les cooccurrences des mots sont prises entre les gouverneurs des groupes syntaxiques ou les constituants d'un même groupe syntaxique. Une étape de filtrage syntaxique est donc nécessaire avant toute représentation de textes.

Divers modèles basés sur l'analyse des cooccurrences

La notion de cooccurrences a été utilisée dans divers autres modèles pour accomplir des tâches variées, notamment dans les systèmes de recherche d'informations basés sur l'expansion de requêtes [Les 1969, Rij 1977, Buz&al 2001]. Ces travaux reposent sur l'hypothèse suivante : si un terme d'indexation est utile pour la discrimination des documents pertinents par rapport aux documents non pertinents, alors tout terme d'indexation étroitement associé est aussi susceptible d'être utile pour cette discrimination.

Cependant, l'expansion de requêtes par l'utilisation de modèles de cooccurrences n'a pas apporté les améliorations espérées. En particulier, H.J. Peat et P. Willett [Pea&Wil 1991] montrent que ces modèles ne sont pas particulièrement adaptés à la recherche documentaire. Dans [Sch&Ped 1994], les auteurs proposent une méthode assez proche de celle de *DSIR*. La construction d'un thesaurus automatique est effectuée en calculant les similarités entre termes par leurs profils de cooccurrences. Dans [Qiu&Fre 1993, Qiu&Fre 1995], les auteurs proposent la notion de « *Similarity Thesaurus* » en calculant la similarité « terme × terme » par la comparaison des profils de leurs distributions dans les documents. Ce thesaurus est alors utilisé pour faire une expansion des requêtes en calculant, à chaque fois, les termes les plus proches de la requête dans son ensemble (et non seulement de chaque terme de la requête). Cette dernière est représentée par le vecteur moyen des profils de distribution des termes qu'elle contient dans la collection de documents. [Dag&al 1997, Dag&al 1999] proposent d'estimer la probabilité de cooccurrence de deux mots U_1 et U_2, lorsque ces derniers n'apparaissent pas ensemble dans un même contexte dans le corpus. Cette opération est réalisée par une combinaison des estimations de cooccurrences entre U_2 et les mots les plus proches d'U_1 (la proximité étant calculée par une mesure de similarité spécifique).

Il existe également d'autres approches de représentation de textes proches des modèles présentés dans cette section, il s'agit des modèles basés sur les graphes. Nous pouvons citer les travaux de A. Lelu [Lel 2003] qui utilise des éléments de la théorie des graphes pour la représentation et la comparaison de textes. Le langage *UNL* (*Universal Networking Langage*) est un formalisme permettant de représenter la sémantique de chaque document par un graphe. Les informations écrites en

langage naturel sont converties en *UNL* puis traduites dans des langages cibles [Uch&Zhu 2005a, Uch&Zhu 2005b].

5. La fouille de données

Dans cette section, nous revenons sur l'une des étapes principales du processus d'extraction de connaissances à partir de documents textuels : La fouille de données, et nous présentons quelques unes des techniques existantes. Même s'il existe de nombreuses techniques de fouille de données disponibles à l'heure actuelle[2], les travaux de recherche autour des documents ont particulièrement abordé les approches suivantes :

- **La recherche de règles d'association** : Le problème de la recherche de règles d'association introduit par R. Agrawal et al. en 1993 [Agr&al 1993], est souvent appelé « problème du panier de la ménagère » (*Market Basket Problem*) car les transactions opérées par les clients d'un magasin et dont la trace est stockée représentent une application typique pour le processus de découverte de connaissances. Dans un tel contexte, une règle d'association peut être par exemple : « 85% des clients qui achètent du beurre et du café achètent aussi du lait ». La recherche de règles couvre un large champ d'applications telles que la conception de catalogues en ligne dans un contexte de e-commerce, la promotion de ventes, le suivi de clientèle, la gestion des stocks, etc. Dans le cas de données de type textuel, les approches de règles d'association sont utilisées généralement pour rechercher des corrélations entre termes de documents [Fel&Hir 1998]. Par exemple, en considérant que l'on dispose d'un ensemble de documents réécrits sous la forme {identification du document, ensemble de mots clés}, l'objectif est d'abord de rechercher les mots fréquemment corrélés entre eux puis de générer les règles d'associations.

- **La recherche de motifs séquentiels** : En 1996, la problématique de la recherche de règles d'association est étendue pour détecter des comportements typiques dans le temps et le concept de motifs séquentiels est introduit [Agr&Sri 1995]. La recherche de tels motifs consiste à extraire des ensembles d'objets couramment associés sur une période de temps spécifiée. Il est alors possible d'extraire des relations

[2] Plus de 27 techniques différentes sont répertoriées dans (**http://www.kdnuggets.com**) et pour chacune d'entre elles, il existe de très nombreux algorithmes plus ou moins adaptés.

temporelles comme par exemple « 36% des clients qui achètent une télévision, achètent un lecteur de DVD dans les deux ans qui suivent et un Home-Cinéma 6 mois après » ou « 30% des abonnés d'une vidéothèque qui ont emprunté Marius, empruntent Fanny un mois plus tard, puis César quelques semaines après ». La problématique de l'extraction de motifs séquentiels est en fait une extension de celle des règles d'association. En effet la prise en compte du temps dans les enregistrements à étudier permet une plus grande précision dans les résultats, mais implique aussi un plus grand nombre de calculs et de contraintes. Ce problème posé à l'origine par l'industrie de la vente au détail, intéresse à présent des domaines aussi variés que les télécommunications (détection de fraudes), la médecine (identification de symptômes précédant les maladies) ou encore les domaines financiers. Par rapport aux règles d'association, les motifs offrent dans le traitement des textes de conserver l'ordre d'apparition des différents termes. En appliquant des algorithmes de recherche de motifs, nous obtenons alors non plus des corrélations entre termes de documents mais plutôt des corrélations de succession de termes. Par exemple, dans [Len&al 1997], les auteurs proposent d'utiliser des algorithmes de motifs séquentiels pour extraire des tendances dans des documents.

- **La classification** : Elle consiste à analyser de nouvelles données et à les affecter, en fonction de leurs caractéristiques ou attributs, à telle ou telle classe prédéfinie. Les techniques de classification sont, par exemple, utilisées pour déterminer, pour une banque, si un prêt peut être accordé, en fonction de la classe d'appartenance d'un client. La classification est très utilisée pour traiter des données de type textuel car il existe de nombreux domaines d'applications (classification automatique de documents, de mails, de news, etc.). Ces dernières années des approches particulièrement efficaces ont été proposées. Un état de l'art de ces approches est proposé dans [Ber 2003] et [Seb 2006]. La classification peut également bénéficier d'autres approches de fouille de données. Ainsi, dans [Wan&al 1999], les auteurs proposent d'effectuer une classification à partir des résultats obtenus via des algorithmes d'extraction de règles et S. Jaillet propose de classifier à partir de motifs séquentiels [Jai 2005].

- **Le clustering** : Le problème du clustering (appelé aussi classification non supervisée ou segmentation) consiste à regrouper des enregistrements qui semblent similaires dans une même classe. Il est complémentaire à

celui de la classification car le but ici est de rechercher les différentes classes possibles d'appartenance en fonction des différents attributs ou critères qui caractérisent les données. Les applications concernées incluent notamment la segmentation de marché, la segmentation démographique (pour identifier par exemple des caractéristiques communes entre populations), etc. Dans le cas de documents textuels, le clustering est particulièrement utilisé pour regrouper des documents en fonction de leur contenu. Nous revenons plus en détail sur ces aspects dans le chapitre IV en proposant une présentation des principes généraux.

6. Discussion

Dans ce chapitre, nous avons présenté les problématiques étudiées ainsi que les différents travaux existants autour de ces problématiques. Dans cette discussion, nous revenons tout d'abord sur le processus d'extraction de connaissances à partir de documents textuels en examinant son adéquation par rapport à nos problématiques. Nous revenons ensuite sur les problèmes de modélisation des documents.

Par rapport aux problématiques générales que nous abordons, le premier constat est que ce travail s'inscrit dans le cadre d'un processus d'extraction de connaissances à partir des données textuelles. En effet, nous souhaitons aider l'utilisateur à mieux appréhender les contenus de l'ensemble de documents dont il dispose (cet objectif est également partagé par celui du processus). Les conséquences sont nombreuses comme nous avons pu le voir tout au long de ce chapitre : nous ne pouvons pas manipuler tous les documents dans leur intégralité, nous avons besoin de modéliser les « éléments » réellement caractéristiques de documents, nous avons besoin d'utiliser des algorithmes de fouille de données pour obtenir la connaissance, etc. Parmi les travaux existants autour du processus, nous avons choisi d'en décrire plus particulièrement trois : le prétraitement, la modélisation et la fouille. Ces choix n'ont pas été fait au hasard. En ce qui concerne le premier point, nous avons pu constater dans la section 3 que les approches morphosyntaxiques suivies d'une lemmatisation étaient très efficace à l'heure actuelle, le problème réside dans le choix de méthodes pertinentes. Nous reviendrons sur cet aspect dans le système *IC-DOC* que nous proposons au chapitre IV. En ce qui concerne la fouille de données, il existe de nombreuses méthodes qui sont plus ou moins adaptées en fonction des buts visés. Par exemple, si nous souhaitons classer automatiquement des documents, les travaux menés autour de la classification ont montré qu'ils étaient

tout à fait adaptées. Dans notre contexte, nous nous intéressons à la problématique inverse : regrouper ensemble et de manière automatique des informations (éléments textuels : termes ou relations textuels) de chacune des thématiques de la collection. Pour cela, nous retiendrons, dans notre système des algorithmes de clustering.

Revenons à présent à la problématique de la représentation des documents. Nous avons proposé dans ce chapitre un état de l'art synthétique sur les modèles de représentation de collections de documents textuels. Parmi ces derniers, les plus utilisés sont les modèles vectoriels. Par rapport à notre problème, le modèle vectoriel standard (*MS*) via sa présentation en « sac de mots » ne prend pas en considération toute notion de similarité textuelle ou de distance entre les différents mots. Ce modèle exclut également toute analyse grammaticale ou contextuelle des mots. Le *MS* a été enrichi par le modèle *LSI* qui procède à des transformations sur la matrice *M* (documents × unités linguistiques). Cependant, les valeurs de la matrice transformée ne sont que des pondérations proches de *TF/IDF*.

Des représentations plus intelligibles tentent de prendre en considération la sémantique des documents. Il s'agit des représentations conceptuelles ou basées thésaurus. L'inconvénient majeur de ces représentations est qu'elles s'attachent à des vocabulaires statiques. Ceci est dû au fait qu'un réseau de concepts ou un thésaurus permettent uniquement d'explorer à partir d'un concept les mots qui s'y rattachent et inversement. Ces représentations sont dépendantes de la langue et ne s'adaptent pas aux bases documentaires à vocabulaires particuliers (tel que des textes techniques ou des bases documentaires issues du Web).

D'autres types de modèles représentent les documents par des segments de textes (comme les phrases). En utilisant par exemple les phrases comme des unités de représentation, les connaissances syntaxiques sont conservées mais comme nous l'avons vu les connaissances statistiques sont dégradées à cause du trop grand nombre de combinaisons à considérer.

Les modèles basés sur les associations de termes ont fait leur preuve pour détecter rapidement les relations textuelles. Dans le cadre de ces approches, les modèles de similarités textuelles et la notion de cooccurrences sont les plus utilisés pour l'analyse du contenu. Dans un contexte proche, cas de la recherche documentaire, la recherche de cooccurrences a également été largement étudiée. Elle consiste à rechercher les associations de termes les plus fréquentes dans les documents afin de retrouver rapidement les documents pertinents qui peuvent répondre aux requêtes de l'utilisateur. Par exemple, *R. Besançon* a proposé un modèle de

filtrages syntaxiques de cooccurrences pour la représentation de documents [Bes&al 2001].

Les modèles présentés dans ce chapitre ne prennent pas en considération, à la fois, la notion de cooccurrences avec la notion de partage de contextes pour l'extraction et la représentation des connaissances textuelles. Ce qui implique une pénalisation d'une partie importante des relations entre termes.

Dans les chapitres suivants nous répondons à ces différentes limites en proposant un nouveau modèle de représentation et nous montrons comment ce modèle, couplé à des outils existants de prétraitements et de fouille, peut être utilisé pour répondre à nos problématiques.

Chapitre III – Modèle de représentation

Au cours du chapitre précédent, nous avons présenté la problématique étudiée ainsi que les principaux travaux relatifs à la modélisation de documents dans un contexte d'extraction de connaissances à partir de textes. Nous avons également vu que les approches basées sur la cooccurrence étaient adaptées à notre contexte car elles permettent de rechercher les associations qui peuvent exister entre différents termes. Dans ce chapitre, nous proposons une nouvelle approche de représentation de collections de documents textuels. Cette dernière reprend la notion de cooccurrences tout en l'étendant. Ainsi, contrairement aux approches précédentes, nous ne pénalisons pas les termes non discriminants s'ils peuvent servir à enrichir la connaissance via les cooccurrences qui existent dans les documents, nous exploitons ensuite cette connaissance pour représenter les contenus textuels en introduisant le critère de « partage de contextes ». Nous verrons au cours de ce chapitre les différents traitements qui sont effectués à partir d'un ensemble de textes initiaux afin d'obtenir des matrices représentant le plus de connaissances possibles.

Le chapitre est organisé de la manière suivante. Dans la section 1, nous proposons un premier modèle de représentation de collections de documents qui est basé sur la notion de cooccurrences. Au cours de la section 2, nous étendons l'approche précédente en intégrant un nouveau critère important : « Le partage de contextes ». Enfin, nous concluons ce chapitre par une conclusion et discussion sur l'approche proposée.

1. Vers un modèle de représentation

Avant de présenter notre approche pour obtenir un nouveau modèle de représentation des collections de documents, nous proposons un ensemble de définitions préliminaires. Dans la section 1.2, nous décrivons le principe général retenu et proposons une illustration de l'approche. L'algorithme général est présenté dans la section 1.3 où nous revenons sur la méthode retenue pour déterminer les termes significatifs.

1.1. Définitions préliminaires

Nous travaillons sur un ensemble de documents textuels. A partir de ce dernier, nous considérons qu'une étape de prétraitements est réalisée pour obtenir une collection de documents sur laquelle nous effectuons nos analyses. Au cours du chapitre suivant nous présenterons la méthode que nous avons retenue dans le cas du système *IC-DOC*, notamment pour les prétraitements. A l'issue de cette

phase, nous disposons d'une collection de documents, où une collection est définie de la manière suivante :

Définition 2 (Collection de Documents)

> Soit $D=\{D'_1, \ldots, D'_i, \ldots D'_n\}$ un ensemble de n documents. Une collection de documents, $C=\{D_1, \ldots, D_i, \ldots D_n\}$, correspond à l'ensemble des n documents transformés par les étapes de prétraitements. Ainsi un document D_i dans C correspond au document initial D'_i transformé en une liste de termes. Ces termes respectent la ponctuation initiale du document D'_i.

Dans la suite de ce chapitre, nous considérerons que le terme « document » correspond à un document réécrit après la phase de prétraitements. Ainsi, si les prétraitements sont réalisés à l'aide d'une approche de lemmatisation et d'analyse morphosyntaxique, le document réécrit sera composé des différents lemmes ordonnés en respectant la ponctuation initiale du document.

Dans le cas où les prétraitements sont basés sur une approche de stemming, le document réécrit sera composé de stems ordonnés respectant la ponctuation.

De manière à illustrer les différents concepts et définitions de ce chapitre, nous considérons trois documents issus de thématiques différentes : « Économie du Football », « Création d'Entreprises » et « Sécurité sur Internet ». Des extraits de ces documents sont proposés en annexe A.

Exemple 4 :

> Considérons l'un des documents D'_1 proposé en annexe A. Considérons l'extrait de texte ci-dessous du document D'_1.
>
> « *Le football professionnel français a pris ce jour deux décisions importantes. Le lancement du Challenge de l'offensive et la réforme de la Coupe de la Ligue qui montrent sa capacité à innover.* »
>
> Si nous considérons des prétraitements basés sur une approche de lemmatisation et d'analyse morphosyntaxique, l'extrait est transformé de la manière suivante.
>
> « *football professionnel français prendre décision important. Lancement Challenge offensive réforme Coupe Ligue montrer capacité innover.* »

Dans le cas de prétraitements basés sur une approche de stemming, nous obtenons :

« *football professionnel franc pris décis import. Lanc Challeng offens réform Coup Ligu montr capac innov.* »

Chaque document est à présent composé d'une liste de termes et nous pouvons définir l'ensemble de tous les termes de la collection.

Définition 3 (Termes d'une collection)

Soit une collection de n documents, $C=\{D_1, \ldots, D_i, \ldots D_n\}$. L'ensemble des termes de C est noté TC et est tel que $TC = TD_1 \cap \ldots \cap TD_i \ldots \cap \ldots TD_n$ où TD_i correspond à l'ensemble des termes obtenus après l'étape de prétraitements. TD_i est tel que $TD_i = \{T_1, \ldots, T_j, \ldots T_m\}$ où T_j avec $j = 1 \ldots m$ correspond au terme obtenu par prétraitements du document initial D_i.

L'une des étapes essentielles lors de l'extraction de connaissances est de repérer non seulement la fréquence d'un terme dans un document mais également dans l'ensemble des documents avec sa distribution.

Définition 4 (Fréquence d'un terme)

Pour chaque terme, il est possible de connaître d'une part sa fréquence d'apparition dans la collection, appelée FTC mais également le nombre de documents qui contiennent ce terme, appelé FTD. Soit la collection $C = \{D_1, \ldots, D_i, \ldots, D_n\}$, et T un terme tel que $T \in TC$. $F_1, \ldots, F_i, \ldots, F_n$ correspondent respectivement aux fréquences d'apparition du terme T dans $D_1, \ldots, D_i, \ldots, D_n$. La fréquence FTC du terme T correspond au nombre total d'occurrences du terme T dans la collection C, elle est définie de la manière suivante :

$$FTC_T = \sum_{i=1}^{n} F_i$$

La fréquence FTD du terme T correspond au nombre de documents de C qui contiennent T.

De manière à illustrer les différentes définitions, considérons l'exemple suivant :

Exemple 5 :

Considérons les documents situés en Annexe A[3]. L'ensemble des termes des documents après prétraitements est le suivant : *TC* = {Créateur, repreneur, entreprise, Cci, Finistère, organiser, challenge, espoir, économie, français, soutenir, conseil, général, opération, viser, récompenser, secteur, commerce/services, personne, industrie/services, jeune, demandeur, emploi, établissement, scolaire, projet, réaliser, élève, cadre, scolarité, réussite, individuel, Brest, Morlaix, Quimper, Cornouaille, entendre, saluer, audace, oser, aventure, création, reprise, connaître, insatiable, nécessité, entreprendre, droit, auteur, sécurité, internet, major, insister, adopter, mesure, restrictif, télécharger, acte, protestation, industrie, techno-fasciste, défendre, loi, numérique, filtrage, frontière, pays, Chine, Corée, penser, défense, constater, regret, intermédiaire, presser, détourner, œuvre, profit, passionner, informatique, suivre, intéresser, groupe, étudiant, école, ingénieur, Esiea, concours, intituler, Securitech, savoir, devenir, élément, phare, communication, moderne, société, craindre, exploitation, donnée, sensible, malhonnête, espionnage, industriel, pression, récupération, base, client, particulier, souhaiter, voir, personnel, dévoiler, intégralité, sécuriser, réseau, requérir, fort, connaissance, mécanisme, danger, pouvoir, menacer, technique, démarche, permettre, infiltrer, voiler, face, protéger, machine, exploiter, actualité, football, relancer, spectacle, championnat, administration, décider, lancer, saison, professionnel, prendre, décision, important, lancement, réforme, coupe, ligue, montrer, capacité, innover, incidence, classement, sportif, évaluer, pertinence, intérêt, nouveau, barème, hidalgo, équipe, offensif, commission, organisation, compétition, officialiser}.

Considérons les termes suivants de *TC* : économie, football, offensif, sportif, challenge. Les fréquences d'occurrences FTC_i correspondantes respectivement à ces différents termes sont : 3, 2, 6, 2, 6. Les fréquences en documents FTD_i des différents termes sont respectivement : 3, 1, 1, 1, 2. Nous pouvons constater que les termes « offensif » et « challenge » sont les plus fréquents dans la collection *C*. Les termes « football » et « sportif » sont les moins fréquents. Le terme « économie » apparaît dans tous les documents de la collection *C* tandis que les termes « football », « offensif » et « sportif » apparaissent chacun dans un seul document.

[3] Dans la suite de ce chapitre, nous considérons, par souci de lisibilité, que les termes ont été obtenus à l'aide d'une étape de prétraitements basée sur une analyse morphosyntaxique et une lemmatisation.

