الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي

Université Amar Telidji Laghouat
Faculté des Sciences

جامعة عمار ثليجي - الأغواط
كلية العلوم

DEPARTEMENT D'INFORMATIQUE

THÈSE

Présentée par

Mustapha BOUAKKAZ

Pour l'obtention du diplôme de

DOCTORAT EN SCIENCES

Spécialité: INFORMATIQUE

THÈME

L'agrégation Sémantique OLAP

Soutenue publiquement le 28/01/2017

DEVANT LE JURY

Mohamed Bachir YAGOUBI	Président	Professeur	Université de Laghouat
Abderrahmane LAKAS	Examinateur	Professeur	Université des Emirats Arabes Unis
Abdelouahab MOUSSAOUI	Examinateur	Professeur	Université de Sétif 1
Hacene BELHADEF	Examinateur	MCA	Université de Constantine 2
Youcef OUINTEN	Encadreur	MCA	Université de Laghouat
Sabine LOUDCHER	Co-Encadreur	Professeur	Université de Lyon 2 - France

Remerciements

Cette thèse est le fruit de cinq années d'efforts incessants, mais aussi d'échanges bénéfiques et de collaborations fructueuses. Ce travail n'aurait pas pu aboutir sans le concours précieux et généreux de personnes qui partagent la même passion pour la recherche scientifique. C'est avec un énorme plaisir que je remercie aujourd'hui tous ceux qui m'ont soutenu durant ces cinq années de travail pour faire réussir cette thèse.

Je tiens à remercier mon directeur de thèse, Monsieur Youcef Ouinten, Maître de conférence à l'université de Laghouat, pour avoir accepté de diriger mes travaux de recherches. Je le remercie pour la patience, la gentillesse et la disponibilité dont il a fait preuve. Qu'il trouve ici l'expression de ma très grande gratitude.

Je tiens aussi à exprimer ma plus profonde gratitude à Madame Sabine Loudcher, Professeur à l'université Lyon 2 et co-encadrante de ma thèse pour la proposition de ce sujet et pour l'intérêt qu'elle a manifesté à l'égard de mes travaux de recherches ainsi que pour son soutien et sa patience. Je la remercie également de m'avoir accueilli au laboratoire ERIC lors des stages que j'ai effectués dans le cadre de la réalisation de cette thèse.

Je tiens à remercier Mr Mohamed Bachir Yagoubi, Professeur à l'université de Laghouat, Mr Lakas Abderrahmane, Professeur à Université d'Alain, Émirats Arabes Unis, Mr Abdelouahabe Moussaoui, Professeur à l'université de Setif 2 et Mr Hacene Belhadef, Maître de conférence à l'université de Constantine 2, pour l'intérêt qu'ils ont porté à mes travaux en examinant ce mémoire et pour l'honneur qu'ils me font en participant à ce jury.

Je remercie également tous les membres du laboratoire LIM ainsi que tous les enseignants du département d'informatique de l'Université de Laghouat.

Dédicace

.

A mes très chers parents qui m'ont aidé, soutenu et encouragé.

A ma femme et ses parents qui ont été de mon côté durant la réalisation de cette thèse.

A Yasmine et Nesrine...

Résumé : L'analyse en ligne OLAP propose des outils et des approches pour analyser les données stockées dans les entrepôts de données. Les opérateurs OLAP sont efficaces lors de l'exploration des données simples, cependant ils deviennent limités et inadéquats quand il s'agit de données complexes. Dans cette thèse nous nous focalisons sur les données textuelles et leur agrégation dans le contexte OLAP. Il devient nécessaire de faire évoluer l'OLAP pour qu'il s'adapte à la particularité des données textuelles. Cette évolution ne sera concrétisée que par (1) la proposition de nouvelles approches OLAP adaptées aux données complexes ; (2) la prise en considération de l'aspect sémantique des données en gardant l'esprit de l'OLAP. L'une des solutions est la combinaison entre l'OLAP, la fouille de données, la recherche d'information et la théorie des graphes. Cette solution sert à enrichir le processus d'analyse en ligne par d'autres approches issues de domaines différents. Dans ces travaux de thèse, nous nous sommes focalisés sur les approches d'agrégation des données textuelles. Nous avons proposé deux approches pour ce problème. La première se base sur une technique issue de la fouille de données. La deuxième proposition se base sur les fondements de la théorie des graphes. Nous avons appliqué nos propositions sur un domaine d'application, les réseaux sociaux, afin de résoudre le problème de la prédiction des liens sémantiques. Pour valider nos différentes contributions, nous avons réalisé une plateforme logicielle pour mesurer la performance de nos approches par rapport à celles proposées dans la littérature.

Mots clés : agrégation, analyse en ligne, données complexes, fouille de données, recherche d'information, OLAP social.

Abstract :

The online analytical processing OLAP offers many tools and approaches to analyze the data stored in the data-warehouses. The OLAP operators are effective when data are numerical ; however, they become limited and not suitable for unstructured data such as texts. Due to the fast growing of textual data, a need for new approaches that take into account the textual content of data in OLAP analysis emerges. One of the solutions is the combination between the four domains : OLAP, data mining, information retrieval and graph theory. In fact, this solution serves to enrich the online analytical processing by other approaches from different areas. This thesis focuses on the textual aggregation problem in the OLAP context. As a solution, we have proposed two approaches ; the first is based on a technique derived from the data mining, whereas the second proposal is based on the concepts of the graph theory. The latter is applied on social networks in order to solve the problem of the semantic links' prediction between members. To validate our different contributions, we have developed a software platform to measure the performance of our approaches compared to the other proposals existed in the literature.

Keywords : Aggregation, Complex data, Data mining, Information retrieval, Online Analytical Processing, Social OLAP

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

1.1 Contexte général

Dans cette dernière décennie, et face aux fluctuations économiques et à la concurrence croissante, le processus de prise de décision est devenu primordial pour les chefs d'entreprise. Ce processus repose sur la maîtrise des informations pertinentes qui circulent au sein de l'entreprise et les outils disponibles. L'évolution des systèmes d'information et de communication a accru la masse d'informations accessibles de façon exponentielle, ce qui rend leur exploitation difficile.

Pour surmonter ce problème du volume, les systèmes d'aide à la décision ont été mis en place au sein des entreprises. Ces systèmes permettent un traitement synthétique de l'information pour faciliter la prise de décision. Les résultats obtenus par ces systèmes sont utilisés par les décideurs en vue d'élaborer une vision globale pour piloter les entités économiques dont ils sont responsables et pour confirmer ou adapter leurs stratégies de conduite avec l'environnement économique.

Les technologies clefs dans ces systèmes d'aide à la décision sont celles des entrepôts de données et de l'analyse en ligne (OLAP : Online Analytical Processing). Un entrepôt est une collection de données, orientées sujet, intégrées, non volatiles et historisées, organisées comme support d'un processus d'aide à la décision [Inmon 1996]. Les données sont extraites, nettoyées, transformées en un format unique qui les prépare à l'analyse. Au niveau conceptuel, les données sont modélisées de façon multidimensionnelle avec des axes d'observation et des indicateurs à observer. Au niveau logique, le modèle multidimensionnel est exprimé par un schéma en étoile, en flocon de neige ou en constellation avec les notions de faits, mesures, dimensions et hiérarchies. Pour analyser les données de l'entrepôt de façon multidimensionnelle et interactive, l'OLAP offre aux décideurs la possibilité d'agréger, de visualiser et d'explorer les données à l'aide d'opérateurs [Codd 1993]. L'OLAP contient des opérateurs pour résumer les données sous forme d'agrégats et des opérateurs pour les visualiser. Ces opérateurs sont dits de navigation et sont décomposés en opérateurs d'agrégation, de sélection et

de structuration. L'ensemble de ce processus est désigné par le terme d'entreposage des données.

De nos jours, le processus d'entreposage des données « simples », généralement des données numériques, est relativement bien maitrisé. Cependant, avec le développement de la technologie Web 2.0 et la profusion des données hétérogènes, les données disponibles sont devenues plus en plus complexes et sont hors de portée des systèmes d'aide à la décision traditionnels, faute d'outils ou de méthodes appropriés. Ces données complexes sont des données provenant de sources diverses et pouvant être multi-format, multi-structure, multi-modal et riches en sémantique. Elles peuvent être composées de textes, d'images, de sons et de vidéos. Des études récentes menées par Delphi Group [1] et Insee Group [2] affirment que 80% des informations des entreprises sont représentées sous forme des documents textuels [Hassan 2013].

Les données textuelles sont classées en trois types spécifiques, à savoir les données structurées, semi-structurées et non structurées. La majorité des travaux proposés dans la littérature mettent en évidence l'exploitation des données structurées et semi-structurées qui sont généralement enregistrées en format XML ou HTML. Cependant, peu de travaux prennent en considération les données non structurées telles que les notes, les comptes rendus et les rapports. La prise en compte des données non structurées dans le processus d'analyse en ligne nécessite différents aménagements de l'analyse OLAP pour l'adapter aux données complexes.

La manipulation des grands volumes d'information textuelle fait appel à une discipline appelée la Recherche d'Information (RI). Elle s'intéresse au développement des systèmes permettent de retrouver une information pertinente afin de satisfaire un besoin en information exprimé sous forme de requête. Ces systèmes sont appelés des systèmes de recherche d'information (SRI). Dans un SRI, le contenu textuel des documents est représenté par une représentation intermédiaire sous forme d'une matrice ou d'un graphe pour qu'on puisse le manipuler par la théorie des graphes.

D'un autre côté, la fouille de données (data mining) est une discipline qui a pris une place plus en plus importante dans cette dernière décennie. Elle est devenue une nécessité imposée par les besoins des décideurs de synthétiser et valoriser le volume de données qu'elles collectent. La fouille de données a été employée avec succès dans plusieurs domaines. Aucun domaine d'application n'est a priori exclu. Cette discipline se base sur la statistique et les techniques d'apprentissage automatique. Dans la lit-

1. http ://www.delphigroup.com/
2. http ://www.insee.fr/

térature, les techniques de fouille de données sont classées en trois catégories : (i) les techniques de visualisation et de description ; (ii) les techniques de structuration et de classification et (iii) les techniques d'explication et de prédiction.