Examinons à présent comment les termes peuvent apparaître dans un même contexte. Comme nous l'avons vu précédemment, il existe de nombreuses manières de gérer le contexte (fenêtre de mots, paragraphe, phrase). Nous avons retenu dans notre approche une notion de contexte liée à la phrase, i.e. deux termes appartiennent à un même contexte s'ils apparaissent dans la même phrase. A partir de l'ensemble des termes d'une collection donnée, il est donc possible de définir des cooccurrences de termes. Dans le cadre de notre approche, ces dernières sont définies comme suit.

Définition 5 (Cooccurrence contextuelle)

> Deux termes T_l, et T_m d'une collection C, i.e. $T_l \in TC$ et $T_m \in TC$, qui apparaissent dans une même phrase appartiennent au même contexte de cooccurrence. Une cooccurrence CO est notée ($CO : T_l — T_m$). L'ensemble des cooccurrences de la collection C est noté $CO2C$. Le nombre de cooccurrences de la collection C, NCO, correspond à la cardinalité de $CO2C$, i.e. $NCO = |CO2C|$.

Nous avons vu précédemment que nous pouvions définir la notion de fréquence d'un terme. Nous étendons cette notion à celle de cooccurrence dans la définition suivante :

Définition 6 (Fréquence d'une cooccurrence)

> La fréquence FCC d'une cooccurrence CO dans une collection C correspond au nombre d'occurrences total de CO dans la collection. Soit T_l et T_m deux termes de la collection $C = \{D_1, \ldots, D_i, \ldots, D_n\}$. Supposons que T_l et T_m soient en cooccurrence, i.e. ($CO : T_l — T_m$). $O_1, \ldots, O_i, \ldots, O_n$ correspondent aux fréquences d'apparition de la cooccurrence ($CO : T_l — T_m$) dans $D_1, \ldots, D_i, \ldots, D_n$. $FCC_{(CO : Tl — Tm)}$ est définie comme suit :

$$FCC_{(CO : Tl — Tm)} = \sum_{i=1}^{n} O_i$$

Les notions de cooccurrences et de fréquences de cooccurrences sont illustrées dans l'exemple suivant :

Exemple 6 :

Considérons l'extrait de document suivant « Nous le savons tous : La sécurité informatique est devenue un élément phare de la communication moderne » dont la transformation est : « savoir : sécurité informatique devenir élément phare communication moderne. » après les prétraitements. L'ensemble des termes associés à cet extrait est le suivant *TC* = {savoir, sécurité, informatique, devenir, élément, phare, communication, moderne}. L'ensemble des cooccurrences associées à cet exemple est le suivant : *CO2C* = {(CO : sécurité—informatique), (CO : sécurité—devenir), (CO : sécurité—élément), (sécurité—phare), (sécurité—communication), (sécurité—moderne), (informatique—devenir), (informatique—élément), (informatique—phare), (informatique—communication), (informatique—moderne), (devenir—élément), (devenir—phare), (sécurité—communication), (devenir—moderne), (élément—phare), (sécurité—communication), (élément—moderne), (phare—communication), (phare—moderne), (sécurité—phare), (communication—moderne)}. Nous avons $|CO2C| = 22$. En outre nous pouvons constater que le terme « savoir » n'est en cooccurrence avec aucun autre élément de *TC*, i.e. il n'existe pas d'autres termes qui apparaisse dans la même phrase avec celui-ci. Nous pouvons également remarquer que la cooccurrence (CO : sécurité—informatique) apparaît 3 fois dans des contextes différents, i.e. dans des phrases différentes. Nous avons donc *FCC*(CO : sécurité— informatique) = 3.

1.2. Principe général de représentation

Dans cette section, nous présentons l'approche générale que nous proposons. Tout d'abord, à partir d'un ensemble de documents, une étape de prétraitements est effectuée de manière à transformer cet ensemble en une collection de documents (C.f. Définition 2). L'objectif de cette transformation est le même que dans les approches traditionnelles d'extraction de connaissances à partir de textes.

A partir de cette collection, nous extrayons les fréquences globales des termes dans les collections de documents (*FTC* et *FTD*). Etant donné que nous disposons de tous les termes et des documents associés, nous recherchons les ensembles de cooccurrences de ces termes, i.e. les termes tels qu'ils apparaissent dans une même phrase.

Toutes ces informations sont alors utilisées pour créer une matrice qui stocke les informations sur les occurrences des termes et les cooccurrences. Cette matrice, appelée *MATCO*, est définie de la manière suivante :

Définition 7 (Matrice de cooccurrences)

Soit *TC* l'ensemble des termes d'une collection *C*, notés $T_1,\ldots,T_i,\ldots,T_N$ où $i=1..N$, $N = |TC|$, i.e. le nombre de termes de la collection *C*. Nous notons *VTC* le vecteur de termes de la collection *C* construit comme suit.

For $i=1$ **to** N **do** $VTC(i)=T_i$

La matrice de cooccurrences de *C*, notée *MATCO* de *TC* correspond à une matrice de *N* lignes et *N* colonnes. La ligne *i* de la matrice correspond au $i^{ème}$ terme du vecteur *VTC* et la colonne *j* de la matrice correspond au $j^{ème}$ terme du vecteur *VTC*. Cette matrice est symétrique, sa diagonale correspond aux *FTC* des termes du vecteur *VTC*, le reste des éléments correspondent soit aux *FCC* des cooccurrences de l'ensemble *CO2C* ou à la valeur 0 en cas d'absence de cooccurrences entre un terme correspondant à une ligne *i* avec un terme correspondant à une ligne *j*. Plus formellement, *MATCO* est telle que :

If $(i \neq j)$ $MATCO\ (i,j) = FCC(CO : Ti — Tj)$

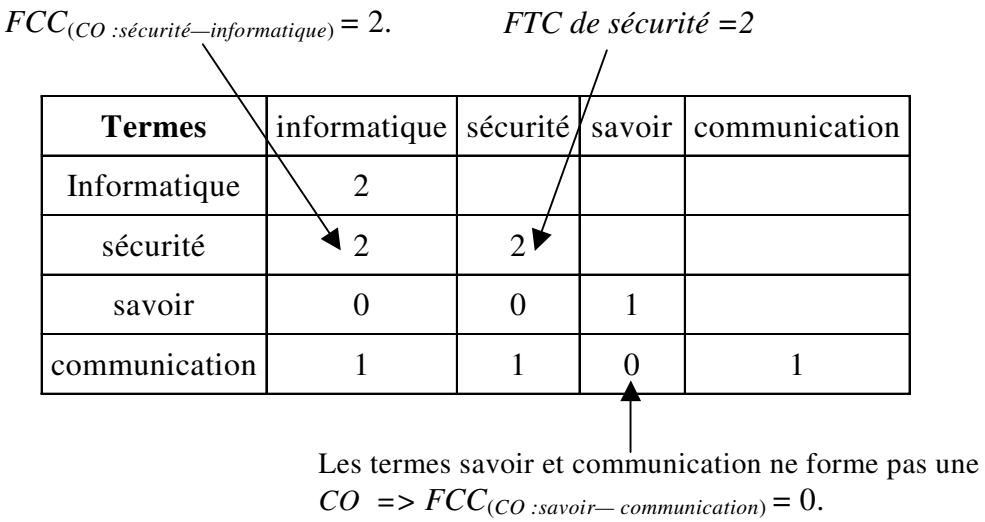
Else $MATCO\ (i,j)=FTC_{Ti}=FTC_{Tj}$

Remarque : Les cooccurrences de l'ensemble *CO2C*, i.e. ensemble des cooccurrences de la collection, peuvent être représentées par un graphe $GC = (TC, GTC)$, où *GTC* est un ensemble d'arrêtes représentant les relations de cooccurrences de l'ensemble *CO2C*.

Exemple 7 :

Dans cet exemple, nous illustrons la construction de la matrice *MATCO*. Considérons l'extrait de document prétraité suivant « Passionner sécurité informatique. savoir : sécurité informatique devenir élément phare communication moderne ». L'ensemble des termes est *TC* = {passionner,

sécurité, informatique, savoir, devenir, élément, phare, communication, moderne}. Les fréquences *FTC* associées respectivement aux termes de l'ensemble *TC* sont : 1, 2, 2, 1, 1, 1, 1, 1, 1. La Figure 4 représente une partie de la matrice *MATCO* associée aux différents termes.



Termes	informatique	sécurité	savoir	communication
Informatique	2			
sécurité	2	2		
savoir	0	0	1	
communication	1	1	0	1

Figure 4. Illustration de la matrice *MATCO*

Le graphe *GC* associé à cette partie de la matrice *MATCO* est présenté dans la Figure 5.

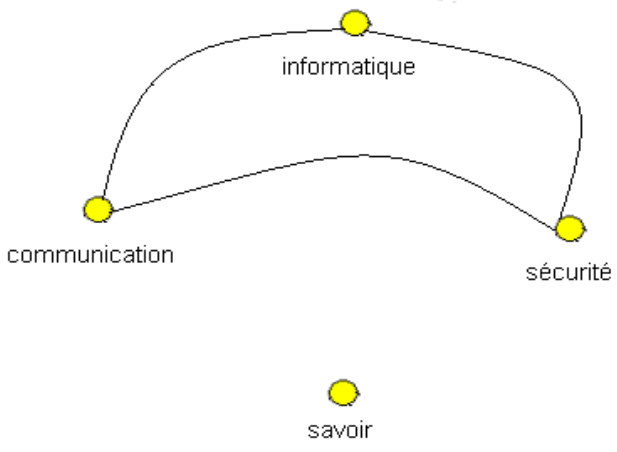


Figure 5. Exemple de graphe de cooccurrences

A partir de la matrice de cooccurrences, il est possible d'obtenir une nouvelle matrice réduite qui correspond à une matrice ne considérant que les termes représentatifs. L'objectif principal de cette matrice est d'une part de réduire le nombre de dimensions à considérer et d'autre part de se focaliser sur les termes les plus représentatifs de la collection de documents.

Définition 8 (Matrice de cooccurrences réduite)

Soit TC l'ensemble des termes de la collection C. Soient les deux sous ensembles $E_1 \subset TC$ et $E_2 \subset TC$, où $M = |E_1|$ et $K = |E_2|$. La matrice de cooccurrences réduite, notée $RMATCO$ de E_1 sur E_2 correspond à une matrice de M lignes et K colonnes. La ligne i de la matrice correspond à un terme $T_i \in E_1$ et la colonne j de la matrice correspond à un terme $T_j \in E_2$. Elle est définie par :

$i=1$;

Foreach $T_k \in E_1$ **do**

For *compteur* de 1 à N **do**

If $VTC(compteur) = T_k$ **then** $pos_1 = compteur$;

j=1 ;

Foreach $T_m \in E_2$ **do**

For *compteur* de 1 à N **do**

If $VTC(compteur) = T_m$ **then** $pos_2 = compteur$;

$RMATCO(i,j) = MATCO(pos_1, pos_2)$;

j++;

i++ ;

Remarque : Le fait de considérer les termes représentatifs a pour conséquence de réduire l'ensemble des cooccurrences de la collection. Soit $CO2C_{réduit} \subset CO2C$, l'ensemble des cooccurrences réduites, cet ensemble peut également être

représenté par un graphe *RGC* = (*RTC*, RG*TC*), où *RTC* ⊂ *TC* résulte de l'union des deux ensembles *E1* et *E2* (*RTC* = *E1* ∪ *E2*), *RGTC*⊂ *GTC*.

Dans l'exemple suivant, nous illustrons la construction de la matrice *RMATCO*.

Exemple 8 :

Considérons l'ensemble de documents donné en Annexe A. La Figure 6 représente une partie de la matrice *MATCO* correspondant aux quinze termes suivants : *VTC*={Challenge, entreprise, savoir, informatique, sécurité, classement, conseil, français, économie, Finistère, créateur, sportif, championnat, football, décision}. Par exemple, la ligne 11, colonne 2 correspond à la fréquence $FCC_{(CO : créateur - entreprise)}$ qui est égale à 7.

Termes	Challenge	entreprise	savoir	informatique	sécurité	classement	conseil	français	économie	Finistère	créateur	sportif	championnat	football	décision
Challenge	6														
entreprise	0	5													
savoir	0	0	4												
informatique	1	0	1	4											
sécurité	1	0	0	3	4										
classement	3	0	0	0	0	3									
conseil	1	4	0	0	0	0	3								
français	1	0	0	0	0	0	0	3							
économie	1	0	0	0	0	0	0	2	3						
Finistère	1	4	0	0	0	0	2	1	1	3					
créateur	0	7	0	0	0	0	2	0	0	2	3				
sportif	2	0	0	0	0	3	0	0	0	0	0	2			
championnat	2	0	0	0	0	2	1	0	0	0	0	1	2		
football	0	0	0	0	0	0	0	2	1	0	0	0	0	2	
décision	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1

Figure 6. Matrice *MATCO* de l'exemple 8

Considérons à présent le sous ensemble *E* ⊂ *TC*, réduit aux termes suivants : *E*= {challenge, conseil, économie, créateur, football, décision}. La matrice *RMATCO* de *E* sur *E*, obtenue à partir de *MATCO* de *TC* est décrite dans la Figure 7. Par exemple, pour obtenir la valeur correspondante à la

ligne 3, i.e. « économie », et la colonne 5, i.e. « football », nous recherchons ces termes dans le vecteur *VTC*. Leurs positions sont respectivement 9 et 14. Nous récupérons ensuite la valeur correspondante à la ligne 9 et la colonne 14 de la matrice *MATCO* de *TC* (ou inversement à la ligne 14 et la colonne 9 puisque la matrice est symétrique) et nous obtenons : $RMATCO(2,1) = MATCO(9,14) = MATCO(14,9) = 1$.

Termes	challenge	conseil	économie	créateur	football	décision
challenge	6					
conseil	1	3				
économie	1	0	3			
créateur	0	2	0	3		
football	0	0	**1**	0	2	
décision	0	0	0	0	1	1

Figure 7. Illustration de la matrice *RMATCO* de *E* sur *E*

1.3. Algorithmes

Cette section décrit plus formellement l'algorithme général (C.f. Algorithme 1) défini pour obtenir les matrices de cooccurrences. Tout d'abord, nous considérons que les documents initiaux sont prétraités (fonction *pretraitement*(D')) et nous obtenons une collection de documents. Les termes de la collection sont extraits via la fonction *getTermes*(C). Les vecteurs $VFTC = <FTC_1, \ldots, FTC_i, \ldots, FTC_n>$ et $VFTD = <FTD_1, \ldots, FTD_i, \ldots, FTD_n>$ correspondant respectivement aux vecteurs de fréquences des termes dans la collection (calculées par la fonction *OccFreqTerme*) et en documents (calculées par la fonction *DocFreqTerme*). Ils sont obtenus comme expliqué dans la section précédente. La matrice *MATCO* (C.f. Définition 7) est obtenue à partir de l'ensemble des termes *TC* et de l'ensemble des cooccurrences des termes *CO2C*. Enfin nous recherchons quels sont les termes les plus représentatifs, i.e. *RTC*, (fonction *TermesReprésentatifs)* à partir des fréquences des termes. Ces derniers nous permettent d'obtenir la matrice réduite *RMATCO*.

Algorithme général d'analyse

Input: Un ensemble de documents $D'=\{D'_1, \ldots, D'_i, \ldots, D'_n\}$;

Output: Une collection $C=\{D_1, \ldots, D_i, \ldots, D_n\}$;

Une matrice *MATCO* ;

Un ensemble de termes *RTC* ;

Une matrice réduite *RMATCO*.

Begin

// Prétraitement de l'ensemble des documents D' pour obtenir une collection C

C= pretraitement(D') ;

// Extraction de l'ensemble des termes TC ;

TC=getTermes(C) ; nombredetermes= |TC| ;

// Calcul des fréquences des termes dans la collection

VFTC←∅ ;

Foreach $T_i \in TC$ **do**

FTC_{Ti} = *OccFreqTerme* (T_i) ;

Update (VFTC(i)) ;

// Calcul des fréquences des termes en documents

VFTD←∅ ;

Foreach $T_i \in TC$ **do**

FTD_{Ti} = *DocFreqTerme* (T_i) ;

Update (VFTD(i)) ;

// Extraction de l'ensemble des cooccurrences contextuelles CO2C

CO2C=geTCoOccurrences(TC, VTFC, C) ;

// Construction de la matrice MATCO de TC

MATCO←∅ ;

Foreach $(CO : T_i — T_j) \in CO2C$ **do**

Calculer $FCC_{(CO : Ti — Tj)}$;

Update(*MATCO*(i,j)) ;

// Construction de l'ensemble des termes représentatifs

RTC = TermesReprésentatifs (TC, VFTC, VFTD, MATCO) ;

// Construction de la matrice RMATCO de RTC sur RTC

$RMATCO \leftarrow \varnothing$;

Foreach $(CO : T_i — T_j) \in RCO2C$ **do**

Récupérer $FCC_{(CO : Ti — Tj)}$;

Update(*RMATCO*(*i,j*)) ;

End

Algorithme 1. Algorithme général d'analyse

Considérons à présent comment les termes représentatifs sont obtenus.

Les termes représentatifs correspondent à un sous ensemble de l'ensemble des termes *TC*, i.e. il s'agit d'une réduction de l'ensemble *TC*. Comme nous l'avons vu dans le chapitre précédent, le critère de sélection de termes le plus répondu dans la littérature est le pouvoir de discrimination des termes introduit par *G*. Salton [Sal 1983]. Ce critère consiste à sélectionner les termes les plus fréquents appartenant à un nombre réduit de documents, i.e. prise en considération du nombre de documents contenant un terme donné (fréquence en documents) en plus de sa fréquence d'occurrence. Ce critère permet de garantir l'indexation de la globalité des contenus textuels.

Dans le cadre de notre approche, nous considérons également un nouveau critère, que nous appelons « critère de représentativité ». Il permet une représentativité du contenu lui-même, i.e. le sens enfoui dans les collections de textes, et prend en considération les fréquences *FCC* de cooccurrences de termes. Dans le chapitre suivant, nous montrerons via une application que ces deux critères sont complémentaires pour mieux représenter les informations enfouies dans les collections.

L'ensemble des termes représentatifs, *RTC*, résulte en fait de l'union de deux ensembles. Le premier ensemble consiste à choisir des termes en fonction de leurs

fréquences d'occurrences FTC_i en tenant en compte de leurs distributions FTD_i dans la collection *C*. Le second ensemble résulte d'un choix de termes en fonction des cooccurrences les plus fréquentes dans la collection *C*.

Notre méthode de choix des termes représentatifs de la collection *C* est décrite dans l'algorithme *TermesReprésentatifs* (C.f. Algorithme 2). Elle est basée sur deux seuils α et β qui sont exprimés respectivement en fonction de la moyenne *MoyT* des FTC_i des termes T_i pondérées par les FTD_i et de la moyenne normalisée *MoyCO* des *FCC* des cooccurrences contextuelles *CO*. Soit *M* le nombre des termes de la collection et $N = |CO2C|$ le nombre de cooccurrences de la collection. *MoyT* et *MoyCO* peuvent être données comme suit.

$$MoyT = \frac{\sum_{i=1}^{M} (FTC_i / FTD)}{M}$$

$$MoyCO = \frac{\sum_{j=1}^{N} FCC_i}{N^2}$$

Algorithme Termes Représentatifs

Input: *TC* ensemble des termes de la collection *C* ;

Vecteurs *VFTC* et *VFTD* ; $n = |CO2C|$;

Matrice MATCO de TC ;

Output: *RTC,* l'ensemble des Termes Représentatifs de *C.*

Begin

$RTC \leftarrow \varnothing$;

Foreach $T_i \in TC$ **do**

// *Les paramètres* α *et* β *correspondent aux seuils de sélection des termes*

If $(FTC_i / FTD_i) > \alpha$

then $RTC = RTC \cup \{ T_i \}$;

Foreach $(CO : T_i — T_j) \in CO2C$ **do**

If $(FCC_{(CO : Ti — Tj)} / n) > \beta$

then $RTR = RTR \cup \{ T_i \} \cup \{ T_j \}$;

End

Algorithme 2. Recherche des termes représentatifs de la collection

A l'issue de cette étape, nous obtenons donc une matrice réduite *RMATCO* de cooccurrences des termes représentatifs. Dans la section suivante, nous montrons comment étendre *RMATCO* pour prendre en compte les partages de contextes.

2. Partage de contextes

2.1. Définitions préliminaires

Avant d'exposer l'algorithme de représentation des termes de l'ensemble *RTC*, nous complétons les définitions de la section précédente.

Définition 9

Soient les deux termes représentatifs $T_i \subset RTC$ et $T_j \subset RTC$, nous notons $(T_i \wedge T_j)$ l'ensemble des termes représentatifs appartenant au moins une fois à l'un des contextes de cooccurrences de T_i et à l'un des contextes de cooccurrences de T_j. L'ensemble est défini de la manière suivante :

$$(T_i \wedge T_j) = \{T_k \in RTC / (CO : T_i — T_k) \wedge (CO : T_j — T_k)\}$$

Définition 10

Soient les deux termes représentatifs $T_i \subset RTC$ et $T_j \subset RTC$, nous notons $(T_i \wedge \neg T_j)$ l'ensemble des termes représentatifs appartenant au moins une fois à l'un des contextes de cooccurrences de T_i et qui n'appartiennent à aucun

des contextes de cooccurrences de T_j. L'ensemble est défini de la manière suivante :

$$(T_i \wedge \neg T_j) = \{T_k \in RTC \,/\, (CO : T_i — T_k) \wedge \neg(CO : T_j — T_k)\}$$

où $\neg(CO : T_j — T_k)$ signifie que le couple de termes $<T_j, T_k>$ ne forme pas une cooccurrence contextuelle *CO*.

Remarque : par analogie, nous notons $(\neg T_i \wedge T_j)$ l'ensemble des termes représentatifs appartenant au moins une fois à l'un des contextes de cooccurrences de T_j et qui n'appartiennent à aucun des contextes de cooccurrences de T_i.

De manière à illustrer ces deux définitions, considérons l'exemple suivant :

Exemple 9 :

Soit le graphe illustré Figure 8 où les sommets représentent des termes représentatifs et les arrêtes représentent des relations de cooccurrences entre termes.

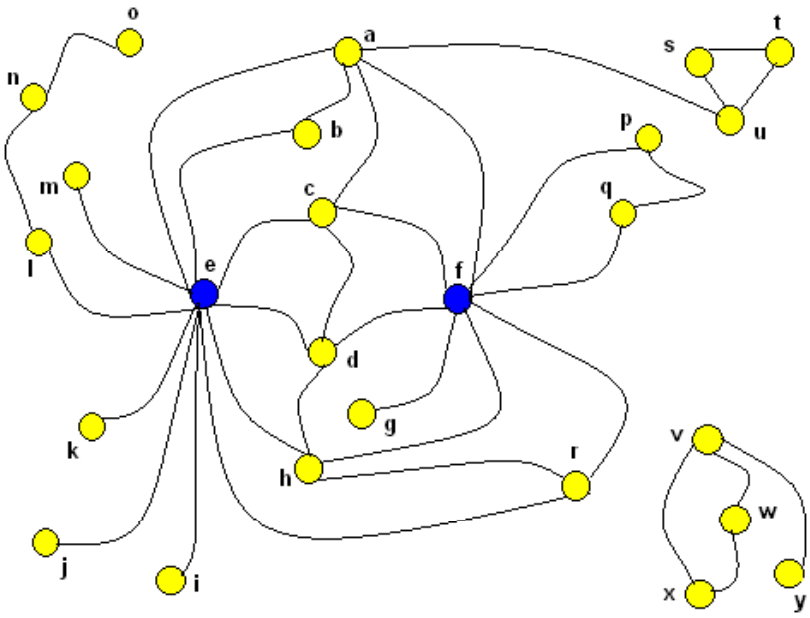


Figure 8. Un exemple de graphe

En adéquation avec les deux définitions précédentes, les ensembles $(e \wedge f)$, $(e \wedge \neg f)$ et $(\neg e \wedge f)$ associés au couple de termes représentatifs $< e, f >$ sont les suivants :

- $(e \wedge f) = \{ a, c, d, h, r \}$
- $(e \wedge \neg f) = \{b, m, l, k, j, i\}$
- $(\neg e \wedge f) = \{g, p, q\}$

2.2. Critère de partage de contextes

Dans l'approche précédente, nous avons retenu comme critère : le critère de cooccurrence contextuelle, i.e. deux termes sont en cooccurrences s'ils appartiennent à la même phrase. Ce critère est utile pour regrouper ensemble deux termes qui sont suffisamment proches pour être reliés ensemble. Considérons à présent l'exemple de collection suivant composé des extraits de documents présentés en annexe A : « *Les majors ont défendu dans la loi sur l'économie numérique, le filtrage aux frontières sur internet. Les CCI du Finistère organisent pour la sixième fois le challenge des espoirs de l'économie française* ». Parmi les cooccurrences que l'on peut trouver nous avons, par exemple : (*CO* : économie—internet) et (*CO* : challenge—économie). Nous constatons, qu'étant donné que deux termes apparraissent en cooccurrences s'ils appartiennent à la même phrase, nous ne pouvons pas déterminer qu'il existe une relation importante entre « internet » et « challenge ». L'une des solutions pour obtenir ce type de liens est de relâcher la contrainte de fenêtre de termes limitée à une phrase. Nous reviendrons sur ce point lors de la discussion de ce chapitre. Nous proposons un nouveau critère appelé « critère de partage de contextes ». Ce dernier permet d'étudier la relation entre un terme *A* et un terme *B* ainsi que les possibilités de rattachement de ces derniers s'ils ont des termes communs en cooccurrences. Nous pouvons comparer cette approche à la notion de « l'ami de mon ami est mon ami ». Via le critère de « partage de contextes », nous avons la possibilité de détecter « les amis proches des amis lointains » ou « ceux qui ne sont pas du tout amis ».

A l'aide des deux principes ou critères cités auparavant « cooccurrences contextuelles » et « partage de contextes », nous représentons les termes de l'ensemble *RTC* par deux matrices. La première matrice, notée $MatR_1$, reprend les données de la matrice *RMATCO* en mettant l'accent sur l'importance des relations de cooccurrences d'un terme donné de l'ensemble *RTC* par rapport à son occurrence dans la collection. La deuxième matrice ($MatR_2$) prend en

considération le critère de « partage de contextes » entre les termes représentatifs, détaillé dans cette section en se basant sur les définitions préliminaires.

L'Algorithme 3 décrit comment les matrices $MatR_1$ et $MatR_2$ sont calculées, la matrice *RMATCO* est considérée comme une matrice intermédiaire, elle est libérée (supprimée) à l'issue de la représentation de l'ensemble *RTC*.

Algorithme Représentation de l'ensemble *RTC*

Input: $RTC = \{T_1,...,T_m\}$;

RMATCO de RTC sur RTC ; $m = |RTC|$;

Output: Matrices $MatR_1$ et $MatR_2$;

Begin

For (i=1; i<= m; i++) **do**

For (j=1; j<= m; j++) **do**

If $RMATCO(i,j) \neq 0$

then $MatR_1(i,j) = RMATCO(i,j) / RMATCO(i,i)$

else $MatR_1(i,j) = 0$;

$l \leftarrow 1$;

Foreach $T_k \in RTC$ **do**

$c \leftarrow 1$;

Foreach $T_m \in RTC$ **do**

Construire l'ensemble $Ens_1 = (T_k \wedge T_m)$;

Construire l'ensemble $Ens_2 = (T_k \wedge \neg T_m)$;

$MatR_2(l,c) = |Ens1| / (|Ens_1| + |Ens_2|)$;

c++;

l++ ;

End

Algorithme 3. Représentation de l'ensemble *RTC*

Exemple 10 :

Considérons la matrice *MATCO* de *TC* associée à la collection de documents mis en Annexe A. Une partie de cette matrice est représentée dans la Figure 6 de l'Exemple 8. Considérons également la matrice *RMATCO* (C.f. Figure 7). Rappelons que cette dernière est associée aux termes suivants : challenge, conseil, économie, créateur, football, décision. A partir de la matrice intermédiaire *RMATCO*, la matrice $MatR_1$ est construite par l'Algorithme 3, cette matrice reflète les cooccurrences d'un terme donné par rapport à ses occurrences, i.e. le critère de « cooccurrences contextuelles ». La matrice $MatR_1$ est représentée dans la Figure 9. Nous remarquons, par exemple, dans cette figure que le terme « créateur » apparaît deux fois dans deux contextes différents avec « conseil » parmi ses 3 occurrences dans la collection *C*. Nous remarquons également que le terme « décision » est apparu une seule fois dans le même contexte que « football » parmi sa seule occurrence dans la collection *C*.

Termes	challenge	conseil	économie	créateur	football	décision
challenge	1	1/6	1/6	0	0	0
conseil	1/3	1	0	2/3	0	0
économie	1/3	0	1	0	1/3	0
créateur	0	2/3	0	1	0	0
football	0	0	1/2	0	1	1/2
décision	0	0	0	0	1	1

Figure 9. La Matrice $MatR_1$

La matrice $MatR_2$ est construite à l'aide de l'Algorithme 3 basé sur la Définition 9 et la Définition 10. Par exemple, pour calculer la valeur de la case correspondante à la ligne 1 « challenge » avec la colonne 5 « football », l'ensemble (challenge $\wedge$ football) est construit suivant la Définition 9, ce dernier est formé d'un seul élément « économie ». Puis l'ensemble (challenge $\wedge \neg$football) est construit également en se basant sur la Définition 10. Ce dernier est composé également d'un seul élément « conseil ». Donc suivant l'Algorithme 3, l'élément de $MatR_2$

correspondant à la ligne 1 et la colonne 5 vaut 1/(1+1) = 1/2. La matrice $MatR_2$ est illustrée dans la Figure 10.

Termes	challenge	conseil	économie	créateur	football	décision
challenge	1	0	0	1/2	1/2	0
conseil	0	1	1/2	0	0	0
économie	0	1/2	1	0	0	1/2
créateur	1	0	0	1	0	0
football	1/2	0	0	0	1	0
décision	0	0	1	0	0	1

Figure 10. La matrice $MatR_2$

Nous pouvons remarquer que les deux matrices $MatR_1$ et $MatR_2$ ne sont pas symétriques. En effet, pour la matrice $MatR_1$, les fréquences d'occurrences de deux termes T_1 et T_2 en cooccurrence peuvent être différentes. Par exemple, avec les fréquences d'occurrences différentes des termes « football » et « décision », nous avons $MatR_1$ (5,6)= 1/2 tandis que $MatR_1$ (6,5)=1. En effet, parmi les occurrences du terme « football » ces cooccurrences sont concentrées avec « décision » tandis que l'inverse n'est pas vrai, i.e. parmi les occurrences de « décision » elles sont concentrées avec « football ». Pour la matrice $MatR_2$, cela est dû au fait que pour deux termes T_1 et T_2 en cooccurrences, les cooccurrences de T_1 ne sont pas forcément les mêmes que les cooccurrences que T_2. Par exemple, le terme « challenge » (correspondant à la ligne 1 de la matrice $MatR_2$) n'a pas les mêmes « cooccurrences » que le terme « créateur » (correspondant à la colonne 4 de la matrice $MatR_2$), i.e. $MatR_2(1,4)=1/2$; $MatR_2(4,1)=1$. ($MatR_2(1,4) \neq MatR_2(4,1)$).

Analysons à présent la relation « challenge–créateur ». Ces deux termes ne sont pas en cooccurrence, i.e. $MatR_1(1,4)= MatR_1(4,1)= 0$. Donc le critère de « cooccurrences contextuelles » ne nous apporte pas beaucoup d'informations à part que ces deux termes n'ont pas de relation à priori. Observons à présent la matrice $MatR_2$ pour les mêmes termes. Nous avons $MatR_2(1,4)=1/2$ et $MatR_2(4,1)=1$. Ceci signifie que les deux termes « challenge » et «créateur » renferme une relation malgré le fait qu'ils ne

sont pas en relation directe de cooccurrences. En effet, le terme « challenge » est apparu dans le même contexte que « conseil » et le terme « créateur » est apparu dans un autre même contexte que « conseil ». Donc les deux termes « challenge » et « créateur » sont susceptibles d'être « amis ». Le fait que $MatR_2(1,4)$ est différente de $MatR_2(4,1)$ et qu'une valeur est plus élevée que l'autre, peut s'interpréter de la manière suivante : le terme « challenge » est susceptible d'être ami intime de « créateur » tandis que le terme « créateur » est seulement susceptible d'être ami du terme « challenge ». En résumé le critère de « partage de contextes » permet d'étudier des « relations » complémentaires au critère de « cooccurrences contextuelles ».

3. Conclusion et discussion

Dans ce chapitre nous avons proposé une nouvelle approche de représentation des collections de documents textuels. Après une étape de prétraitements, nous obtenons une collection de documents sur laquelle nous effectuons différentes opérations pour représenter le mieux possible la connaissance. A l'issue de cette approche nous disposons de deux matrices une matrice $MatR_1$ reflétant les fréquences de termes et de cooccurrences les plus représentatives mais également d'une matrice $MatR_2$ basée sur la notion de « partage de contextes ».

L'originalité de cette approche est d'une part d'être indépendante des algorithmes de prétraitements associés tant qu'ils respectent la relation d'ordre entre les différents termes des documents. D'autre part, elle propose de prendre en compte lors de la première phase de traitement non seulement les fréquences des termes (nombre d'occurrences des termes dans la collection *FTC* et nombre de documents contenant les termes *FTD*) mais également celles des cooccurrences. En effet, dans les approches traditionnelles d'utilisation de cooccurrences comme celle du modèle *DSIR* présentée dans le chapitre précédent, les cooccurrences ne sont établies qu'après avoir effectué une étape de recherche des termes significatifs afin de relever quels sont les termes représentatifs qui sont en cooccurrences. En considérant directement les fréquences de cooccurrences dans *MATCO*, i.e. entre tous les termes de la collection, nous n'éliminons pas de connaissances a priori comme dans le cas de *DSIR* car nous conservons les cooccurrences entre termes qui ne sont pas forcément discriminants. De manière à illustrer la différence, considérons l'exemple suivant. Soient les trois textes D_1= « *Java est un langage de programmation objet.*», D_2= « *Java est le plus utilisé dans les Ecoles d'Ingénieurs.*», D_3= « *La programmation objet fait référence au langage Java.* ». Avec l'approche *DSIR*, le terme Java apparaissant dans les trois textes, i.e. *TF/IDF*=3/3=1, il ne sera pas retenu car non discriminant. Dans le cadre de notre

approche, le terme Java ne sera pas considéré comme un terme discriminant pour les mêmes raisons. Par contre, il sera retenu car il existe deux fois la cooccurrence Java—Objet, i.e. la fréquence de cooccurrence Java—Objet est importante par rapport au reste des cooccurrences.

Au cours de la section 2.2, nous avons défini un nouveau critère appelé le critère de « partage de contextes ». Nous revenons dans cette discussion sur le choix de définir un nouveau critère plutôt que d'étendre la notion de contexte. Dans le chapitre précédent, nous avons vu qu'il existait de nombreuses manières de définir le contexte : phrase, fenêtre de mots, paragraphes. Nous avons retenu dans notre approche la notion de phrase. Ce choix est guidé par le fait que des mots qui apparaissent dans la même phrase sont « sémantiquement proches » [Her 2004]. L'utilisation d'une fenêtre de mots pourrait également être utilisée mais la difficulté principale est la définition de la taille de la fenêtre : comment savoir a priori la taille à fixer ? Le critère que nous avons défini permet d'une part de contourner ce problème car la taille est fixée en fonction du contenu du document. L'autre avantage est bien entendu de réduire l'espace de recherche. En effet, en ne considérant que des termes qui apparaissent dans les cooccurrences nous n'avons plus qu'à rechercher les termes qui sont situés à l'intersection des cooccurrences afin de vérifier leurs associations.

Dans le chapitre suivant, nous montrons comment le modèle de représentation que nous avons proposé dans ce chapitre peut être utilisé dans deux domaines d'applications complémentaires. Nous verrons également que les résultats, obtenus par notre approche, peuvent facilement être adaptés pour donner un aperçu significatif des connaissances incluses dans les collections de documents.

Chapitre IV – Le Système IC-DOC

Ce chapitre décrit le système *IC-DOC* et deux domaines d'applications complémentaires de notre modèle de représentation de collections de documents. Dans le premier cas, nous montrons comment notre modèle décrit dans le chapitre précédent peut être utilisé pour aider à identifier des groupes de clusters de thématiques différentes. Le second type d'application permet de cartographier l'information utile et de visualiser la connaissance enfouie dans les contenus textuels.

Le chapitre est organisé de la manière suivante. Dans la section 1, nous présentons le système *IC-DOC* développé et décrivons ses principaux modules. L'objectif de la section 2 est de présenter l'utilisation du système *IC-DOC* pour identifier des clusters de thématiques. Au cours de cette section, nous revenons sur les principes généraux du clustering et présentons quelques unes des approches existantes. Dans la section 3, nous nous intéressons à la visualisation des connaissances extraites et nous présentons comment le système *IC-DOC* est étendu pour intégrer cette composante. Au cours de ces deux dernières sections, nous présentons également quelques expériences menées. Enfin, dans la section 4, nous concluons ce chapitre en résumons l'intérêt de notre modèle mis en œuvre via le système *IC-DOC*.

1. Le système IC-DOC

L'objectif du système *IC-DOC* est de proposer un environnement d'extraction de connaissances pour des données textuelles issues de thématiques différentes. Ses principes généraux sont expliqués dans la Figure 11. Ils sont similaires au processus classique d'extraction de connaissances à partir de textes que nous avons présenté dans le chapitre II.

La démarche se décompose en quatre phases principales. Tout d'abord, à partir des documents initiaux, un prétraitement est nécessaire pour pouvoir manipuler les contenus textuels et réduire les données à conserver lors des étapes suivantes. Dans la seconde phase, nous modélisons les documents à l'aide des différentes matrices que nous avons présenté dans le chapitre précédent. Un algorithme de fouille de données est ensuite utilisé pour extraire la connaissance. Enfin, l'interprétation et l'exploitation des résultats obtenus est facilité par un outil de cartographie et de visualisation des connaissances.

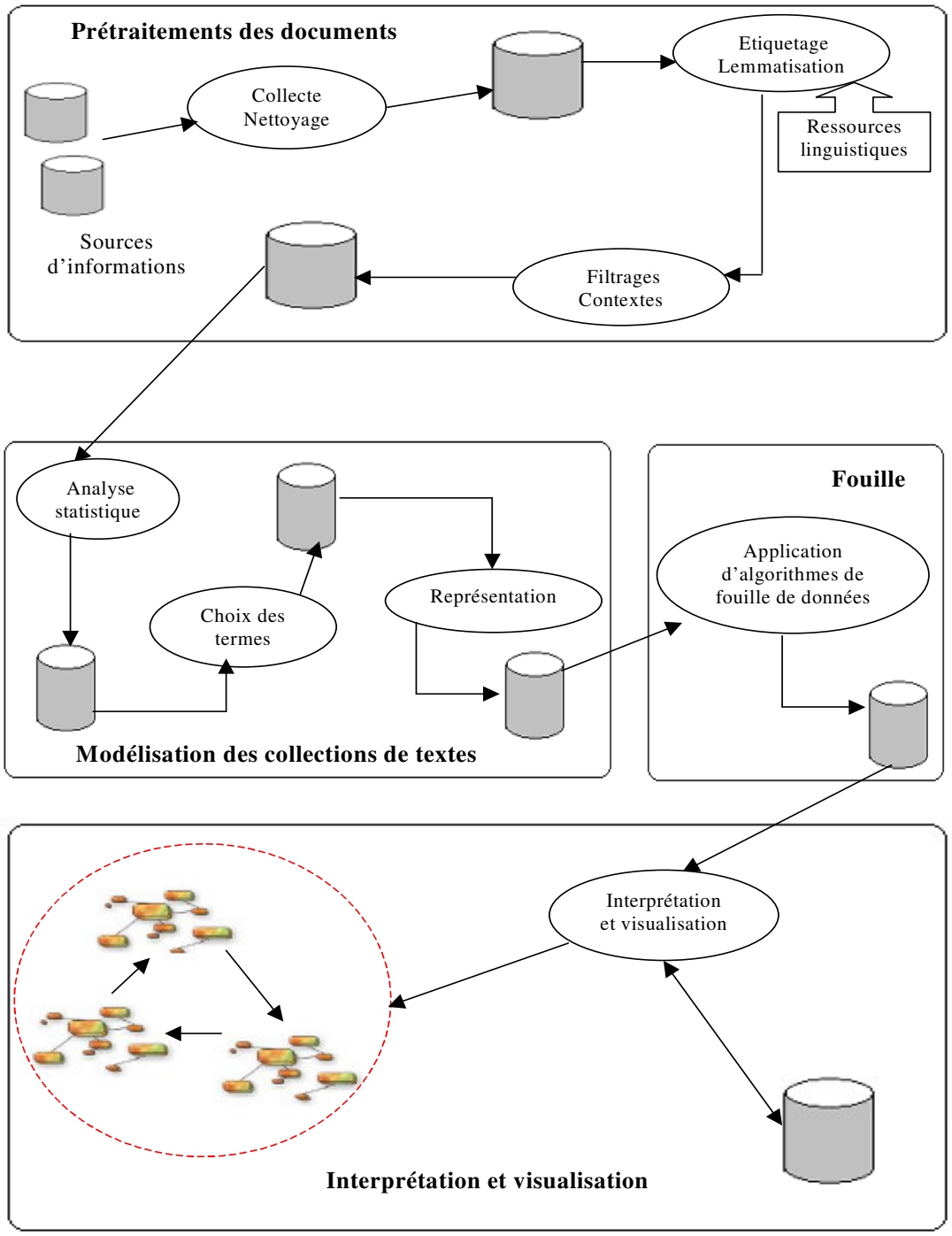


Figure 11. Architecture générale du système IC-DOC

Dans la suite de cette section nous décrirons les deux modules de prétraitements et de modélisation des documents. Ensuite, nous présenterons brièvement le module de fouille et le module d'interprétation et de visualisation dans la mesure où nous reviendrons sur ces derniers de manière détaillée dans les sections 2 et 3. Nous

préciserons également la manière dont les modules s'enchaînent. Enfin, pour chaque module, nous décrierons les choix d'implémentations effectués au sein de notre prototype.