Dans ce contexte, la combinaison entre la technologie OLAP avec la RI et les techniques d'extraction des connaissances (Data Mining) peut être considérée comme étant une solution envisageable. Cette combinaison offre une solution possible pour améliorer le processus d'aide à la décision. Cet assemblage permet de tirer profit des points forts et de combler les points faibles de chaque domaine [BenMessaoud 2004]. De plus, l'assemblage de l'OLAP, la RI et la fouille de données est capable d'apporter des réponses satisfaisantes au problème de l'analyse en ligne des données complexes. Il permet d'étendre les capacités des opérateurs OLAP classiques aux données complexes et enrichir les possibilités d'analyse et d'extraction des connaissances à partir des données complexes.

1.2 Problématique

L'intérêt pour l'analyse OLAP s'est développé énormément ces dernières années. Les entreprises se sont rendues compte de l'efficacité de la technologie OLAP dans l'analyse et l'exploration des données. Avec la profusion des données complexes, l'analyse en ligne doit s'adapter à la nature spécifique des données complexes. Cependant, les opérateurs OLAP tels que les opérateurs d'agrégation sont définis pour des données simples, et ils deviennent inadéquats quand il s'agit de données textuelles. Plusieurs questions se posent : Comment agréger dans l'OLAP les données complexes telles que les données textuelles ? Comment on peut faire évoluer l'OLAP vers une analyse sémantique des données complexes issues des réseaux sociaux ? et comment on peut appliquer l'agrégation textuelle pour résoudre le problème de la prédiction des liens sémantiques dans les réseaux sociaux tout en gardant l'esprit de l'OLAP ?

Ces interrogations deviennent cruciales, dès lors que la spécificité de ce type de donnée nécessite des traitements différents et implique l'utilisation des nouvelles approches d'entreposage et d'analyse en ligne par rapport à celles utilisées dans les systèmes décisionnels classiques.

Le processus d'entreposage des données textuelles constitue une problématique déjà connue dans la littérature [Tournier 2008], [Aknouche 2014]. Dans le cadre de cette thèse on cherche à analyser les documents textuels dans un but de les agréger et d'extraire les connaissances véhiculées par ces documents. Celles-ci nécessitent en

premier lieu à considérer chaque document comme étant une donnée élémentaire et la deuxième étape consiste à traiter le document textuel comme étant un ensemble de mots (simples ou composés) ayant des poids.

Généralement dans un document textuel les successions de mots dans un texte véhiculent des informations sémantiques non accessibles au premier abord. Plusieurs techniques ont été proposées dans la littérature, notamment dans le domaine de la recherche d'informations (RI) pour extraire des informions utiles à partir des données textuelles. En revanche, dans celui relevant de l'entreposage et de l'analyse en ligne des documents textuels, peu de travaux se sont intéressés à cette problématique.

La difficulté de l'agrégation des données textuelles et de l'extraction de la sémantique qu'elles véhiculent constitue toujours un problème ouvert pour la communauté scientifique. Dans ce contexte, l'utilisation des techniques relevant à la fois des domaines de la fouille de textes [BenMessaoud 2004], la recherche d'information, et de la théorie des graphes a constitué une des pistes prometteuses pour analyser des données textuelles dans un environnement multidimensionnel.

1.3 Nos contributions

Eu égard à ce qui précède, la réponse à la problématique posée dans le cadre de cette thèse est assujettie d'une part, au besoin de nouveaux opérateurs OLAP qui prennent en considération les données textuelles ; d'autre part, à l'évolution de l'OLAP vers une analyse sémantique des données textuelles.

Dans ce cadre, nous exposons un état de l'art des approches d'agrégation textuelle dans le contexte OLAP. Ceci nous a conduit à une nouvelle classification de ces approches : les approches basées sur la structure cube de données et les approches basées sur le contenu textuel. Cette dernière catégorie est divisée en trois sous-catégories celles des approches basées sur les connaissances linguistiques, celles des approches basées sur les connaissances externes et celles des approches basées sur les techniques statistiques. Toutes ces approches vont être détaillées dans notre travail afin de positionner nos contributions.

Après la synthèse des travaux proposés dans la littérature, nous proposons deux approches originales pour l'agrégation des données textuelles. Il s'agit d'un nouveau processus décisionnel qui permet aux décideurs d'obtenir des informations agrégées et pertinentes. Ces approches relèvent essentiellement des domaines de la RI, de la théorie des graphes et de la fouille de données [Bouakkaz 2016a].

Ces approches sont constituées de deux processus, un processus de prétraitement et d'épuration, et un processus d'agrégation adapté aux données textuelles. (i) Le processus de prétraitement et d'épuration des données textuelles, permet grâce aux techniques de la RI de procéder à l'extraction, au filtrage et à l'indexation des données textuelles. (ii) le processus d'agrégation adapté aux données textuelles, a comme objectif de pallier les limites des opérateurs OLAP classiques inadaptés aux données textuelles.

La première contribution baptisée GOTA, est une fonction qui agrège des données textuelles en utilisant la technique de fouille de données qui s'appelle $K-means$ avec une nouvelle distance adoptée appelée Google similarity distance [Bouakkaz 2015].

La deuxième contribution baptisée TAG, est une fonction d'agrégation automatique des données textuelles qui se base sur l'utilisation de la théorie des graphes, plus précisément, l'exploitation des circuits dans un graphe pour extraire les agrégats [Bouakkaz 2014].

Nous proposons un nouvel outil d'extraction des mots-clés à partir des documents textes. Il s'agit d'un prototype logiciel permettant de déterminer les mots-clés dans les documents en se basant sur le site Web "Microsoft Research Academic". Ces mots-clés serviront par la suite comme paramètre d'entrée dans notre plate-forme. Cette plate-forme d'agrégation des données textuelles dans le contexte OLAP est baptisée OLAP-TAS. Son objectif est de concrétiser nos différentes propositions et de les comparer avec quatre autres approches implémentées dans OLAP-TAS. Pour aborder le problème lié au corpus d'évaluation, nous avons évalué les différentes approches sur des collections de données réelles présentées dans les sections 5.2.1 et 5.2.2 [Bouakkaz 2016b].

Après la définition d'un domaine d'application de nos approches d'agrégation textuelle, qui est celui des réseaux sociaux, nous présentons une manière d'évolution et d'élargissement de l'OLAP vers le Sociale OLAP. Nous dressons un état de l'art sur les approches de prédiction des liens afin de positionner notre contribution et nous proposons une nouvelle approche baptisée DIAMANT pour la prédiction des liens sémantiques basés sur l'agrégation de contenu textuel publié par les membres du réseau.

Les résultats de nos contributions ont été validé par des publications dans des conférences et une revue internationale [Bouakkaz 2016a], [Bouakkaz 2014], [Bouakkaz 2015] et [Bouakkaz 2016b].

1.4 Organisation du manuscrit de thèse

Cette thèse est organisée comme suit : Dans le chapitre 2, nous introduisons les concepts de base utilisés le long de cette thèse. Nous exposons les notions d'entreposage de données, l'évolution des entrepôts de données classiques aux entrepôts de données complexes, les concepts de base des systèmes de recherche d'information et leurs mesures de performances.

Le chapitre 3 présente la formalisation de notre problématique d'agrégation des données textuelles par des opérateurs OLAP adéquats. Nous exposons un état de l'art général des approches d'agrégation textuelle. Nous proposons une nouvelle classification des approches proposées dans la littérature pour l'agrégation des données textuelles. Nous discutons, pour chacune de ces approches, le contexte et les travaux réalisés. Nous présentons également une synthèse permettant de positionner nos contributions au regard de l'existant [Bouakkaz 2016a].

Dans le chapitre 4, nous évoquons les objectifs et les motivations qui nous ont poussé à faire le couplage entre l'OLAP et la fouille de données d'une part et l'OLAP et la théorie des graphes d'une autre part. Nous exposons en détail les formalismes et les différentes étapes de nos approches. Ce chapitre inclut aussi des exemples d'applications pour simuler nos propositions [Bouakkaz 2014][Bouakkaz 2015][Bouakkaz 2016a].

Nous évoquons, dans le chapitre 5, notre outil d'extraction des mots-clés à partir des documents textuels en se basant sur le site Web «Microsoft Research Academia», et notre plateforme d'agrégation textuelle, baptisée OLAP-TAS, où nos approches GOTA et TAG ainsi que quatre autres approches proposées dans la littérature sont implémentées. Nous présentons les corpus d'évaluation utilisés (ITInnovation et OHSUMed). Nous exposons les différents indicateurs utilisés pour mesurer la qualité des approches implémentées, ainsi que les résultats expérimentaux évaluant la performance de nos approches d'agrégation textuelle dans le contexte OLAP [Bouakkaz 2016b].

Nos travaux pouvant être appliqués à l'analyse des réseaux sociaux, nous consacrons le chapitre 6 à une introduction aux réseaux sociaux et leur analyse. Nous décrivons le problème de la prédiction des liens avec un état de l'art sur les approches proposées dans la littérature. Ce chapitre inclut aussi un exposé sur les problèmes résolus par la prédiction des liens.

Le chapitre 7 présente l'évolution de l'OLAP classique vers le social OLAP. Nous détaillons notre proposition qui sert à faire la prédiction des liens sémantiques dans les réseaux sociaux en se basant sur l'agrégation du contenu textuel introduit par les

membres du réseau. Nous introduisons le jeu des données sur lequel nous déroulons notre approche et les cinq autres approches implémentées dans notre plateforme. Les résultats que nous avons obtenus confirment que notre proposition améliore considérablement la qualité des liens prédits par rapport aux autres approches.

Enfin, le chapitre 8 conclut cette thèse en présentant un bilan général de l'ensemble de nos contributions et en évoquant de nouvelles perspectives de recherche.

Chapitre 2

Concepts de base

Sommaire

2.1 Les systèmes d'aide à la décision

Le développement des technologies de l'information et de la communication a fait évoluer la masse d'informations disponibles aux utilisateurs de façon exponentielle. De nos jours, les entreprises ont rencontré un problème de maîtrise de l'information qui circule au sein de leur système d'information. Cette explosion de volume de l'information rend difficiles leur exploitation et leur manipulation par les décideurs. Pour surmonter ce problème, les systèmes d'aide à la décision ont été adoptés par les entreprises. Ils permettent aux décideurs d'effectuer des analyses pour bien piloter leurs entreprises. Les résultats de ces systèmes aident les décideurs à confirmer leurs lignes de conduite ou à changer périodiquement leur stratégie de gestion selon les changements de l'environnement économique.