1.1. Prétraitement des documents

Le prétraitement des documents est réalisé dans *IC-DOC* en suivant les étapes suivantes :

1. Après une collecte de documents, un nettoyage est effectué pour transformer les documents (traitement des majuscules, des numéros, …, etc.) ou éliminer les éléments non représentatifs (e.g. balises html dans le cas de documents issus du Web) de façon à ne retenir que des contenus textuels.

2. Une analyse morphosyntaxique est ensuite effectuée à l'aide de la bibliothèque de fonctions de « Cordial Analyser ». Cette dernière est une *DLL* développée en langage *C++* que nous exploitons via notre propre interface de programmation.

3. Les documents sont aussi lemmatisés en utilisant les techniques de « Cordial Analyser » via la même interface de programmation.

Nous avons vu dans le chapitre II que la lemmatisation permet de diminuer fortement le nombre de mots à analyser en éliminant toutes les flexions et les dérivations grammaticales. Elle ramène ainsi chaque mot en une forme unique. Après la phase de lemmatisation de la collection de documents textuels via *IC-DOC*, chaque document est transformé en une suite de lemmes. Nous avons choisi d'utiliser la lemmatisation plutôt que d'utiliser le stemming pour sa puissance de distinction entre les catégories des mots. Via les algorithmes de stemming, le nom « porte » et le verbe « porter » sont tous les deux ramenés à la même forme « port », et donc nous perdons la distinction entre « le porte parole », « projet porté » ou « port USB ». Via la lemmatisation et les étiquettes morphosyntaxiques, nous distinguons les noms des verbes, donc le verbe « porter » n'aura pas la même forme que le nom « porte ». La Figure 12 illustre un exemple d'étiquetage morphosyntaxique et de lemmatisation réalisé à l'aide de la bibliothèque de fonctions « Cordial Analyser ».

……

==== DEBUT DE PHRASE ====

Le	1	le	ADMS (Article défini masculin singulier)
Centre	2	centre	NCMS (Nom commun masculin singulier)
du	3	du	AIMS (Article indéfini masculin singulier)
Val	4	val	NCMS (Nom commun masculin singulier)
est	5	être	Verbe indicatif présent 3ème personne singulier
sans	6	sans	Prep (préposition)
Parking	7	parking	NCMS (Nom commun masculin singulier)

……

==== FIN DE PHRASE ====

==== DEBUT DE PHRASE ====

……

Figure 12. Un exemple d'étiquetage et de lemmatisation

1.2. Modélisation des documents

Après les prétraitements des documents par le premier module du système *IC-DOC*, le second module : modélisation de documents, se charge de la représentation des collections de textes suivant le modèle décrit dans le chapitre III. Ce module est composé d'une collection d'outils exploitant les structures de données fournis par le module de prétraitements [Mok&al 2006, Mok&al 2005a, Mok&al 2004b]. Parmi ces outils, *CO2MAT* (*Contextual Cooccurrence Matrix*) permet de construire la matrice de cooccurrences *MATCO* de *TC.* A partir de cette matrice l'ensemble *RTC* des termes représentatifs du contenu sont extraits suivant l'Algorithme 2 décrit dans le chapitre III. Enfin, les données de la matrice *MATCO* sont réduites sur l'ensemble *RTC* pour représenter l'ensemble des contenus textuels à l'aide de l'Algorithme 3 du chapitre précédent.

1.3. Fouille de données

Nous avons vu dans le chapitre II qu'à partir de la modélisation de documents, plusieurs techniques de fouille de données peuvent être appliquées. Dans le cadre du système *IC-DOC*, le module de fouille exploite la représentation des collections de

textes fournie par le module 2 et applique des techniques de fouille aux éléments associés aux deux matrices $MatR_1$ *et* $MatR_2$, i.e. l'ensemble *RTC* des termes représentatifs.

Dans le cadre de nos expérimentations, nous nous sommes intéressés aux regroupements thématiques associés à une collection de documents. Ce module a donc été développé pour pouvoir fonctionner avec des algorithmes de clustering. Nous détaillerons cet aspect dans la section 2.

1.4. Interprétation et visualisation

Après l'application d'un algorithme de fouille de données via le module 3. Ce dernier renvoie les résultats obtenus au module 4 d'interprétation et de visualisation. Ce module exploite les résultats de la fouille pour permettre de visualiser la connaissance. La première phase du module de visualisation consiste à interpréter les résultats fournis par le module de fouille. Cette interprétation est basée principalement sur deux mesures de pertinence largement utilisées : la précision et le rappel [Poi 2003]. Ces mesures sont définies comme suit.

Soit l'ensemble de données analysées et les résultas fournis par le système illustrés dans la Figure 13. Après analyse, les données sont découpées en deux ensembles, données extraites et non extraites par le système. Chaque ensemble est découpé lui-même en deux sous ensembles, données répondant à l'objectif et données ne répondant pas à l'objectif. Pour les données extraites par le système, les deux sous ensembles sont appelés *VP* (*Vrais Positifs*) et *FP* (*Faux Positifs*). Pour celles non extraites par le système, les deux sous ensembles sont appelés *VN* (*Vrais Négatifs*) et *FN* (*Faux Négatifs*).

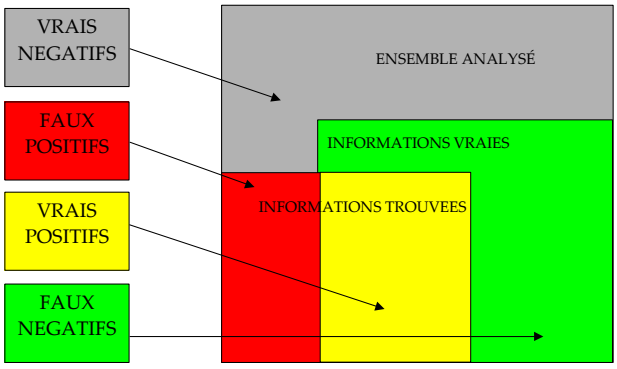


Figure 13. Exemple d'interprétation des données analysées

En se basant sur les données illustrées sur la Figure 14, les mesures de précision (pourcentage de ce qui est correct, retourné par le système : sélectivité) et de rappel (pourcentage de ce qui est correct, retourné par le système, par rapport à ce qui devrait être retourné : sensibilité) sont données par les formules suivantes :

$$\text{Précision} = VP/(VP+FP)$$

$$\text{Rappel} = VP/(VP+FN)$$

Nous montrons dans la section 3 plus en détail, l'utilité de ce module pour cartographier et visualiser des informations sur le contenu global de la collection de documents ou sur certaines thématiques de la collection.

2. Identification de clusters de thématiques

Dans cette section, nous détaillons le module fouille de données que nous avons réalisé pour le système *IC-DOC* et qui permet de rechercher des groupements thématiques dans une collection de documents. L'approche de fouille retenue étant basée sur des algorithmes de classification non supervisée (*clustering*), nous en présentons tout d'abord les principes puis décrivons quelques unes des principales approches existantes. Nous revenons dans la section 2.2 sur les modifications apportées au prototype *IC-DOC* et sur les choix effectués. Au cours de la section 2.3, nous présenterons quelques unes des expérimentations menées.

2.1. Classification non supervisée : une introduction

2.1.1. Principes de base

Au sens large, la classification non supervisée (segmentation, groupement, clustering) [Jai&Dub 1998, Jai&al 1999, Ber 2002] peut être définit comme un processus permettant d'organiser des objets (dont on ne connaît pas la classe) en groupes dont les éléments sont proches. Un groupe (cluster) est donc une collection d'objets qui sont supposés être similaires entre eux et être différents des objets appartenant aux autres groupes. Dans la plupart des cas les objets sont représentés par des vecteurs de caractères (attributs) les décrivant. Le processus de clustering

consiste alors à regrouper ces vecteurs en respectant au mieux les propriétés suivantes :

- Homogénéité dans les groupes : les données appartenant à un même cluster doivent être les plus similaires possibles.

- Hétérogénéité entre groupes : les données appartenant à différents clusters doivent être les plus dissemblables possibles.

La Figure 14 illustre le processus de classification non supervisée pour un espace en deux dimensions. Les objets représentés par des points sont décrits par deux caractères X_1 et X_2 (Cf. Figure 14a). Le clustering consiste à déterminer automatiquement les groupes de points les plus « proches » (Cf. Figure 14b).

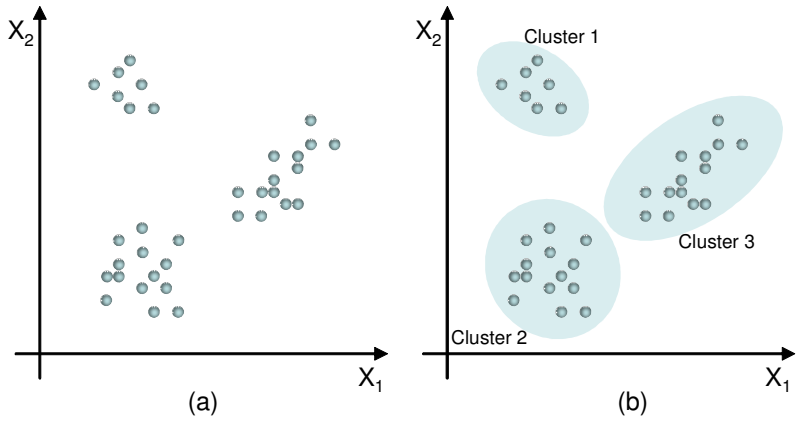


Figure 14 . Illustration du processus de classification non supervisée

La classification non supervisée est très utile dans les domaines pour lesquels les classes ou les catégories ne sont pas définies ou mal caractérisées. Les méthodes de clustering s'appliquent dans de nombreux domaines tels que :

- L'analyse d'images médicales : les méthodes de classification non supervisées sont par exemple utilisées pour déterminer des tumeurs sur des images de scanners.

- La biologie : le clustering est utilisé pour déterminer des groupes de gènes possédant des fonctions proches.

- Le marketing : la classification non supervisée permet de déterminer des ensembles d'individus possédant des comportements de consommation identiques.

Plus formellement, considérons un ensemble *X* d'objets (points) représentés par des vecteurs X_i de d attributs :

$$X_i = \begin{bmatrix} x_i^1 & \cdots & x_i^k & \cdots & x_i^d \end{bmatrix}$$

$$X = \begin{bmatrix} x_1^1 & \ldots & x_1^k & \ldots & x_1^d \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_i^1 & \ldots & x_i^k & \ldots & x_i^d \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_n^1 & \ldots & x_n^k & \ldots & x_n^d \end{bmatrix}$$

La classification non supervisée consiste à attribuer à chaque vecteur X_i un cluster d'appartenance. Le nombre de clusters q peut être défini manuellement ou automatiquement par la méthode de clustering.

Soit *C* l'ensemble de ces clusters :

$$C = \{c_1, \cdots, c_k, \cdots, c_q\}$$

Les clusters peuvent être de deux types : exclusif ou avec recouvrement. Pour le premier type, un vecteur de données ne peut appartenir qu'à un et unique cluster. Pour le second type, un vecteur peut appartenir à plusieurs clusters avec différents degrés d'appartenance.

2.1.2. Notions de proximité, similarité, dissimilarité, distance

Les techniques de clustering que nous allons étudier font référence à la notion de proximité entre deux vecteurs. Dans [Jai&Dub 1998], les auteurs définissent cette notion en introduisant un indice de proximité entre deux vecteurs X_i et X_j noté $d(X_i, X_j)$ qui satisfait les propriétés suivantes :

a) pour une similarité :

$$\forall X_i \quad d(X_i, X_i) \geq \max_k d(X_i, X_k)$$

b) pour une dissimilarité :

- $\forall X_i \quad d(X_i, X_i) = 0$
- $\forall X_i, \ \forall X_j \quad d(X_i, X_j) \geq d(X_j, X_i)$
- $\forall X_i, \ \forall X_j \quad d(X_i, X_j) \geq 0$

Les mesures de dissimilarité les plus utilisées sont les mesures de distance.

La distance euclidienne reste la plus utilisée. Elle est donnée par la formule suivante.

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{d} \left(x_i^k - x_j^k \right)^2 \right]^{1/2} = \left\| X_i, X_j \right\|_2$$

avec $X_i = \left[x_i^1 \quad \cdots \quad x_i^k \quad \cdots \quad x_i^d \right]$ et $X_j = \left[x_j^1 \quad \cdots \quad x_j^k \quad \cdots \quad x_j^d \right]$ deux vecteurs de dimension d.

Cette distance est un cas particulier (p=2) de la distance de Minkowski donnée par la formule suivante.

$$d_p(X_i, X_j) = \left[\sum_{k=1}^{d} \left(x_i^k - x_j^k \right)^p \right]^{1/p} = \left\| X_i, X_j \right\|_p$$

Ces mesures posent un problème lorsque les échelles des données ne sont pas homogènes. En effet, les grandes échelles sont favorisées dans le calcul de la distance. Pour palier ce problème, les données peuvent être normalisées (par rapport à la variance ou à une plage fixée) ou en utilisant la distance de Mahalanobis dans le cas où les densités suivent une loi normale. Cette dernière est calculée à l'aide de la formule suivante :

$$d_M\left(X_i, X_j\right) = \left(X_i - X_j\right)\Sigma^{-1}\left(X_i - X_j\right)^T \text{ avec } \Sigma \text{ matrice des covariances.}$$

D'autres mesures de dissimilarité, telle que la mesure du cosinus, utilisées pour le clustering sont proposées dans la littérature. Celles-ci n'étant pas utilisées dans notre approche, ces mesures ne seront pas détaillées dans ce document. Le lecteur pourra se référer à [Jai&Dub 1998, Jai&al 1999] pour plus d'informations à ce sujet.

2.1.3. Méthodes de clustering

Les algorithmes de clustering peuvent être organisés en deux catégories : les algorithmes hiérarchiques et les algorithmes par partitionnement.

Les algorithmes de clustering hiérarchiques déterminent itérativement les clusters. Soit par agglomérations successives de points et/ou de clusters on parlera alors d'algorithmes ascendants, soit par dichotomies successives de l'ensemble des objets on parlera alors d'algorithmes descendants.

Les algorithmes par partitionnement conduisent directement à des partitions des objets initiaux, en affectant les éléments à des centres provisoires de classes, puis en recentrant ces classes, et en affectant de façon itérative ces éléments.

Nous présentons dans cette section la méthode de clustering k-means utilisée dans le cadre du système *IC-DOC*. Cette dernière est largement utilisée dans le domaine de la fouille de données et de textes. Elles est rapide, robuste et donnent généralement de bons résultats pour des espaces de grandes dimensions.

La méthode des k-means

La méthode des k-means a été proposée par B. MacQueen en 1967 [Mac 1967]. La simplicité et la robustesse de l'algorithme de classification font de k-means une des méthodes de clustering la plus utilisée et la plus performante.

Le principe de base consiste à supposer que l'espace X de n points de dimension d peut être groupé en q clusters (q<n). Les clusters sont décrits par leurs centres :

$$V_k = \left[v_k^1, v_k^2, ..., v_k^j, ..., v_k^p\right],\ 1 \le k \le q \text{ dans le même espace que } X$$

$$X = \begin{bmatrix} x_1^1 & \dots & x_1^k & \dots & x_1^d \\ \dots & \dots & \dots & \dots & \dots \\ x_i^1 & \dots & x_i^k & \dots & x_i^d \\ \dots & \dots & \dots & \dots & \dots \\ x_n^1 & \dots & x_n^k & \dots & x_n^d \end{bmatrix}$$

Notons $d(X_i, V_k)$ la distance entre le point X_i et le centre V_k. Le point X_i est affecté au cluster dont le centre est le plus proche (au sens de *d*). Nous notons m_k la moyenne des vecteurs dans le cluster *k*. L'algorithme général des k-means est le suivant.

Algorithme des k-means

Input: L'ensemble des vecteurs *X* ; la matrice de distances *D* ;

Le nombre de clusters souhaité *q* ;

Output: Ensemble des clusters : $C = \{c_1, \cdots, c_k, \cdots, c_q\}$;

Begin

// Initialiser la position des centres : $V_k = \left[v_k^1, v_k^2, ..., v_k^j, ..., v_k^p \right]$, $1 \le k \le q$;
Calculer les m_k ;

While il existe des changements sur les m_k **do**

Chaque point X_i est affecté au cluster le plus proche ;

Calculer les nouveaux m_k ;

End

End

Algorithme 4. Clustering par la méthode des k-means

La Figure 15 illustre le déroulement de l'algorithme sur un espace en deux dimensions. Le nombre de clusters fixé est *q*=4. Les positions initiales des centres sont

toutes [0,0] ($V_k = [0,0]$, $1 \le k \le 4$). Les cercles représentent les positions successives des centres des quatre clusters.

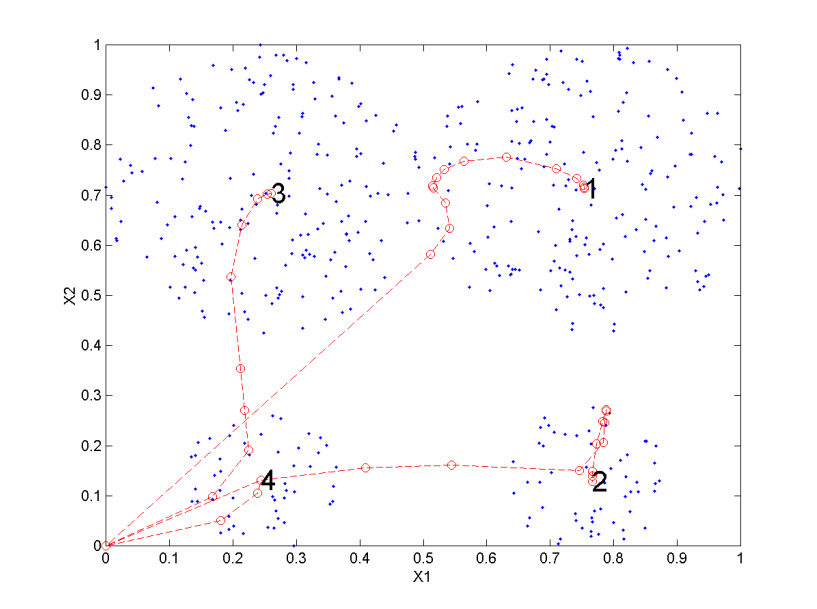


Figure 15. Exemple d'application du k-means

La position initiale des centres influe très fortement sur le résultat final du clustering. Pour palier ce problème, plusieurs positions initiales doivent être essayées. Seuls les clusters les plus stables sont alors conservés.

Approche floue du clustering

Les sous-ensembles flous permettent une représentation simple des incertitudes et imprécisions liées aux informations et aux connaissances. Leur principal avantage est d'introduire le concept d'appartenance graduelle à un ensemble alors qu'en logique ensembliste classique cette appartenance est binaire (i.e. un élément appartient ou n'appartient pas à un ensemble). Plusieurs méthodes de clustering flou ont été proposées dans la littérature. Nous citerons en exemple la méthode du « fuzzy c-means » proposée par J.C. Bezedek en 1981 [Bez 1981] et la méthode du « subtractive clustering » introduite par S. L. Chiu en 1994 [Chi 1994]. Ces deux méthodes ont été largement utilisées et font l'objet de nombreuses publications.

2.1.4. Mesures de validité

Une des difficultés majeure dans la mise en œuvre d'une méthode de classification non supervisée réside dans la validation des résultats obtenus. En effet, la classe des individus n'étant pas connue, seules des méthodes de validation par évaluation de l'homogénéité dans les groupes et d'hétérogénéité entre groupes peuvent être envisagées. Il existe dans la littérature un nombre important de définitions d'indices de validité d'un clustering [Hal&al 2002]. Nous présentons dans cette partie l'indice de Dunn [Dun 1974] qui est parmi les plus souvent utilisés et que nous retiendrons ensuite dans notre application.