Ces systèmes reposent sur des entrepôts de données qui correspondent à des bases de données dédiées à la prise de décision. Généralement, ces entrepôts n'utilisent

pas les modèles de données classiques (entité-association et relationnel), mais des modèles multidimensionnels où les données sont organisées sous forme des faits et des dimensions. Les modèles multidimensionnels (Etoile, Constellation et Flocon) sont adaptés aux analyses OLAP effectuées par le décideur [Inmon 1996].

Définition 1 *Un système d'aide à la décision est l'ensemble des outils matériels et logiciels qui permettent de collecter, de stocker et d'analyser des données issues du système d'information des entreprises dans le dessein de faciliter la prise de décision par les décideurs [Inmon 1996].*

Les utilisateurs de ces systèmes ne sont, généralement, pas des informaticiens mais des experts d'un domaine de l'entreprise (commercial, production, personnel). Ils sont chargés d'analyser les données décisionnelles pour assurer le pilotage de l'entreprise.

2.1.1 L'architecture de systèmes d'aide à la décision

Les systèmes d'aide à la décision permettent aux décideurs d'avoir une vision transversale de l'activité de l'entreprise et sa position dans son secteur. Ils se fondent sur la collecte et le stockage des données en provenance de différentes sources (base de données locales, emails...). Dans notre contexte, l'architecture de ces systèmes est composée de deux niveaux de stockage à savoir : les entrepôts et les magasins de données, et deux niveaux fonctionnels : les outils ETL (Extract Transform Load) et les outils de restitution et d'analyse [Inmon 1996].

Un entrepôt de données (data warehouse) constitue l'espace centralisé où les données décisionnelles, identifiées comme pertinentes pour l'aide à la décision, sont stockées de manière homogène, le plus souvent historisées, agrégées et organisées suivant un modèle assurant la gestion efficace des données (cohérence, validité et fraîcheur des données). Bill Inmon [Inmon 1996] définit l'entrepôt de données comme suit :

Définition 2 *: Un entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historiées, organisées pour le support d'un processus d'aide à la décision.*

Les Magasins de données (data marts), sont élaborés comme un extrait de l'entrepôt adapté à une classe de décideurs ou à un usage particulier et regroupent les données utiles pour un sujet d'analyse. Les données sont organisées suivant un modèle facilitant l'interrogation et l'analyse des données adaptées aux traitements décisionnels. Les magasins de données basées sur la modélisation multidimensionnelle sont créés

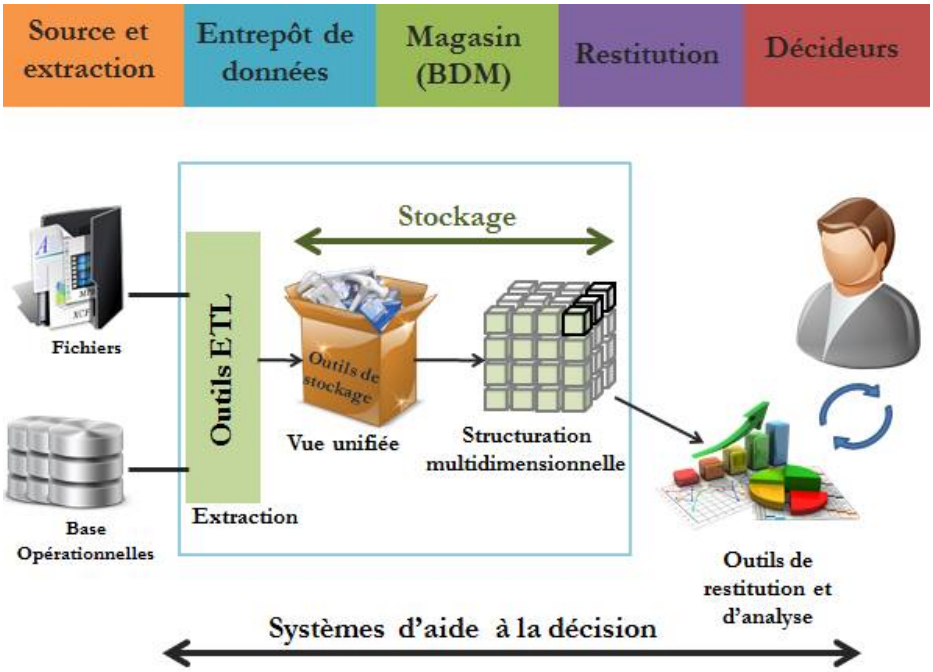

FIGURE 2.1 – Architecture d'un système d'aide à la décision

pour supporter le processus de prise de décision et les opérations d'analyse en ligne OLAP (Online Analytical Processing) [Codd 1993]. Les magasins de données sont représentés sous forme d'un cube ou d'un hyper-cube où les données sont vues sous forme d'un point d'intersection entre plusieurs axes dans un espace multidimensionnel. Chaque donnée représente une cellule du cube. Pour plus de visibilité, ces dimensions représentent les axes d'analyse et forment plusieurs niveaux de granularité. La figure 2.2 présente un exemple de cube de données qui permet l'analyse des publications scientifiques. Le cube est formé du mots-clés des articles scientifique en cellules qui représentent les indicateurs d'analyse disponibles du sujet d'analyse Article, appelés faits, et de quatre axes de dimensions représentant respectivement les Auteurs, les Conférences, le temps et le Contenu. Chacune de ces dimensions dispose de plusieurs niveaux de détail (Auteur, Laboratoire, université ...) et permettent d'obtenir une vision plus ou moins détaillée lors des analyses.