L'indice de Dunn a pour objectif d'identifier des clusters compacts et bien séparés. L'expression de l'indice de Dunn est la suivante :

$$D_q = \min_{i=1,\cdots,q} \left\{ \min_{j=1,\cdots,q} \left\{ \frac{d\left(C_i, C_j\right)}{\max_{k=1,\cdots,q} diam\left(C_k\right)} \right\} \right\}$$

Avec :

- q le nombre de clusters ;
- $d\left(C_i, C_j\right)$ la fonction de dissimilarité entre deux clusters C_i et C_j, définie par $d\left(C_i, C_j\right) = \min_{x \in C_i, y \in C_j} d\left(x, y\right)$ i.e. la plus petite dissimilarité entre tous les points du cluster C_i et tous les points du cluster C_j ;
- $diam\left(C\right)$ le diamètre du cluster C qui peut être considéré comme une mesure de dispersion du cluster. $diam\left(C\right) = \max_{x,y \in C} d\left(x, y\right)$;

Si le jeu de données contient des clusters compacts et bien séparés, les distances entre les clusters sont supposées être grandes et les diamètres des clusters sont supposés être petits. Une grande valeur de l'indice de Dunn indiquera donc la présence de clusters compacts et bien séparés.

2.2. Clustering dans IC-DOC

Nous avons vu précédemment que le module de fouille de données utilisait les données issues du module de Modélisation. Cependant, pour pouvoir appliquer un algorithme de clustering, il est nécessaire de définir des mesures de dissimilarités (ou similarités) entre les différents termes à regrouper. A cet effet, la première étape du module de fouille du système *IC-DOC* concerne le calcul des mesures de dissimilarités (ou similarités) entre les termes de l'ensemble *RTC* à partir des deux matrices $MatR_1$ *et* $MatR_2$. La Figure 16 illustre les étapes de ce module de fouille. Dans notre contexte, la matrice des mesures de dissimilarités, notée *KDMAT*, est calculée comme suit.

Soit $T_i \in RTR$, $T_j \in RTC$ *et* $m = |RTC|$; la dissimilarité textuelle entre T_i *et* T_j est donnée par :

$$KDMAT(i,j) = \alpha * Dist_1(i,j) + (1-\alpha) * Dist_2(i,j)$$

où $Dist_1(i,j)$ et $Dist_2(i,j)$ sont des distances euclidiennes calculées à partir des matrices $MatR_1$ et $MatR_2$ comme suit :

$$Dist_1(i,j) = \sqrt{\sum_{k=1}^{m} [MatR_1(i,k) - MatR_1(j,k)]^2}$$

$$Dist_2(i,j) = \sqrt{\sum_{k=1}^{m} [MatR_2(i,k) - MatR_2(j,k)]^2}$$

Le paramètre α est donné par les expérimentations, ce dernier permet de pondérer entre le critère de « cooccurrences contextuelles » et le critère de « partage de contextes » [Mok&al 2006, Mok&al 2005a, Mok&al 2005b]. Si par exemple ce dernier est égal à 0.2, ceci signifie que le critère de cooccurrences contextuelles contribue à hauteur de 20 % à la pertinence des résultats tandis que le critère de partage de contextes contribue à hauteur de 80%.

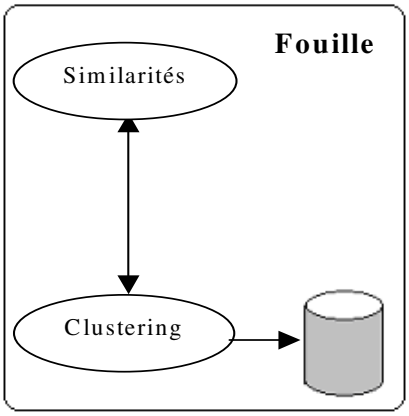


Figure 16 . Module de fouille

La deuxième phase du module de fouille consiste à appliquer un algorithme de clustering. Ce dernier n'est qu'un outil « utilitaire » dans le cadre de notre application. Notre choix s'est porté sur un algorithme classique : « le k-means », simple, efficace, robuste et très utilisé. Soit *C* une collection de documents textuels composée de *N* thématiques, nous appliquons le « k-means » de façon à obtenir *N* clusters liés à chacune des thématiques de la collection *C.* Le paramètre *k* de l'algorithme 4 décrit dans la section 2.1.3 est donc fixé à *N*. Nous présentons dans la section suivante quelques expérimentations menées sur différentes collections de documents.

2.3. Expérimentations

De manière à valider notre modèle sur des collections de documents, différentes expérimentations ont été réalisées dans l'objectif de montrer la pertinence et la capacité du modèle proposé pour l'identification de clusters de thématiques, i.e. une caractérisation thématique indépendamment des poids donnés aux thématiques dans les collections de documents.

Données

Dans le cadre de nos expérimentations, nous avons souhaité utiliser notre approche sur différents types de données. Tout d'abord nous expérimentons notre modèle sur des collections de documents composées de trois thématiques différentes: Économie, Informatique et Cinéma. Les compositions des différentes collections de documents sont illustrées dans la Figure 17. Les secondes expérimentations concernent des thématiques proches liées au domaine du cinéma.

Enfin, nous avons utilisé des données issues d'un projet de transfert de technologie et pour lesquelles nous avions à la fois des thématiques très proches et des thématiques différentes. L'objectif de ces expérimentations était de valider notre approche mais également d'en tester les limites.

Documents analysés par étape	**Economie** ***Nb_Doc***	**Informatique** ***Nb_Doc***	**Cinema** ***Nb_Doc***
C1	10	10	10
C2	40	40	40
C3	100	100	100
C4	10	40	100
C5	40	100	10
C6	100	10	40
C7	40	10	100

Figure 17. Compositions des collections de documents

Méthode

Après l'extraction des différents termes représentatifs *RTC* à partir de chacune des collections de documents, les mesures de dissimilarités sont calculées comme décrit dans la section 2.2 et nous appliquons l'algorithme de clustering suivant la méthode décrite dans la section précédente. Les résultats de pertinence du modèle sont obtenus par les mesures de *Précision* et de *Rappel* (C.f. la section 1.4) sur les termes représentatifs extraits pour chacune des thématiques dans chaque collection de documents. Soit *S* l'ensemble des *RTC* d'une thématique extraits par le système dans une collection de documents ; soit *V* l'ensemble des *RTC* de la thématique dans la collection de documents, dans le cadre de nos expérimentations, la précision et le rappel sont calculés comme suit :

$$\text{Précision} = |S \cap V| \ / \ |S|$$

$$\text{Rappel} = |S \cap V| \ / \ |V|$$

La précision détermine la quantité d'informations correctement extraites pour chacune des thématiques ; le rappel détermine la quantité d'informations

correctement extraites par rapport à ce qui devrait être extrait réellement pour chacune des thématiques.

Résultats

Les résultats obtenus lors des premières expérimentations sont illustrés Figure 18. L'objectif de l'expérimentation sur la collection C_1 (10 documents pour chacune des thématiques) est de montrer que les résultats pour une thématique ne sont pas significatifs dans le cas d'une quantité très réduite de documents, en raison de données pauvres sur les thématiques dans la collection de documents, ce qui se traduit par des chutes de précisions ou de rappels.

Comme illustré Figure 18 la précision ou le rappel ne peuvent chuter pour une thématique que dans le cas de thématiques pauvres dans une collection comme dans la collection *C6* et *C7* pour informatique, *C5* pour cinéma ou *C4* pour économie. Les chutes de précisions sont illustrées Figure 19. Dans tous les autres cas la précision dépasse les 75% et le rappel dépasse les 50% pour chacune des thématiques.

Résultats	*Nombre des TR*	***Economie***		***Informatique***		***Cinéma***	
		Précision	***Rappel***	***Précision***	***Rappel***	***Précision***	***Rappel***
C1	533	0.996	0.852	1.000	0.223	0.490	0.387
C2	1526	0.915	0.671	0.985	0.683	0.821	0.533
C3	2383	0.943	0.675	0.997	0.616	0.902	0.557
C4	1631	0.660	0.559	0.994	0.664	0.958	0.565
C5	1510	0.868	0.649	0.996	0.611	0.316	0.348
C6	1391	0.992	0.905	0.920	0.080	0.751	0.513
C7	1394	0.982	0.721	0.575	0.381	0.982	0.622

Figure 18. Résultats des expérimentations

Nous avons fait varier les poids des thématiques dans les collections de documents pour montrer la pertinence des résultats indépendamment des poids des thématiques dans les collections, les résultats illustrés sur la Figure 18 sont aussi bien pertinents dans le cas de thématiques équilibrées (collection *C2* ou *C3*) ou dans les autres cas (*C4*, *C5*, *C6* ou *C7*).

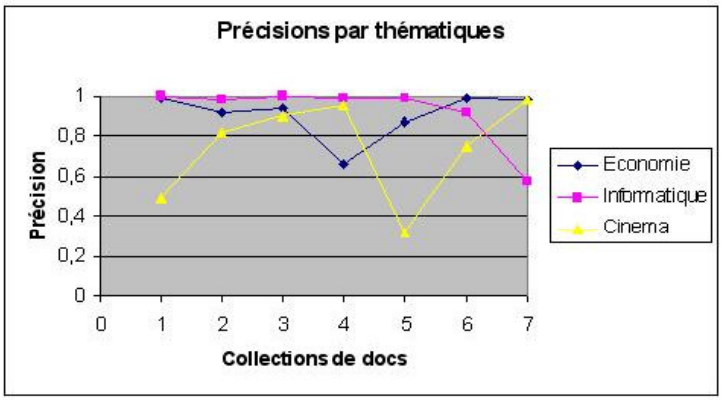


Figure 19. Précisions par thématiques

Nous avons mené également une expérimentation sur une collection de documents textuels composée de thématiques proches autour de la thématique générale du « cinéma » : thématique acteur, thématique réalisation, thématique scénario [Pla&al 2005]. La précision obtenue pour les trois thématiques dépasse les 60 % (en moyenne 63 %) et le rappel dépasse les 50 % (en moyenne 51.5 %). Nous signalons que nous avons utilisé l'indice de Dunn, présenté dans la section 2.1.4 comme mesure de validité du clustering. Rappelons que l'indice de Dunn est basé sur l'identification de clusters compacts et bien séparés. Nous signalons que l'indice de Dunn moyen correspondant aux résultats, présentés dans cette section, est équivalent à 2.33. L'objectif de cet indice est de maximiser les distances inter-cluster et de minimiser les distances intra-cluster.

Nous présentons enfin une expérimentation menée avec des données issues d'articles de presse (un extrait d'article, au format XML, utilisé lors des expérimentations est proposé en annexe B). L'objectif de cette expérimentation est d'étudier l'application de notre modèle lorsque les thématiques se chevauchent fortement. La collection est composée de cinq thématiques : « Politique, Économie des entreprises, Transport, Bourse / Marchés, Finance ». La composition de la collection est illustrée Figure 20.

Thématiques	Nombre de documents
Politique	282
Économie des entreprises	131
Transport	122
Bourse / Marchés	201
Finance	166

Figure 20. Contenu de la collection de documents

Après la phase de nettoyage de la collection des balises XML, nous appliquons le même processus de prétraitements suivant le modèle proposé dans cette thèse. Nous appliquons également le clustering et nous analysons les résultats de la même manière que les expérimentations exposées au début de cette section. La Figure 21 illustre les précisions et rappels obtenus pour chacune des thématiques de la collection. Ces résultats sont obtenus pour un indice de Dunn de 1.87.

Thématiques	Précision	Rappel
Politique	0.698	0.560
Economie des entreprises	0.505	0.478
Transport	0.679	0.521
Bourse / Marchés	0.475	0.310
Finance	0.631	0.437

Figure 21. Résultats par thématiques

Les résultats illustrés Figure 21 montrent que la précision est proche de 70% pour les thématiques « Politique » et « Transport ». La précision dépasse les 50 % pour les thématiques « Economie des entreprises » et « Finance ». Elle est proche de 50% pour la thématique « Bourse / Marchés ». Le rappel dépasse les 50% pour les thématiques « Politique » et « Transport » et il est proche de 50 % pour la thématique « Economie des Entreprises ». Le rappel diminue un peu pour les thématiques « Finance » et « Bourse / Marchés ». En effet, ces deux dernières thématiques se chevauchent considérablement et partagent un vocabulaire très proche. Ce qui se traduit, dans le cadre de notre expérimentation, par la diminution des taux de rappel pour ces deux thématiques et la diminution du taux de précision pour la thématique « Bourse / Marchés ».

Les expérimentations présentées dans cette section montrent les capacités de notre modèle pour la représentation de différentes collections de documents. Notre modèle peut être pénalisé lorsque les thématiques sont trop proches. De manière à mieux comprendre les raisons, nous avons examiné plus en avant les collections de documents et lorsque les thématiques sont très proches, nous avons noté qu'il y avait plus de 60 % de vocabulaire commun (Ex. 61.08% du vocabulaire de la thématique « Bourse/Marches » est commun avec celui de la thématique « Finance »). Dans ce cas, nous pensons qu'il est possible d'améliorer ces résultats et d'obtenir des résultats encore plus pertinents en utilisant des algorithmes de clustering flou. Ces derniers permettent d'associer des termes ou associations de termes à plusieurs clusters et non pas forcément à un seul comme nous le faisons actuellement. Nous

revenons sur cette discussion dans le chapitre V où nous étudions les possibilités d'extension du système *IC-DOC*.

3. Cartographie et visualisation de connaissances textuelles

Dans cette section, nous montrons comment le système *IC-DOC* peut être adapté pour faire de la cartographie et de la visualisation de connaissances textuelles. Après un rapide survol des objectifs des outils de cartographie, nous précisons comment nous avons réalisé ce module dans *IC-DOC*. Nous présentons ensuite une application menée avec notre système.

3.1. Objectifs des outils de cartographie et de visualisation

Les outils de cartographie et de visualisation de connaissances textuelles contribuent à l'efficacité et la pertinence des processus mis en œuvre en offrant aux utilisateurs des représentations intelligibles des contenus et facilitant l'interaction [Clo&al 2006, Joa&al 2006, Mot&al 2003]. Il existe différents systèmes de catégorisation, voire de cartographie de documents et d'informations tel que *Kartoo* [Chu&al 2003] ou *Mapstan* [Spi 2002]. Ces derniers retrouvent des liens entre les différents documents ou sites Web et représentent ces liens sous formes de cartes de navigation [Wis&Van 2004]. Cependant les informations du contenu de la collection de documents ou de chacun des documents sont généralement peu représentées. En se basant sur le modèle de représentation présenté dans le chapitre III, nous illustrons, dans cette section, l'intérêt de ce dernier dans le cadre d'une cartographie visuelle des informations enfouies dans les contenus textuels.

3.2. Cartographie visuelle dans IC-DOC

Après une phase d'interprétation de données fournies par le module de fouille, la deuxième étape du module 4 du système *IC-DOC* concerne la visualisation [Mok&al 2006, Mok&al 2004a]. La Figure 22 illustre les étapes de ce module. Pour faciliter la représentation des résultats obtenus et pour une cartographie visuelle, nous avons développé un outil de visualisation des connaissances textuelles que nous avons intégré à ce module. La visualisation des informations est faite via des cartes dynamiques d'informations, appelées aussi *Knowledge Dynamic Maps (KDMs)*. Une *KDM* est un graphe $G = <X, U>$ où X est un ensemble de N sommets modélisant N termes représentatifs et U un ensemble d'arêtes représentant les relations textuelles. Les sommets du graphe sont des hyperliens (liens dynamiques) ayant deux

fonctionnalités. La première permet d'organiser d'une manière automatique le graphe autour d'un thème central. La seconde permet d'atteindre une nouvelle *KDM*. La dimension d'une *KDM* correspond au nombre de ses termes représentatifs [Mok&al 2006, Mok 2004a]. La Figure 23 illustre un exemple de carte dynamique d'informations. Nous présentons dans la section suivante une mise en œuvre de notre outil.

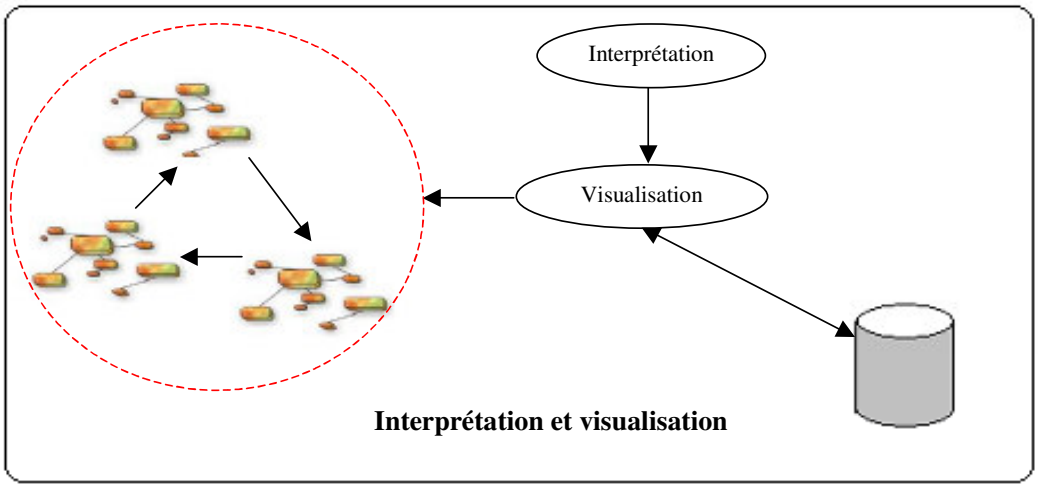


Figure 22. Module d'interprétation et de visualisation

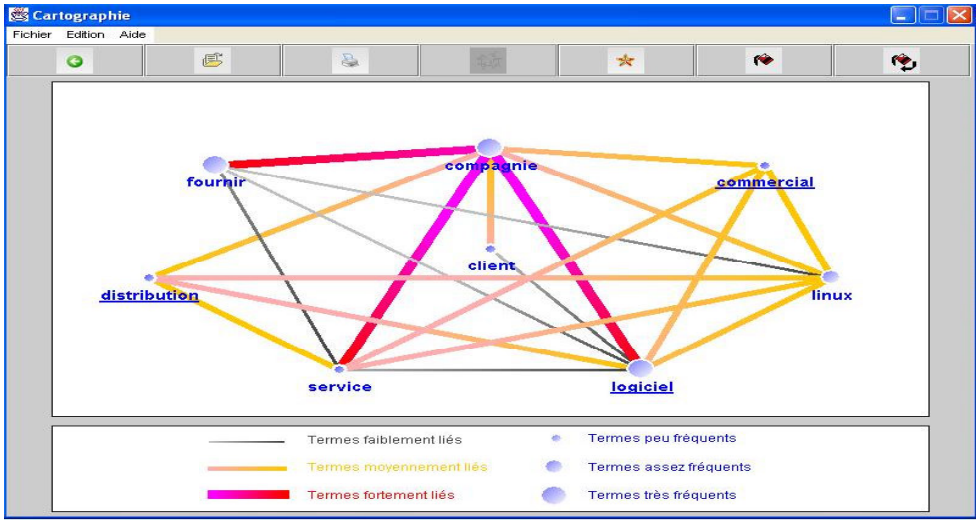


Figure 23. Exemple de carte dynamique d'informations

3.3. Application

L'application présentée dans cette section est composée de deux parties. L'objectif de la première partie est de montrer l'apport de notre méthode d'extraction de termes en prenant en considération les cooccurrences contextuelles. Dans la seconde partie nous illustrons l'intérêt de notre modèle dans le cadre d'une visualisation interactive d'informations sur le contenu.

A cet effet, nous avons appliqué dans un premier temps notre approche, via l'outil *CO2MAT* du module 2 du système *IC-DOC*, sur une collection d'environ 1000 documents de presse [Mok&al 2004c, Mok&Are 2003]. La collection de documents porte sur la thématique « politique internationale ». Ces documents sont issus de journaux internationaux avant le déclenchement de la deuxième guerre en Irak.

En suivant le modèle de représentation présenté dans le chapitre III, la première étape du choix des termes représentatifs consiste à générer l'ensemble des termes en ne prenant en considération que les fréquences *FTC* des termes dans la collection et le nombre de documents *FTD* contenant chaque terme. L'ensemble RTC_1 des termes générés par la première étape et non détectés par la seconde ainsi que les termes les plus fréquents de cette étape est le suivant :

RTC_1 = {Guerre, Irak, Américain, Bush, Faire, Aller, Saddam, Président, Irakien, France, ONU}.

La deuxième étape consiste à extraire des termes représentatifs en prenant en considération les fréquences de cooccurrences *FCC*. Les cooccurrences contextuelles les plus fréquentes de l'étape 2 ainsi que celles qui ont participées à l'extraction de nouveaux termes non détectés à l'étape 1 sont décrites dans la Figure 24.

1	CO : Irak – Guerre	6	CO : Bush – Faire	11	CO : Président – Bush
2	CO : Bush – Guerre	7	CO : Faire – Guerre	12	CO : Président – Saddam
3	CO : Irak – Etats Unis	8	CO : Aller – Guerre	13	CO : Irak – Président
4	CO : Irak – Américain	9	CO : Bush – Irak	14	CO : Irak – Irakien
5	CO : Bush – Aller	10	CO : Faire – Irak	**15**	CO : Irak – Inspection

Figure 24. Les cooccurrences contextuelles pertinentes

L'ensemble des termes représentatifs correspondant à la Figure 24 est le suivant :

RTC_2 = {Etats-Unis, Irak, Irakien, Guerre, Américain, Bush, Aller, Faire, Président, Saddam, Inspection}.

L'ensemble *RTC* des termes représentatifs résultant de l'union de RTC_1 et RTC_2 est donc :

RTC= {Guerre, Irak, Américain, Bush, Faire, Aller, Saddam, Président, Irakien, *France, ONU, Etats-Unis, Inspection*}.

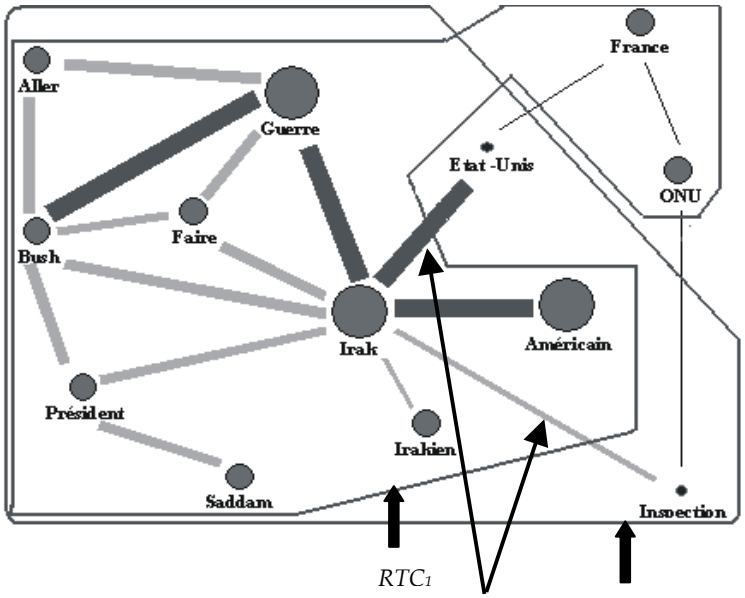
$RTC_1 — RTC_2$= {France, ONU}, les termes appartenant à RTC_1 et non à RTC_2.

$RTC_2 — RTC_1$= {Etats-Unis, Inspection}, les termes appartenant à RTC_2 et non à RTC_1.

Nous remarquons que la simple utilisation des fréquences des termes nous prive d'informations sur les Etats-Unis et les Inspections (ensemble $RTC_2 — RTC_1$). La prise en considération des cooccurrences seule mène au même problème, ce qui se traduit par la suppression d'informations sur l'ONU et la France (ensemble $RTC_1 — RTC_2$). Il est clair que les deux étapes de sélection des termes représentatifs sont complémentaires. La Figure 25 illustre la carte d'informations associée à ces termes représentatifs du contenu.

Dans notre modèle, malgré que le terme Etats-Unis ait une fréquence faible dans la collection de documents, ce dernier a été sélectionné comme terme représentatif. Cela grâce à la cooccurrence contextuelle importante entre Irak et Etats-Unis *{CO : Irak—Etats-Unis}.* De même pour le terme Inspection qui a été sélectionné par rapport à la cooccurrence contextuelle entre Irak et Inspection *{CO : Irak—Inspection}.* Ces nouveaux termes permettent de représenter de nouvelles cooccurrences utiles pour la représentation du contenu de la collection (ces cooccurrences ne figurent pas dans la Figure 24), comme les cooccurrences contextuelles entre la « France et les Etats-Unis » ou bien « Inspection et ONU ». Ces nouvelles cooccurrences sont faibles dans la collection mais apportent de nouvelles informations sur le contenu de la collection.

Une première consultation de cette carte donne un aperçu général sur le contenu de la collection. On remarque que la presse est concentrée sur la guerre en Irak. On remarque aussi que la presse écrite parle souvent du président des Etats Unis. Il existe des informations sur les initiatives Etats-Unis / France et il y a des informations sur l'Organisation des Nations Unies (ONU), etc.



Cooccurrences contextuelles participant à la détection des termes représentatifs Etats-Unis et Inspection.

Figure 25. Choix et représentation des termes de l'ensemble RTC

La deuxième partie de cette section concerne l'aspect visuel des connaissances. Cette visualisation peut être considérée comme une étape complémentaire à l'identification de clusters de thématiques présentée précédemment. En effet, après l'identification des clusters liés aux thématiques, nous lançons le processus de clustering pour chaque thématique. Soit *NB* le nombre de thématiques de la collection de documents et *DK* la dimension d'une *KDM* définie a priori de façon à ne pas surcharger l'utilisateur. Dans une première étape, nous appliquons le clustering comme détaillé dans la section 2. A l'issue de cette étape nous obtenons *NB* clusters où chaque cluster correspond aux termes représentatifs d'une thématique (sous ensemble de *RTC*).

La seconde étape du processus de clustering consiste en l'application du clustering aux termes de chaque thématique. Pour une thématique donnée, nous appliquons donc l'algorithme k-means(*DK*) à l'ensemble des termes représentatifs de cette thématique. Nous obtenons *DK* clusters. Ces clusters sont représentés sous forme d'une *KDM*, que nous appelons *KDM* principale, cette dernière est définie de la manière suivante.

Chaque sommet de la *KDM* principale représente un des *DK* clusters obtenus. Un cluster peut être représenté par le terme ayant le poids le plus fort. Les liens entre les sommets de cette *KDM* principale peuvent représenter les distances entre termes, i.e.

entre sommets. Nous rappelons que ces distances sont stockées dans la matrice *KDMAT*.

Le processus de clustering de cette seconde étape est itératif. Pour chacun des *DK* clusters dont la taille dépasse *DK*, nous relançons le processus de clustering. L'objectif du processus de clustering itératif est de permettre d'éclater un cluster en plusieurs autres clusters, i.e. la *KDM* principale en plusieurs autres *KDMs*. Ce processus permet une visualisation interactive et une navigation par le contenu. En effet, nous avons vu dans la section précédente que les sommets d'une *KDM* peuvent renvoyer à d'autres *KDMs*. Dans notre application, un sommet d'une *KDM* renvoie à une autre *KDM* si la taille du cluster représenté par ce sommet dépasse *DK*.

De façon à illustrer l'intérêt de cette visualisation des connaissances textuelles, nous avons étendu le système *IC-DOC* en lui intégrant le sous système de *Visualisation de Connaissances Textuelles (VICOTEXT*) [Mok 2004b]. Nous avons appliqué la seconde étape du processus de clustering détaillé dans cette section sur une collection portant sur la thématique « voyages et aventures ». Nous illustrons dans la suite de cette section ce processus via des copies écrans, de façon à montrer au mieux l'intérêt de notre modèle dans une application de visualisation interactive de connaissances textuelles.

La Figure 26 illustre la carte d'informations principale (*KDM* principale) et l'environnement général du sous système *VICOTEXT*. Les figures 27 et 28 illustrent de manière générale les interactions de l'utilisateur. En effet, la Figure 27 illustre la carte d'informations autour du thème « santé » (nouvelle *KDM* autour de « santé ») et la Figure 28 illustre l'auto organisation de la même *KDM* autour du thème « jeu » suite aux différentes interactions de l'utilisateur.

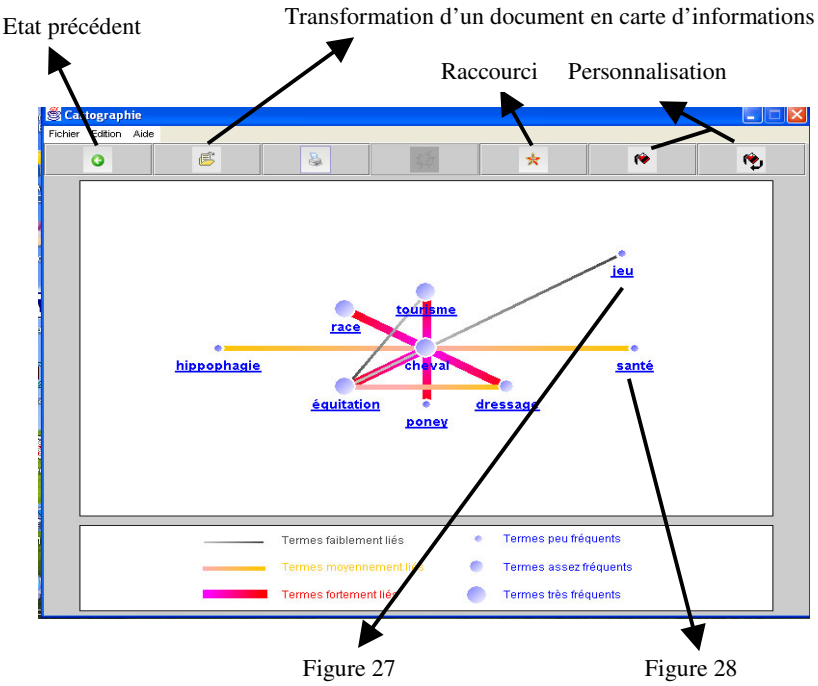


Figure 26. Environnement et carte d'informations du contenu global

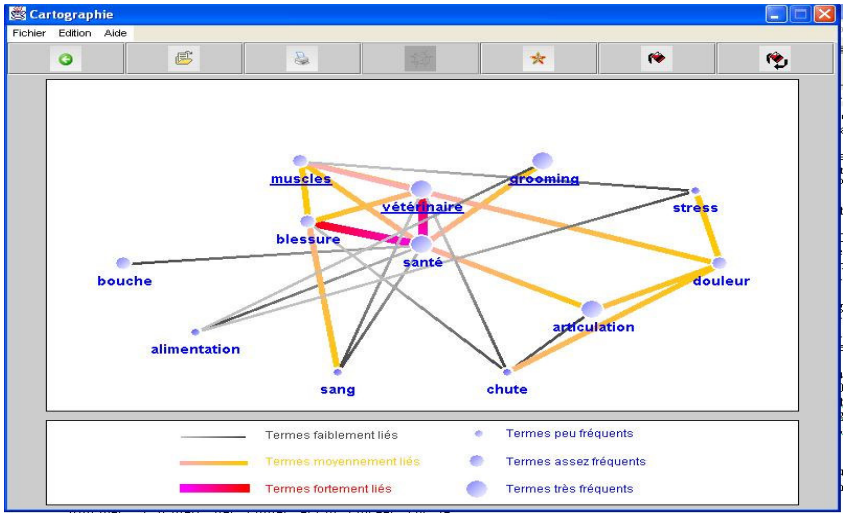


Figure 27. Carte d'informations autour du thème *santé*

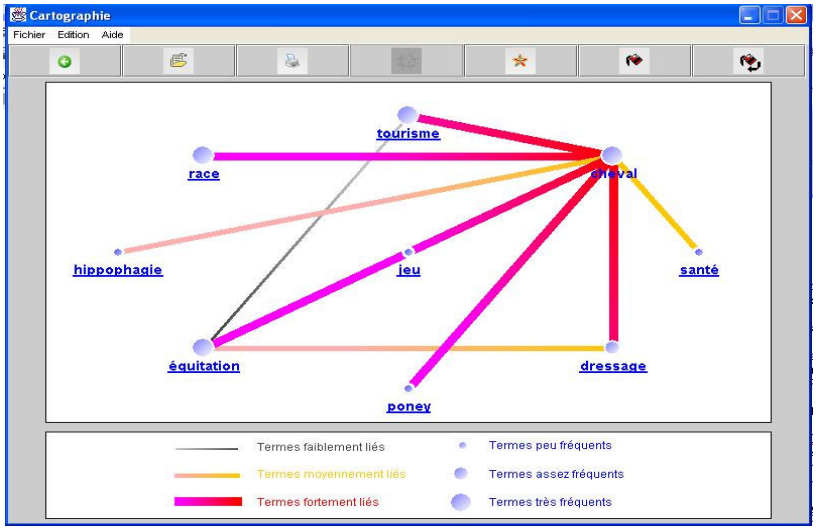


Figure 28 . Auto-organisation de la carte autour du thème *jeu*

4. Conclusion

Nous avons présenté dans ce chapitre le système *IC-DOC* ainsi que le fonctionnement de ses différents modules. L'implémentation de ces derniers suit la chaîne des traitements de la fouille de données textuelles. Ce système nous a permis de mettre en œuvre notre modèle de représentation de collections de documents textuels.

Nous avons expérimenté notre modèle au travers de deux applications complémentaires.

La première application concerne l'identification de clusters de thématiques à partir de collections de textes. La série d'expérimentations réalisée sur diverses collections de documents a montré la pertinence de notre modèle pour aider à une caractérisation thématique du contenu. Les résultats obtenus confirment également que notre approche est indépendante des poids donnés aux thématiques dans les collections de documents.

La seconde application avait pour but de montrer l'intérêt de notre modèle dans le cadre d'une cartographie visuelle de connaissances textuelles. Au travers de cette application, nous avons montré dans un premier temps l'utilité de notre modèle pour synthétiser les contenus textuels ou pour avoir un aperçu général sur le contenu global. Dans un second temps, nous avons exposé une application de visualisation de connaissances textuelles. Ce type d'application peut être considéré comme une étape complémentaire au processus d'identification de clusters de thématiques. La

visualisation des connaissances est réalisée via des cartes dynamiques d'informations. Nous avons étendu dans cette deuxième application le système *IC-DOC* de façon à illustrer au mieux l'utilité de ce dernier. Nous revenons plus en détails dans le dernier chapitre sur l'utilité du système *IC-DOC* et les perspectives de notre modèle de représentation.

Chapitre V – Conclusions et Perspectives

Dans ce chapitre, nous concluons ce mémoire en revenons sur les différentes contributions et en présentant les extensions possibles de ce travail. Le chapitre est organisé de la manière suivante. Dans la section 1, nous rappelons les principales contributions et revenons sur les discussions réalisées au cours du mémoire. Nous présentons, au cours de la section 2, les perspectives associées à ce travail.

1. Contributions

Au cours de ce mémoire nous avons abordé la problématique de représentation de collections de documents dans le cadre d'un processus d'extraction de connaissances à partir de données textuelles.

Après avoir étudié les principales approches existantes de représentation, nous nous sommes intéressés plus particulièrement à celles liées à la cooccurrence de termes. Cependant, nous avons vu dans le chapitre II et en discussion du chapitre III que ces approches souffrent de certaines lacunes en considérant que les cooccurrences ne sont obtenues qu'après avoir déjà effectué des choix dans les termes à retenir. Au cours du chapitre III, nous avons étudié la possibilité de relâcher la contrainte de contexte, limitée dans notre cas à la phrase, pour pouvoir considérer des cooccurrences intervenant dans un « intervalle plus important ». Cependant, cette approche est difficile à mettre en œuvre car elle nécessiterait de trop nombreuses analyses des documents pour aider à fixer la fenêtre optimale. En outre, cette fenêtre est également dépendante des documents. Notre choix s'est donc porté sur la définition d'un nouveau critère appelé « partage de contextes ». Via ce critère nous étendons les cooccurrences et sommes alors à même de détecter des liens entre termes qui n'apparaîtraient pas autrement.

Pour prendre en compte ces différents critères nous avons défini dans le chapitre III, différents algorithmes qui, à partir des documents sources, permettent d'obtenir deux matrices différentes :

- La matrice $MatR_1$ qui reflète les cooccurrences d'un terme donné par rapport à ses occurrences, i.e. le critère de « cooccurrences contextuelles ».

- La matrice $MatR_2$ qui représente la notion de « partage de contextes ».

Au cours du chapitre IV, nous avons proposé un nouvel environnement d'extraction de connaissances à partir de documents textuels appelé *IC-DOC*. L'originalité de cet environnement est de tirer profit de notre modèle de représentation. Enfin, nous avons considéré deux applications. Dans la première application, nous nous sommes

intéressés à une caractérisation thématique de collections de documents indépendamment des poids donnés aux thématiques. L'idée sous jacente est d'être capable de déterminer les thématiques communes entre différents documents. Pour cela, nous nous sommes intéressés aux approches de classification non supervisée et avons retenu dans nos expérimentations la méthode k-means, simple et robuste, qui nous a permis de mettre en œuvre notre modèle dans une application de caractérisation thématique. Les résultats de nos expérimentations ont montré qu'à l'aide de nos deux matrices couplées à un algorithme de clustering nous pouvions obtenir des clusters cohérents représentatifs des thématiques enfouies dans les collections. Il est important de préciser qu'en ne considérant que les cooccurrences traditionnelles, les résultats de cette caractérisation chutent d'une manière considérable. En effet, le paramètre α réglé lors des expérimentations montre que le critère de « partage de contextes » participe à hauteur de 70 % à la pertinence des résultats tandis que le taux de participation du critère de « cooccurrences contextuelles » ne dépasse pas les 30%.

Le second type d'application que nous avons souhaité aborder concerne la synthèse des contenus de collections de documents textuels. Cette synthèse est réalisée via une cartographie d'informations en utilisant notre modèle de représentation. La visualisation des connaissances sur le contenu des collections des documents est rendue possible via des cartes dynamiques d'informations. Les expérimentations réalisées ont montré l'efficacité du choix des termes représentatifs du contenu en prenant en considération les cooccurrences contextuelles. La pertinence du modèle de représentation nous a permis d'avoir des aperçus significatifs et représentatifs du contenu.

Dans la suite de ce chapitre, nous présentons quelques perspectives de notre modèle de représentation et du système *IC-DOC* à d'autres applications de fouille de textes.

2. Perspectives

2.1. D'autres approches de clustering

De manière à montrer que notre approche pouvait être utile pour caractériser thématiquement des collections de documents, nous avons mené différentes expérimentations décrites dans le chapitre IV avec l'algorithme de clustering k-means. Même si les résultats des expérimentations ont montré que notre approche est pertinente, ces résultats peuvent être améliorés dans le cas de thématiques très proches. Par exemple, considérons les documents suivants (extraits de documents de

la collection « cinéma » présentée dans le chapitre précédent). D_1 = « *On soulignera que la grande qualité de ce film réside dans sa force émotionnelle* ». D_2 = « *L'émotion est née de la complicité et l'ambiguïté de Cassel à l'écran* ». D_3 = « *Cassel semble doué d'une maîtrise de sa personnalité dans ce second film* ». D_4 = « *Le réalisateur a voulu que ce film soit léger afin qu'il puisse être revu plusieurs fois avec plaisir* ». Les documents D_1 et D_2 abordent la thématique « acteur » alors que les documents D_3 et D_4 concernent celle de « réalisation ». Ces thématiques étant sémantiquement très proches, il est difficile de les caractériser. Les approches de clustering traditionnelles vont permettre de regrouper des termes (ou associations de termes) dans un seul cluster alors qu'il est évident qu'ils peuvent dans ce cas appartenir à plusieurs clusters, comme le cas du terme « *film* » ou « *Cassel* » de l'exemple précédent. Au cours du chapitre IV, nous avons vu que le clustering flou pouvait répondre à ce problème. L'une des perspectives à court terme, de ce travail est de poursuivre cette approche. Nous avons déjà réalisé quelques expérimentations qui montrent que les approches floues semblent adaptées, cependant ces expérimentations doivent être poursuivies. En effet, le processus de décision dans le cadre des approches floues est complexe. Il nécessite de définir des règles fiables pouvant aider à l'interprétation des clusters.

2.2. Prise en compte de l'aspect dynamique

Dans la définition de notre modèle, nous considérons que les documents sont déjà collectés et nous partons d'ensembles de documents statiques. Cependant, dans le cas de nombreuses applications « temps réel », cette contrainte est très pénalisante car elle nécessite de relancer la phase d'extraction. Si nous examinons plus attentivement les conséquences par rapport à notre proposition, il est évident que dans un premier temps nous devons maintenir les deux matrices $MatR_1$ et $MatR_2$. La difficulté principale consiste en fait à gérer de manière dynamique les dimensions de ces matrices. En effet, lorsque des termes sont similaires dans les nouveaux documents qui interviennent, il suffit de remplacer les valeurs des matrices. Cependant lorsque de nouveaux termes interviennent, de nouvelles cooccurrences peuvent également intervenir. L'une des solutions pour maintenir ces matrices serait d'associer à chacun des termes un gestionnaire d'historique afin de conserver les occurrences de chacun des termes au cours du temps. Une approche similaire, appelée *tilted time windows*, est actuellement proposée dans la détection d'itemsets dans des flots de données [Gia&al 2003]. Cette notion a initialement été introduite dans [Che&al 2002] et est basée sur l'intuition suivante : les utilisateurs s'intéressent plus souvent aux récents changements avec une granularité fine et aux changements à long terme avec une granularité plus large. Cette hypothèse pourrait également être appliquée à notre contexte où les utilisateurs sont plus intéressés par un regroupement des nouvelles informations qu'aux anciennes.