Les cubes de données sont interrogés et analysés via les opérateurs OLAP qui offrent une souplesse d'interrogation fondée sur les requêtes multidimensionnelles pour agréger les données d'une ou de plusieurs mesures d'un fait suivant les attributs d'une ou plusieurs dimensions. [Ravat 2007].

Les résultats obtenus seront affichés selon diverses structures de visualisation tels

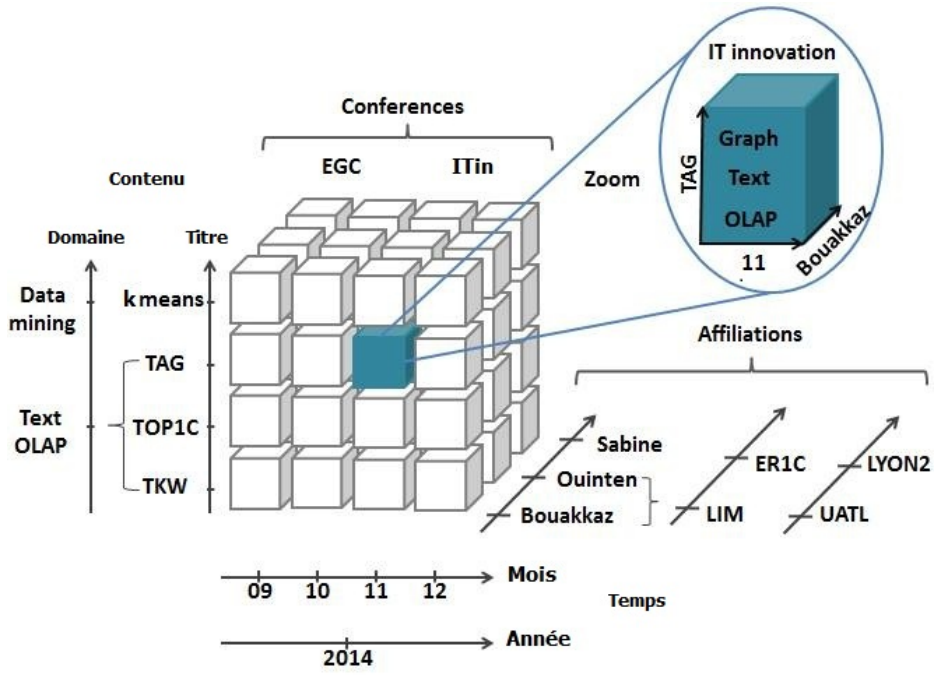

FIGURE 2.2 – Exemple de cube de données

que les courbes et les histogrammes. La structure de visualisation la plus utilisée dans le contexte OLAP est la table multidimensionnelle qui affiche les données selon deux axes [Ravat 2008].

2.1.2 L'analyse en ligne OLAP

L'analyse en ligne (OLAP) des données décisionnelles est une solution pertinente pour les décideurs. Par la suite, le concept système OLAP désignera un système d'aide à la décision dont les données sont stockées dans un entrepôt de données. Les outils d'analyse et de restitution de données sont basés sur la technologie OLAP.

2.1.2.1 Modélisation multidimensionnelle des données

La modélisation multidimensionnelle est une façon d'organisation des données pour les analyser. Elle permet de présenter les données selon un schéma bien défini. Ce schéma est constitué de deux composantes : dimensions et faits [Kimball 1996]. Les dimensions sont des axes d'analyse selon lesquels on veut étudier les données. Une table de faits contient deux types d'éléments, les données observables (les faits) représentant le sujet à analyser et les mesures correspondant aux informations de l'activité

à étudier. Les mesures peuvent être soit : (1) non additives, dans le cas où aucune dimension ne nous permet d'agréger les données. (2) semi-additives, dans le cas où certaines dimensions permettent d'agréger les données. (3) additives, dans le cas où l'ensemble des dimensions permettent d'agréger les données.

Plusieurs travaux ont été proposés dans la littérature pour la modélisation multidimensionnelle, les plus connus et plus utilisés parmi ces travaux sont le schéma en étoile, en flocon de neige et en constellation [Kimball 1996], [Chrisment 2005] et [Ravat 2007].

Schéma en étoile : un schéma en étoile tire son nom de sa représentation graphique qui est présentée sous-forme d'une étoile comme illustré dans la figure 2.3. Un schéma en étoile est constitué d'une table centrale de faits et de plusieurs tables de dimensions. Ce modèle représente de manière non-normalisée les dimensions [Bhide 2016] .

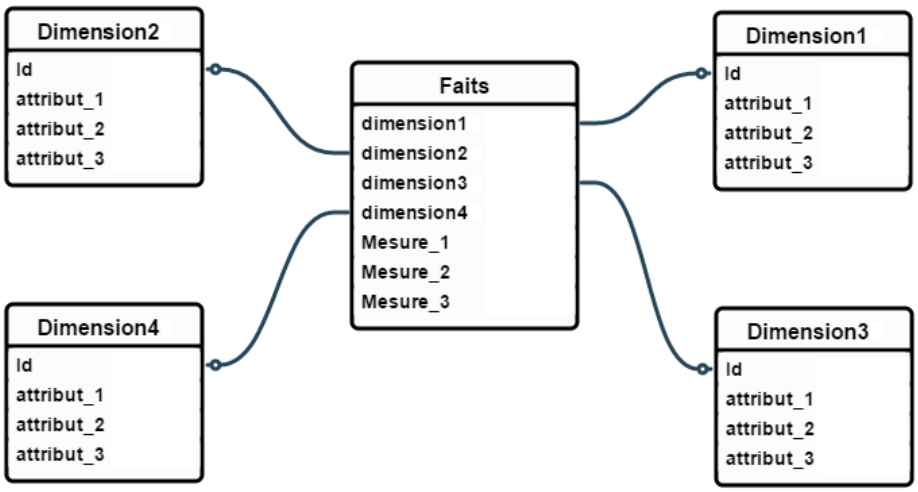

FIGURE 2.3 – Schéma en étoile

Schéma en flocon de neige : un schéma en flocon de neige est une variante du schéma en étoile. La différence réside dans la normalisation des tables de dimension. Ce schéma met les attributs de chaque niveau hiérarchique dans une table de dimensions. L'avantage de ce modèle est la formalisation d'une hiérarchie au sein d'une dimension et permet d'éviter le problème de redondance que l'on peut trouver dans le modèle en étoile [Bhide 2016].

Schéma en constellation : ce schéma est un ensemble de schémas en étoiles et/ou en flocon dans lesquels les tables de faits se partagent certaines tables de dimensions.

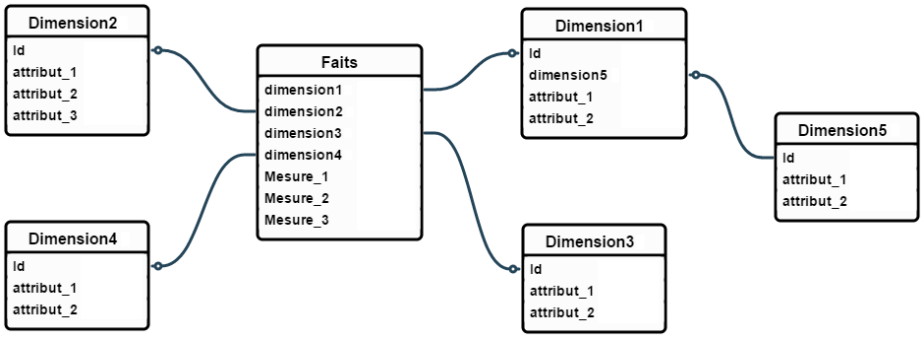

FIGURE 2.4 – Schéma en flocon de neige

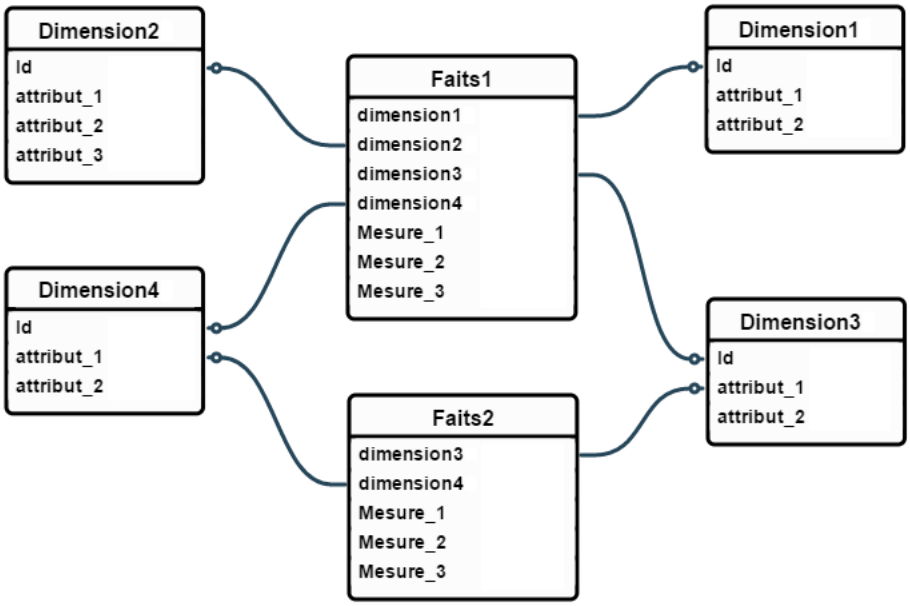

FIGURE 2.5 – Schéma en constellation

2.1.2.2 Manipulation des données multidimensionnelles

Les décideurs ont besoin d'outils de manipulation et d'analyse de données qui soient efficaces et conviviaux pour la prise de décision. Dans la représentation multidimensionnelle sous forme de cube, les cellules fournissent des informations agrégées pour une mesure, et selon plusieurs dimensions. Les décideurs utilisent plusieurs opérateurs permettant de manipuler cette structure multidimensionnelle.