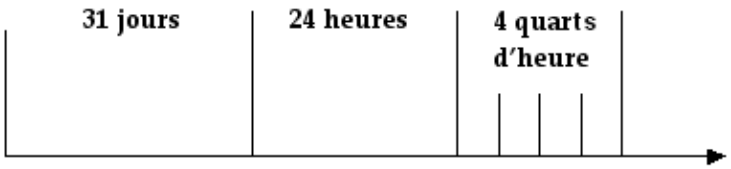


Figure 29. Un exemple de tilted time windows

Le fonctionnement de *tilted-time windows* est illustré Figure 29 : les 4 quarts d'heure plus récents, les dernières 24 heures et enfin 31 jours. A partir de ce modèle, il est possible de représenter les informations apparues pendant la dernière heure avec une précision au quart d'heure près, le dernier jour avec une précision d'une heure, etc. En associant à chaque terme une table de *tilted-time windows*, il est possible de détecter et de connaître les occurrences des termes avec les intervalles de temps associés. La mise à jour des fenêtres temporelles au sein de la table se fait grâce à une opération de décalage et de compactage sur chacun des niveaux de granularités en commençant par la granularité la plus fine. Le fait de compacter les termes implique bien entendu qu'il est indispensable de faire des approximations sur les résultats. Dans le contexte des itemsets cette approximation est basée sur un support intermédiaire.

A partir de ces historiques, nous disposerions donc des éléments constitutifs des matrices tout en assurant une meilleure gestion des dimensions. Par exemple, nous ne pourrions rajouter des dimensions que si le nombre d'occurrences d'un terme devient significatif par rapport à un seuil donné.

Toutefois, même si une approche basée sur les tilted time windows pourrait répondre en partie à notre problématique, le second problème à considérer est comment mettre à jour de manière automatique les clusters obtenus lors de l'étape précédente, i.e. les connaissances acquises. Une solution pour résoudre ce problème pourrait également venir des travaux de recherche sur les flots de données. En effet, les techniques traditionnelles d'extraction de connaissances ou de fouille de données doivent aujourd'hui faire face à une contrainte nouvelle liée au développement des nouvelles technologies : les données manipulées apparaissent sous la forme de flots de manière continue, ordonnée et potentiellement infinie (*data streams*). L'importance de ces flux ne permettant pas de les stocker, il est crucial de développer de nouvelles approches bornant l'utilisation de la mémoire et les différentes entrées-sorties sur les données. Récemment, de nouveaux travaux de recherche, appelés « *Data Stream Mining* », se sont intéressés à la définition d'algorithmes de clustering en temps réel. Nous pouvons citer par exemple : [Ber&Hul 2005], [Cao&al 2006] ou [Agg&al 2003].

Il serait intéressant d'étudier l'intégration de ce type d'algorithmes dans le système *IC-DOC*. Toutefois, cette intégration n'est pas si simple car il faut d'une part la coupler aux matrices et d'autre part assurer que les prétraitements des documents ne pénalisent pas le processus global. En outre, comme nous l'avons vu pour le cas des fenêtres temporelles, la prise en compte des données issues d'un flot nécessite d'approximer les résultats car les données ne peuvent pas être stockées. Bien entendu, même si cette contrainte est indispensable pour certains domaines d'application (e.g. classification de news en ligne), il faut reconnaître qu'il existe de très nombreux domaines où cette dernière peut être relâchée. L'une des questions auxquelles il faut maintenant répondre est la suivante : l'utilisateur peut-il se contenter d'une réponse approximative (même s'il est possible de maîtriser en partie cette approximation) alors qu'il existe une réponse complète à son problème ? En effet, notre objectif initial, comme nous l'avons vu précédemment, est d'offrir la connaissance la plus complète possible et que cette connaissance ne soit pas trop importante pour pouvoir être manipulée par des algorithmes de clustering ou pour pouvoir être visualisée. En intégrant une nouvelle approximation nous minimiserons encore le nombre de connaissances et il serait indispensable d'étudier la qualité des résultats obtenus par rapport aux besoins de l'utilisateur final.

2.3. Partage de connaissances

Le système *IC-DOC* que nous avons développé permet comme nous l'avons vu tout au long de ce mémoire, d'extraire de la connaissance à partir d'une collection de documents textuels. Une perspective de ce système est de pouvoir l'étendre pour faciliter le partage de connaissances entre différents utilisateurs.

Dans ce cadre, l'objectif est d'une part d'étendre le système à la prise en compte du comportement de l'utilisateur mais également d'intégrer celui d'autres utilisateurs. En effet, le premier point est primordial pour pouvoir faire profiter la communauté de l'expérience de l'utilisateur. Par contre, dans le second point, l'objectif est d'offrir de nouveaux outils d'aide au partage de connaissances.

Dans notre cas, l'apprentissage du comportement de l'utilisateur consiste à en établir son profil. Nous pouvons considérer que ce profil est représenté par les documents auxquels il a accédé au cours du temps. Dans [Are&al 2004], nous avons montré qu'il était possible de prendre en compte trois profils distincts : à long, court et moyen terme. Alors que la sémantique des deux derniers correspond intuitivement « aux derniers » documents qui ont été lus, le profil long terme s'intéresse aux différents documents qui ont été accédés au cours du temps et qui constituent un intérêt régulier pour l'utilisateur. Le système que nous avons proposé dans [Are&al 2004]

s'intéresse au comportement à long terme dans un contexte de recherche documentaire. Le principe est le suivant : nous classons les documents consultés et nous analysons l'évolution des classes au cours du temps. Dans le contexte *IC-DOC*, l'idée est de réutiliser ces notions de profils pour une meilleure modélisation de l'utilisateur au cours du temps. Cette approche est complémentaire à celle présentée dans la section précédente car elle considère que les documents arrivent de manière dynamique. Toutefois, dans ce contexte, nous ne souhaitons pas réduire l'espace de recherche mais plutôt suivre l'évolution des clusters au cours du temps. L'idée sous jacente est par exemple qu'un utilisateur intéressé par le langage java va initialement consulter des documents sur les généralités et qu'au cours du temps sa compétence et son expérience vont faire qu'ils ne va pas consulter le même type de documents : il sera intéressé par des documents sur le langage java mais plus particulièrement sur les aspects avancés du langage. Dans un contexte de communautés, nous disposons alors d'une connaissance significative sur le comportement d'un utilisateur. Cette connaissance peut alors être mise en œuvre de différentes manières : nous pouvons offrir aux autres membres de la communauté les documents qui semblent les plus pertinents par rapport à un besoin donné, nous pouvons regrouper entre eux les utilisateurs dont les profils sont proches, etc. Ce qui favorise le partage de connaissances et l'échange d'expériences entre les différents membres de la communauté.

Publications réalisées au cours de cette thèse

Chapitres de livres avec comité de rédaction

[Mok&al 2006] A. Mokrane, G. Dray et P. Poncelet. Fouille de collections de documents en vue d'une caractérisation thématique de connaissances textuelles. *Dans le chapitre VIII de l'ouvrage «Extraction de connaissances : Etat et Perspective »* (F. Cloppet, N. Vincent et J-M Petit éds.), Cépaduès Editions, I.S.B.N 2.85428.707.x. Paru en 2006. pp 269-278.

Revues internationales avec comité de sélection

[Mok&al 2005a] A. Mokrane, G. Dray et P. Poncelet. Caractérisation thématique de collections de documents textuels. *In the International Journal on line of Information Science and Decision Making (ISDM)*. N° 22. I.S.S.N 1265 499X. Paru en 2005.

Conférences internationales avec comité de sélection

[Mok&al 2005b] A. Mokrane, G. Dray et P. Poncelet. Automatic Representation of Textual Document Collections for Thematic Characterization of Contents. Abstract. *In Data Mining 2005*. May 2005, Skiathos, Greece.

[Mok&al 2004a] A. Mokrane, P. Poncelet et G. Dray. Visualisation automatique du contenu d'une base de documents textuels via les hyper cartes d'information. *Actes du 4ème Colloque International de Veille Stratégique, Scientifique et Technologique (V.S.S.T 2004)*. Octobre 2004, Toulouse, France. pp 249-260.

[Mok&al 2004b] A. Mokrane, G. Dray et P. Poncelet. Fouille et cartographie automatique de bases de documents textuels de communautés d'utilisateurs. *Actes de la Conférence Internationale sur les Systèmes Complexes (CISC 2004)*. Septembre 2004, Jijel, Algérie. 8p (CD-ROM).

[Mok 2004a] A. Mokrane. Automatic Extraction of Knowledge Dynamic Maps from large Textual Databases. *Proceedings Volume II of the 21st British National Conference on Databases (BNCOD21)*. July 2004, Edinburgh, UK. 2p.

[Are&al 2004a] R. Arezki, A. Mokrane, G. Dray, P. Poncelet and D. Pearson. A Personnalization Documentary System Based on the Analysis of the History of the User's Actions. *Proceedings of the 6th International Conference On Flexible Query Answering Systems* (*FQAS 2004*), *Lecture Notes in Artificial Intelligence (LNAI)*, Springer Verlag. June 2004, Lyon, France. pp 487-498.

[Mok&al 2004c] A. Mokrane, R. Arezki, G. Dray et P. Poncelet. Cartographie automatique du contenu d'un corpus de documents textuels. *Actes des 7ème Journées internationales d'Analyse statistique des Données Textuelles (JADT 2004)*, *Le poids des mots*, (G. Purnelle, C. Fairon, A. Dister éds.), Presses Universitaires de Louvain (PUL). Mars 2004, Louvain-la-Neuve, Belgique. pp 816-823.

Conférences nationales avec comité de sélection

[Mok 2004b] A. Mokrane. VICOTEXT : un outil de visualisation automatique et dynamique de connaissances textuelles. Démonstration et Poster. *Actes des 15èmes journées francophones d'Ingénierie des Connaissances (IC 2004)*, Lyon. pp 28-29.

[Are&al 2004b] R. Arezki., A. Mokrane , G. Dray , P. Poncelet and D. Pearson Modélisation dynamique et temporelle de l'utilisateur pour un filtrage personnalisé de documents textuels. 4èmes Journées d'Extraction et de Gestion de Connaissance (*EGC 2004*), RNTI. Janvier 2004, Clermont-Ferrand. Volume 2, pp 479-484.

[Mok&Are 2003] A. Mokrane et R. Arezki. Méthodologie de modélisation du contenu global d'un corpus documentaire par un graphe de liens sémantiques. *Actes des Journées Graphes, Réseaux et Modélisation (GRM 2003)*. Décembre 2003, Paris. pp 26-27.

Bibliographie

[Aas&Eik 1999] K. Aas and L. Eikvil. Text Categorization : A Survey. *Technical report.* Norwegian Computing Center. 1999.

[Agg&al 2003] C. Aggarwal, J. Han, J. Wang and P. S. Yu. A Framework for Clustering Evolving Data Streams. *Proceedings of the International Conference on Very Large Data Bases (VLDB'03).* September 2003. Berlin, Germany.

[Agr&al 1993] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Database. *Proceedings of the International Conference on Management of Data (ACM SIGMOD 93).* 1993. Washington, USA. pp 207-216.

[Agr&Sri 1995] R. Agrawal and R. Srikant. Mining Sequential Patterns. *Proceedings of the 11th International Conference on Data Engineering (ICDE 95).* 1995. Tapei, Taiwan. pp 3-14.

[Apt&al 1994] C. Apté, F.J. Damerau, and S.M. Weiss. Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems.* 1994. 12(3) : pp 233-251.

[Ber&Hul 2005] J. Beringer and E. Hullermeier. *Online Clustering of Parallel Data Streams*. Data & Knowledge Engineering, 2005.

[Ber 2003] M. W. Berry. *Survey of Text Mining : clustering, classification and retrieval.* Springer Verlag. 2003. ISBN 0-387-95563-1.

[Ber 2002] P. Berkhin. Survey of Clustering Data Mining Techniques. *Technical Report*. Accrue Software. September 2002. 56p.

[Bes&al 2001] R. Besançon, A. Rozenknop, J.C. Chappelier and M. Rajman. Intégration probabiliste de sens dans la représentation de textes. *Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN)*. 2001, Tours, France. pp 83-91.

[Bez 1981] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithm*. Plenum Press. 1981. New York.

[Bri 1992] E. Brill. A simple Rule-based Part-of-speech Tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP 1992)*. March 1992, Trento, Italy. pp 152-155.

[Buc&al 1992] C. Buckley, G. Salton and J. Allan. Automatic Retrieval with Locality Information using Smart. *Proceedings of the First Text*

Retrieval Conference. Gaithersburg, 1992. pp 59-72.

[Buz&al 2001] J.W. Buzydiowski, H.D. White and X. Lin. Term Co-occurrence Analysis as an Interface for Digital Librairies. *Lecture Notes in Computer Science*. 2002. N°2539. pp 133-144.

[Cao&al 2006] F. Cao, M. Ester, W. Qian, and A. Zhou. Density-based Clustering over an Evolving Data Stream with Noise. *Proceedings of the SIAM Conference on Data Mining (SDM'2006).* 2006.

[Car&al 2001] M.F. Caropreso, S. Matwin, and F. Sebastiani. A Learner Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization. A. G. Chin (edts.). *Text Databases and Document Management : Theory and Practice.* Idea Group Publishing. Hershey 2001. pp 78-102.

[Cha 1984] J. Chauché. Un outil multidimensionnel de l'analyse du discours. *Proceedings of the 22nd conference on Association for Computational Linguistics*. July 1984, Stanford California. pp 11-15.

[Cha 1990] J. Chauché. Détermination sémantique en analyse structurelle : Une expérience basée sur une définition de distance. *TA Information.* 1990. Volume. 31, N°1, pp 17-24.

[Che&al 2002] Y. Chen, G. Dong, J. Han, B. Wah and J. Wang. Multidimensional Regression Analysis of Time-series Data streams. *In Proceedings of the 28th International Conference on Very Large Databases (VLDB 02)*. 2002. Hong Kong, China. pp 322–334.

[Chi 1994] S. L. Chiu. Fuzzy Model Identification Based on Cluster Estimation. *Journal of Intelligent and Fuzzy System*. 1994. Volume 2, pp 267-278.

[Chu&al 2003] W. Chung, H. Chen and J. Nunamaker. Business Intelligence Explorer : A Knowledge Map Framework for Discovering Business Intelligence on the Web. *Proceedings of the 36 Hawaii International Conference on System Sciences* (HICSS'03). 2003. 10p.

[Chu&Han 1989] K. Church et P. Hanks. Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th Annual Meetings of the Association for Computational Linguistics (ACL 1989)*. ACL Press. 1989,Vancouver. pp 76-83.

[Clo&al 2006] F. Cloppet, N. Vincent et J-M Petit. *Extraction de connaissances: Etat et Perspectives*. Cépaduès Editions, Paris, France. Janvier 2006. I.S.B.N 2.85428.707.x. 428p.

[Dag&al 1997] I. Dagan, L. Lee and F. Pereira. (1997). Similarity-based Methods for Word Sense Disambiguation. *Proceedings of ACL 1997*. pp 56-63.

[Dag&al 1999] I. Dagan, L. Lee and F. Pereira. Similarity-based Models of Word Co-occurrence Probabilities. *Machine Learning*. 1999. 34 : pp 43-69.

[Dee&al 1990] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*. 1990. 41(6): pp 391-407.

[Dem&al 1977] A. Dempster, N. Laird and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Jounal of Royal Statistics Society*. 1977. 39: pp 185-197.

[Dou&Kar 2005] B. Dousset et S. Karouach. Manipulation de graphes de grande taille pour l'étude des réseaux d'acteurs et des réseaux sémantiques. *In the International Journal on line of Information Science and Decision Making (ISDM)*. N° 22. I.S.S.N 1265 499X. Juin 2005.

[Dum&al 1998] S. Dumais, J. Platt, D. Heckerman and M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of the 7th ACM International Conference on Information and Knowledge Management (CIKM 1998).* ACM Press. 1998, Nex York, USA. pp 148-155.

[Dun 1974] J. C. Dunn. *Well Separated Clusters and Optimal Fuzzy Partitions*. J. Cybern. 1974. Volume 4, pp 95-104.

[Fay&al 1996] U. M. Fayad, G. Piatetsky-Shapiro, P. Smyth and P. Smyth. *Advances in Knowledge Discovery and Data Mining*. AAAI Press. 1996. Menlo Park, CA.

[Fay&al 1991] C. Fay-Varnier, C. Fouqueré, G. Prigent et P. Zweingenbaum. Modules syntaxiques des systèmes d'analyse du français. *TSI – Techniques et Science Informatiques*. Editions AFCET-Bordas, 1991. Volume 10, N°6, pp 403-425.

[Fel&al 1998] R. Felman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler and O. Zamir. Text Mining at the Term Level. *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 1998)*. J.M. ZYTKOW et M. QUAFALOU (edts.). Lecture Notes in Artificial Intelligence (LNAI). Nantes, 1998. Volume 1510, pp 65-73.

[Fel&Hir 1998] R. Feldman and H. Hirsh. Finding Associations in Collections

of Text. *In R.S. Michalski, I. Bratko and M. Kubat, editors, Machine Learning and Data Mining: Methods and Applications*. John Wiley & Sons. 1998. New York, USA. pp 223-240.

[Fel&Dag 1995] R. Felman et I. Dagan. Knowledge Discovery in Textual Databases (KDT). *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 1998)*, J.M. ZYTKOW et M. QUAFALOU (edts). Lecture Notes in Artificial Intelligence (LNAI). Nantes, 1998. Volume 1510, pp 65-73.

[Fir 1957] J. Firth. A Synopsis of Linguistic Theory 1930-1955. *Studies in Linguistic Analysis*. *Réédité, Selected Paper of J.R. FIRTH, F. PALMER (edts.), Longman*. 1957. pp 82-95.

[Fuh&Buc 1991] N. Fuhr, and C. Buckley. A Probabilistic Learning Approach for Document Indexing. *ACM Transactions on Information Systems*. 1991. Volume 9, pp 223–248.

[Gia&al 2003] G. Giannella, J. Han, J. Pei, X. Yan, and P. Yu. *Mining Frequent Patterns in Data streams at Multiple Time Granularities.* In H. Kargupta, A. Joshi, K. Sivakumar and Y. Yesha (edts.). Next Generation Data Mining, MIT Press. 2003.

[Hab&al 1997] B. Habert, A. Nazarenko and A. Salem. *Les linguistiques de corpus*. Armand Colin/Masson (edts.). 1997. ISBN: 2200017758. 240p.

[Hal&al 2002] M. Halkidi, Y. Batistakis, M. Vazirgiannis. *Cluster Validity Methods: Part I and Part II*. In SIGMOD Record. June 2002.

[Har&al 1989] Z. Harris, M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T. N. Harris and S. Harris. *The Form of Information in Science: Analysis of an Immunology Sublanguage*. Preface by H. PUTNAM. Boston studies in the phlisophiy of science. Kluwer Academic Publishers. Boston, USA, 1989. 590p.

[Har 1988] Z. Harris. *Language and Information*. Columbia University Press. New York, 1988. 120p.

[Hea 1999] M. Hearst. Untangling Text Data Mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*. University of Maryland.1999. pp 3-10.

[Her 2004] N. Hernandez. *Description et détection automatique de structures de textes*, Phd dissertation. Université Paris-Sud XI, décembre 2004.

[Hof 1999] T. Hofmann. Probabilistic Latent Semantic Analysis. *Proceedings of Uncertainity in Artificial Intelligence Conference (UAI 1999)*. 1999, Stockholm. pp 289-296.

[Iha&al 2004] M. A. Ihadjadène. *Méthodes avancées pour les systèmes de recherches d'informations*. Ouvrage collectif sous la direction de M. A. Ihadjadène. Hermès sciences publications. Paris, France. 2004. ISBN : 2-7462-0846-6. 247p.

[Jai&al 2005] S. Jaillet, M. Teisseire et G. Dray. Adéquation des modèles de représentation aux méthodes de catégorisation. *Revue Ingénierie des Systèmes d'Information (ISI), numéro spécial « Fouille de données complexes »*. 2005. 19p.

[Jai 2005] S. Jaillet. *Catégorisation automatique de documents textuels : d'une représentation basée sur les concepts aux motifs séquentiels*. *PhD Thesis*. Université Montpellier II, 2005.

[Jai&al 1999] A.K. Jain, M.N. Murty and P.J. Flynn. *Data Clustering : A Review*. ACM Computing Surveys. 1999. Volume 31-3, pp 264-323.

[Jai&Dub 1998] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall. 1988.

[Joa&al 2006] J. Joachim, J. Kister, Y. Bertacchini et H. Dou. Intelligence économique et système d'information. *In the International Journal on line of Information Science and Decision Making (ISDM)*. N° 24. I.S.S.N 1265 499X. 2ème trimestre 2006.

[Joh 1967] S. C. Johnson. *Hierarchical Clustering Schemes.* Psychometrika. 1967. Volume 2, pp 241-254.

[Kay 1988] D. Kayser. What Kind of Thing is a Concept ? *Computational Intelligence Journal*. 1988. Volume 4, pp 158-165.

[Kod 2000] Y. Kodratoff. Data Mining and Text Mining. *Actes des journées Extraction des Connaissances à partir de Données : Data Mining, OLAP and Data Warehousing*. Tunis, Tunisie, mai 2000. pp 6-9.

[Koh&al 2000] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero and A. Saarela. Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks.* 2000. Volume 11, N° 3. pp 574-585.

[Lan&Lit 1991] T.K. Landauer, M. Littman. A Statistical Method for Language Independent Representation of the Topical Content of the Text Segments. *Actes du 10éme congrès sur l'Information, la Documentation et le Transfert de connaissances (IDT 1991)*. Ed. Adbs & anrt (edts.).

Avignon, 1991. pp 77-85.

[Lan&al 1998] T. Landauer and P. Foltz and D. Laham. Introduction to Latent Semantic Analysis. *Discourse Process*. 1998. 25: pp 259-284.

[Lee 1995] J. H. Lee. Combining Multiple Evidence from Different Properties of Weighting Schemes. *Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 1995)*. Washington, USA, 1995. pp 180-188.

[Lel 2003] A. Lelu. Graphes de similarité entre texts : les constituer et les exploiter. *Journées Graphes, Réseaux et Modélisation (GRM 2003)*. Paris 2003. 2p.

[Len&al 1997] B. Lent, R. Agrawal and R. Srikant. Discovering Trends in Text Databases. *In Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining (KDD 97).* August 1997. Newport Beach, California. pp 227-230.

[Les 1969] M. Lesk. Word-word association in Document Retrieval Systems. *American Documentation*. 1969. 20 : pp 27-38.

[Lew 1992a] D. Lewis. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. N. J. BELKIN, P. INGWERSEN AND A.M PEJTERSEN (edts.). *Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1992).* ACM Press. 1992. pp 37-50.

[Lew 1992b] D. Lewis. *Representation and Learning in Information Retrieval*. *PhD thesis*, Department of Computer Science, University of Massachusetts. 1992, USA.

[Lio&al 2004] J. Lioréns, M. Valasco and A. De Amescua. Automatic Generation of Domain Representations using Thesaurus Structures. Mining Interesting Sequential Patterns for Intellligent Systems. *Journal of the American Society for Information Science and Technology*. Wiley InterScience 2004. Volume 55, Issue 10, pp 846-858.

[Lov 1968] J-B Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics.* 1968. Volume 11, pp 22-31.

[Mac 1967] J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. 1967. Berkeley, University of California Press. Volume 1, pp 281-297.

[Mot&al 2003] J. Mothe, C. Chrisment, B. Dousset and J. Alau. DocCube : Multi-dimensional Visualisation and Exploration of Large Document Sets. *JASIST*. 2003. 54(7): 650-659.

[Naz 2004] A. Nazarenko. Donner accès au contenu des documents textuels, Acquisition de connaissances et analyse de corpus spécialisés. *Mémoire d'Habilitation à diriger les recherches, LIPIN, Université Paris 13*. Villetaneuse, Décembre 2004.

[Par&Raj 2000] P. Paroubek et M. Rajman. Etiquetage morpho-syntaxique. *Dans le chapitre 5 de l'ouvrage « Ingénierie des Langues »*, sous la direction de J-M Pierrel. Collection Information Commande Communication, Editions Hermès Sciences. Paris, octobre 2000. ISBN 2-7462-0113-5. pp 131-148.

[Paz 1999] M.T. Pazienza. *Information Extraction.* LNCS, Springer Verlag. Berlin, 1999. ISBN : 3-540-66625-7. 164p.

[Pea&Wil 1991] H.J. Peat and P. Willett. The Limitations of Term co-occurrence Data for Query Expansion in Document Retrieval Systems. *Journal of the American Sociiety for Information Science*. 1991. 42(5) : pp 378-383.

[Poi 2003] T. Poibeau. *Extraction automatique d'informations, du Text brut au web sémantique*. Hermès sciences publications. Paris, 2003. ISBN : 2-7462-0610-2. 238p.

[Por 1980] M.F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3). 1980. pp 130-137.

[Pla&al 2005] M. Plantié, J. Montmain and G. Dray. Movies Recommander System: Automation of the Information and Evaluation Phases in Multi-criteria Decision-Making Process. In proceedings of the 6th International Conference on Database and Expert Systems Applications (DEXA 2005). August 2005. Copenhaguen, Denmark.

[Qiu&Fre 1993] Y. Qiu and H.P. Frei. Concept-based Query Expansion. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1993, Pittsburgh, USA. pp 160-169.

[Qiu&Fre 1995] Y. Qiu and H.P. Frei. Improving the Retrieval Effectiveness by a Similarity Thesaurus. *Technical Report (TR 225)*. 1995. Department of Computer Science, Zrich, Switzerland.

[Raj&al 2000] M. Rajman, R. Besançon et J-C Chappelier. Le modèle DSIR: une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement automatique des langues, 41(2).* 2000. pp 549-578.

[Rij 1977] C.V. Rijsbergen. A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. *Journal of Documentation*. 1977. 33(2) : 106-119.

[Rob&Spa 1976] S. Robertson and K. Sparck Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*. 1976. 27(3) : pp 129-146.

[Rob 1977] S. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*. 1977. Volume 33, pp 294-304.

[Rob&al 1981] S. Robertson, C.V. Rijsbergen and M. Porter. *Probabilistic Models of Indexing and Searching*. R.N. Oddy, S.E. Robertson, C.V. Rijsbergen, and P.W. Williams (edts.), formation Research and Retrieval. Butterworths. 1981. Chapter 4, pp 35-56.

[Rob&Wal 1994] S. Robertson and S. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994).* Dublin, 1994. pp 232-241.

[Rob&al 1994] S. Robertson, S. Walker, S. Jones, M.H. Beaulieu and M. Gatford. Okapi at TREC-3. *Proceedings of the 3rd Text Retrieval Conference*. 1994. pp 109-126.

[Sal&Les 1965] G. Salton and M.E. Lesk. The SMART Automatic Document Retrieval Systems - An Illustration. *Commun, ACM*. 1965. (6): pp 391-398.

[Sal 1971a] G. Salton. *The SMART Retrieval System. Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs. 1971.

[Sal 1971b] G. Salton. The Performance of Interactive Information Retrieval. *Information Processing Letters*. 1971. (2): pp 35-41.

[Sal 1973] G. Salton. Recent Studies in Automatic Text Analysis and Document Retrieval. *ACM Journal.* 1973. 20(2) : pp 258-278.

[Sal&McG 1983] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. 1983. McGraw Hill.

[Sal&Buc 1988] G. Salton and C. Buckley. Term-Weighting Approaches in

Automatic Text Retrieval. *Information Processing and Management*. 1998. (5): pp 513-523.

[Sal 1989] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley. 1989. ISBN:0-201-12227-8. 530p.

[Sal 1991] G. Salton. The SMART Information Retrieval System after 30 years - Panel. *SIGIR 1991*. pp 356-358.

[Sch&Ped 1994] H. Schütze and J. O. Pedersen. A Co-occurrence-based Thesaurus and Two Applications to Information Retrieval. *Proceedings of the RIAO 1994*. Rockefeller University, New York. pp 266-274.

[Sch&al 1995] H. Schütze, D. Hull and J. O. Pedersen. (1995). A Comparison of Classifiers and Document Representations for the Routing Problem. *Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1995), E.A. FOX, P. INGWERSEN and R. FIDEL (edts.).* ACM Press. July 1995, Washington, USA. pp 229-237.

[Seb 2006] F. Sebastiani. *Classification of Text, Automatic*. In Keith Brown (edts.), The Encyclopedia of Language and Linguistics. Elsevier Science Publishers. 2006. Amsterdam, NL. Volume 14, 2nd Édition pp. 457-462.

[Sin 1997] A. Singhal. *Term Weighting Revisited*. *PhD thesis*. Department of Computer Science, Cornell University. 1997.

[Spa&al 1998] K. Sparck Jones, S. Walker and S. Robertson. A Probabilistic Model of Information Retrieval: Development and Status. *Technical Report (TR 446).* Cambridge University, Computer Laboratory. 1998.

[Spi 2002] E. Spinat. Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ? *Colloque Cartographie de l'information : De la visualisation à la prise de décision dans la veille et le management de la connaissance.* 2002. Paris.

[Tze&Har 1993] K. Tzeras and S. Hartmann. Automatic Indexing Based on Bayesian Inference Networks. *Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1993), R. Korfhage, E. Rasmussen and P. Willett (edts.).* ACM Press. July 1993, New-York, Pittsburgh, USA. pp 22-34.

[Uch&Zhu 2005a] H. Uchida and M. Zhu. UNL 2005 for Providing Knowledge Infrastructure. *Invited speech, Semantic Computing Initiative Workshops*. May 2005. Chiba, Japan. 12p.

[Uch&Zhu 2005b] H. Uchida and M. Zhu. UNL 2005 from Langage Infrastructure toward Knowledge Infrastructure. Special Speech, *Pacific Association for Computational Linguistics (PCLING 2005)*. August 2005, Tokyo, Japan. 16p.

[Wan&al 1999] K. Wang, S. Zhou and S.C. Liew. Building Hierarchical Classifiers using Class Proximity. *In Proceedings of the International Conference on Very Large Databases (VLDB'99)*. 1999. Edimburgh, UK. pp 363-374.

[Wis&Van 2004] F. Wiesman and H.J. Van Den Herik. Information Retrieval by Metabrowzing. *Journal of the American Society for Information Science and Technology*. Wiley InterScience. 2004. 55(7): pp 565-578.

[Zan 2005] A. Zanasi. *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*. Advanced in Management Information Series, TEMIS Text Mining Solutions. Italy, 2005. ISBN. 1-85312-995-x. 368p.

Annexe A

Cette partie contient quelques extraits de documents qui sont utilisés pour illustrer les définitions et concepts introduits dans le cadre de ce mémoire. Le volume de ces extraits de documents est réduit de façon à mieux illustrer les concepts et ne pas surcharger le lecteur. Ces documents sont extraits à partir de journaux internationaux disponibles sur le Web. Les thématiques associées à ces documents sont respectivement :

- L'Economie du football,
- La Sécurité sur Internet,
- La Création d'entreprises.

Extrait de document sur l'Economie du Football :

« Actualités de l'économie du football français.

Afin de relancer le spectacle et l'offensive en championnat, le Conseil d'administration a décidé de lancer, dès la saison prochaine, le Challenge de l'offensive.

Le football professionnel français a pris aujourd'hui deux décisions importantes. Le lancement du Challenge de l'offensive et la réforme de la Coupe de la Ligue qui montrent sa capacité à innover.

Le Challenge de l'offensive n'aura pas d'incidence sur le classement sportif. Il permettra d'évaluer la pertinence et l'intérêt du nouveau barème Hidalgo en récompensant les équipes les plus offensives. La Commission d'organisation des compétitions officialisera à chaque journée de championnat, le classement sportif et le classement du challenge de l'offensive. »

Extrait de document sur la Sécurité sur Internet :

« Droits d'auteurs et sécurité sur internet.

Une chose est sûre, plus les majors insistent pour faire adopter des mesures restrictives, plus les gens qui téléchargent le font comme un acte de protestation contre l'industrie techno-fasciste. Les majors ont défendu dans la loi sur l'économie numérique le filtrage aux frontières sur internet. Seuls des pays comme la Chine ou la Corée du Nord le font. Je pensais que la loi se voulait une défense du droit d'auteur. Je constate avec regret qu'il défend surtout les intermédiaires pressés de détourner nos oeuvres pour leur seul profit.

Vous êtes passionné de sécurité informatique ? Alors ce qui suit va vous intéresser : Un petit groupe d'étudiants de l'école d'ingénieurs ESIEA organise un concours de sécurité informatique intitulé Challenge SecuriTech. Nous le savons tous : la sécurité informatique est devenue un élément phare de la communication moderne. Quelle société ne craint pas l'exploitation de ses données les plus sensibles à des fins malhonnêtes (espionnage industriel, pressions, etc.) ou la "simple" récupération de ses bases clients ? Quel particulier souhaiterait voir ses données personnelles dévoilées en intégralité ? Savoir sécuriser un réseau requiert, avant tout, de

fortes connaissances, aussi diverses soient-elles, des mécanismes informatiques. Il faut connaître les dangers qui peuvent le menacer, et donc connaître les techniques et démarches qui permettraient à une tierce personne de s'y infiltrer. Ce serait se voiler la face que de dire que nous savons protéger une machine sans savoir l'exploiter. »

Extrait de document sur la Création d'Entreprises :

« La récompense des créateurs et repreneurs d'entreprises.

Les CCI du Finistère organisent pour la sixième fois le challenge des Espoirs de l'économie française. Soutenue par le Conseil général du Finistère, cette opération vise à récompenser des créateurs et repreneurs d'entreprise finistériens dans différents secteurs (commerce/services aux personnes, industrie/services aux entreprises, jeune créateur de moins de 31 ans, demandeur d'emploi de longue durée), ainsi qu'un établissement scolaire pour un projet d'entreprise réalisé par ses élèves dans le cadre de leur scolarité.

À travers ces réussites individuelles, les CCI de Brest, Morlaix, Quimper Cornouaille et le Conseil général du Finistère entendent saluer l'audace de celles et ceux qui osent l'aventure de la création ou de la reprise d'une entreprise. Avec l'espoir que d'autres, à leur tour, connaissent cette insatiable nécessité d'entreprendre. »

Annexe B

Cette partie contient un exemple illustrant un extrait d'article de presse au format XML similaire aux articles de la base de données des expérimentations menées dans le cadre du projet de transfert de technologie.

Exemple d'extrait d'article de presse au format XML :

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="ExempleML.xsl"?>
<!DOCTYPE  SyS "exemple">
<ExempleML version="0.2" builder="Test">
        <article        date="2004-06-17"
                        heure="13:36:03"
                        lang="FR"
                        tz="+00:00"
                        pri="3">
                <source name="Le Monde" code="ECO"/>
                <auteur>/A</auteur>
                <titre>Pacific - Mise en Bourse filiale consommation</titre>
                <accroche>USA-GEORGIA PACIFIC-BOURSE </accroche>
                <resume/>
                <copyright>(c) Le monde 2004. </copyright>
                <corps>
<p> WASHINGTON, 17 juin - La filiale produits de consommation du groupe papetier Georgia-Pacific , CP&amp;amp;P, adéposé lundi auprès de la Securities and Exchange Commission, son projet d'introduction en Bourse qui devrait permettre de lever un milliard de dollars.</p> <p> CP&amp;amp;P, qui fabrique du papier de toilette et des mouchoirs en papier, a réalisé un chiffre d'affaires de 12,4 milliards de dollars au titre de son excercice fiscal 2001. Les produits tirés de la mise en Bourse serviront à rembourser une partie de la dette transférée par la maison mère.</p> <p> Aucune précision n'est pour l'instant donnée en terme de nombre d'actions ou de fourchette de prix d'introdution. Ces informations seront incluses dans d'autres documents qui seront remis par la suite aux autorités boursières.</p> <p> Georgia-Pacific gardera une participation non spécifiée dans sa filiale au terme de cette introduction en Bourse. Le groupe papetier prévoit toutefois de distribuer à ses actionnaires le restant de ses titres dans CP&amp;amp;P dans les six mois qui suivront la mise en Bourse.</p> <p> CP&amp;amp;P, dont le siège est à Atlanta, ne sera ainsi plus dépendante des procès sur l'amiante qui ont affecté le cours de Bourse de Georgia OPacific.</p> <p> L'introduction en Bourse sera menée par Goldman Sachs, Banc of America Securities et Morgan Stanley. /DRO</p>
</corps>
                <rubrique>
                        <meta name="slugline" value="USA-GEORGIA PACIFIC-
BOURSE"/>
                        <meta name="IIM:CATEGORY" value="F"/>
```

```
                    <meta name="IIM:SUPPLEMENTALCATEGORY" value="US"/>
                    <meta name="IIM:SUPPLEMENTALCATEGORY" value="IPO"/>
                    <meta name="IIM:SUPPLEMENTALCATEGORY"
value="GP.N"/>
                    <meta name="IIM:SUPPLEMENTALCATEGORY" value="CPX"/>
                    <meta name="N2000:TOPIC" value="US"/>
                    <meta name="N2000:TOPIC" value="IPO"/>
                    <meta name="N2000:TOPIC" value="GP.N"/>
                    <meta name="N2000:TOPIC" value="CPX"/>
                    <meta name="IIM:SUBJECTREFERENCE"
value="IPTC:04000000"/>
                    <meta name="IIM:SUBJECTREFERENCE"
value="IPTC:04016000"/>
                    <meta name="IIM:SUBJECTREFERENCE"
value="IPTC:04016019"/>
                </rubrique>
        </article>
</ExempleML>
```

Résumé. Ce travail de thèse s'inscrit dans le contexte d'extraction de connaissances à partir de documents textuels, appelé *Fouille de textes* (*FdT*) ou *Text Mining (TM).* Ce mémoire s'articule autour des problématiques liées à la modélisation de documents et la représentation de connaissances textuelles. Il s'intéresse à des collections de documents qui abordent des thématiques différentes. Le mémoire s'attache à élaborer un modèle de représentation et un système permettant d'extraire automatiquement des informations sur les différentes thématiques abordées mais également des mécanismes offrant la possibilité d'avoir des aperçus sur les contenus. Il est montré que les approches basées sur les associations de termes sont adaptées à ce contexte. Cependant, ces approches souffrent de certaines lacunes liées au choix du modèle et de la connaissance à retenir. Pour l'élaboration du modèle de représentation, le choix porte sur l'extension de l'approche d'association de termes. A cet effet, la notion de contexte est étudiée et un nouveau critère appelé « partage de contextes » est défini. Via ce critère, il est possible de détecter des liens entre termes qui n'apparaîtraient pas autrement. L'objectif est de représenter le plus de connaissances possibles. Ces dernières sont exploitées pour une meilleure représentation du contenu et des informations enfouies dans les textes. Un système appelé *IC-DOC* est réalisé, ce dernier met en œuvre le modèle de représentation dans un nouvel environnement d'extraction de connaissances à partir de documents textuels. Dans un contexte de veille scientifique, la proposition de ce type de systèmes devient indispensable pour extraire et visualiser de manière automatique l'information contenue dans les collections de documents textuels. L'originalité du système *IC-DOC* est de tirer profit du modèle de représentation proposé. Une série d'expérimentations et de validations sur divers jeux de données sont réalisées via le système *IC-DOC*. Deux applications sont considérées. La première s'intéresse à la caractérisation thématique et la seconde étend la première pour une cartographie visuelle de connaissances textuelles.

Representation of Textual Document Collections : Application to Thematic Characterization

Abstract. The general context of this PhD Thesis is related to the Knowledge Discovery from Textual Documents, called Text Mining. This work treats the problems of Document Modeling and Textual Knowledge Representation. It is interested in document collections related to different thematics. This PhD Thesis works out a Representation Model and a System allowing automatic extraction of information on the thematics. It is shown that the approaches based on Term Associations are adapted to this context. However, these approaches are limited by problems related to the model and knowledge to be retained. For the development of the representation model, the association term model is extended. The concept of context is studied and a new criterion called « context sharing » is defined. The objective is to represent the most possible knowledge which are exploited for a relevant representation of the contents and the information hidden in the texts. A system called *IC-DOC* is implemented, this last exploits the representation model in a new environment of Knowledge Extraction from Textual Documents. A series of experimentations and validations on various textual document collections are realized with *IC-DOC* system. Two applications are considered. The first is interested in the Thematic Characterization. The second extends the first for a Textual Knowledge Cartography.

Discipline. Informatique.

Mots-clés. Fouille de textes, Extraction et découverte de connaissances, Représentation de connaissances, Partage de contextes, Caractérisation thématique, Cartographie d'informations.
Keywords. Text Mining, Knowledge Extraction and Representation, Context Sharing, Thematic Characterization, Information Cartography.

LGI2P. site EERIE, Parc scientifique Georges Besse, 30035 Nîmes cedex 1, France.
Tél. +33 4. 66. 38. 70. 27 Fax. +33. 4. 66. 38. 70. 74