Nombreux sont les travaux de recherche [Agrawal 1997], [Vassiliadis 1999] qui ont classé les opérateurs de manipulation OLAP. Dans [Laurent 2002], l'auteur identifie deux classes d'opérateurs de manipulation de données multidimensionnelles. La première concerne les opérateurs inter-cubes, qui ne changent que la présentation du cube, et la deuxième concerne les opérateurs intra-cubes qui modifient le contenu du cube. D'autres travaux ont proposé une autre typologie des opérateurs de manipulation, selon l'intuition du décideur qui cherche à obtenir, soit une forte interaction en utilisant des opérateurs de restructuration, soit une hiérarchisation de l'information et donc il fait appel aux opérateurs de granularité, soit il peut simplement chercher à adapter les opérateurs relationnels en utilisant les opérateurs dits classiques.

Les divers opérateurs définis pour la manipulation des données multidimensionnelles ne sont pas standardisés. Cependant, la majorité des travaux qui se sont intéressés à la manipulation des cubes de données [Bimonte 2006][Boukraâ 2010][Hassan 2013] partagent la même catégorisation des opérations OLAP, opérateurs relationnels, opérateurs de granularité et opérateurs de restructuration.

Les opérateurs relationnels : les opérateurs dits relationnels sont ceux qui étend les opérations de l'algèbre relationnelle. Les différentes définitions de la littérature précisent que les opérateurs classiques s'opèrent sur le contenu de cube sans aucune influence sur sa structure physique. Ainsi, les opérations relationnelles classiques sont directement traduites en opérations de l'algèbre relationnelle. On peut retenir comme opération classique : la sélection, la projection, le produit cartésien, la jointure, les opérations ensemblistes (union, intersection et différence), la suppression et l'ajout d'une mesure calculée.

Les opérateurs de granularité : ces opérateurs agissent sur la granularité de l'observation des données. ils guident la navigation entre les différents niveaux d'abstraction. Ces opérateurs nécessitent des informations non contenues dans le cube pour passer d'une représentation initiale à une représentation de granularités différentes, comme par exemple les opérations de forage qui permettent la navigation vers le haut

et vers le bas en se basant sur les deux fonctions roll-up et drill down. Ces deux fonctions permettent de remonter ou de descendre dans une hiérarchie de dimension vers un niveau plus agrégé ou détaillé, respectivement.

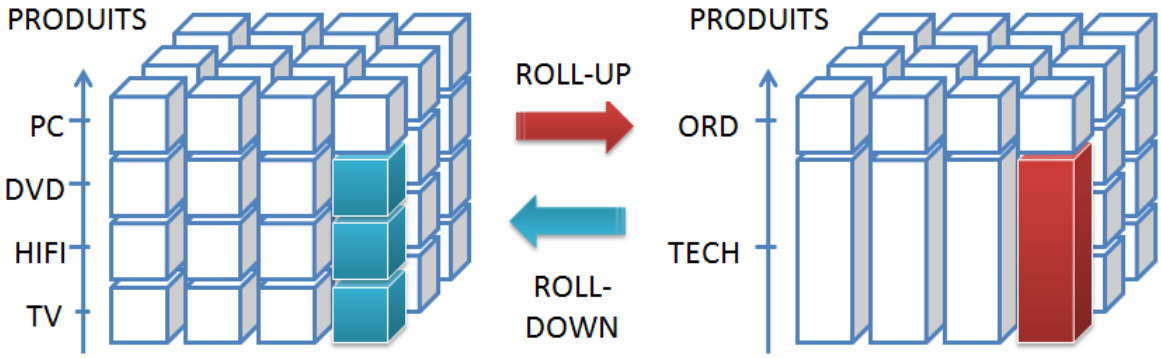

FIGURE 2.6 – Principe du forage

Les opérateurs de restructuration : les opérateurs de restructuration regroupent toutes les opérations élémentaires de changement de point de vue sur un cube. Ainsi, tout cube obtenu par restructuration d'un cube initial contient les informations nécessaires et suffisantes pour générer le cube initial par restructuration réciproque. Les opérateurs de restructuration changent seulement la structure d'un cube sans toucher le contenu. Les principales opérations de restructuration existant dans les systèmes OLAP sont : les opérations de rotation, les opérations d'ordonnancement et les opérations de transformation.
Les opérations de rotation réorientant une analyse en changeant l'axe d'analyse en cours (rotation de dimension).

Les opérations d'ordonnancement permettent de changer l'ordre des valeurs (positions) des paramètres ou celle des dimensions (switch) ou de réordonner les paramètres d'une hiérarchie (nest). Par généralisation, cette dernière permet d'imbriquer un attribut dans une autre hiérarchie. Les opérations de transformation permettent l'ajout d'attributs de dimension en tant qu'indicateur d'analyse (push) ou de convertir un indicateur d'analyse en paramètre (pull).

2.2 L'analyse des documents : au-delà des nombres

dans cette dernière décennie, l'analyse des données d'un entrepôt de données et les systèmes OLAP reposent sur les données numériques qui sont relativement bien

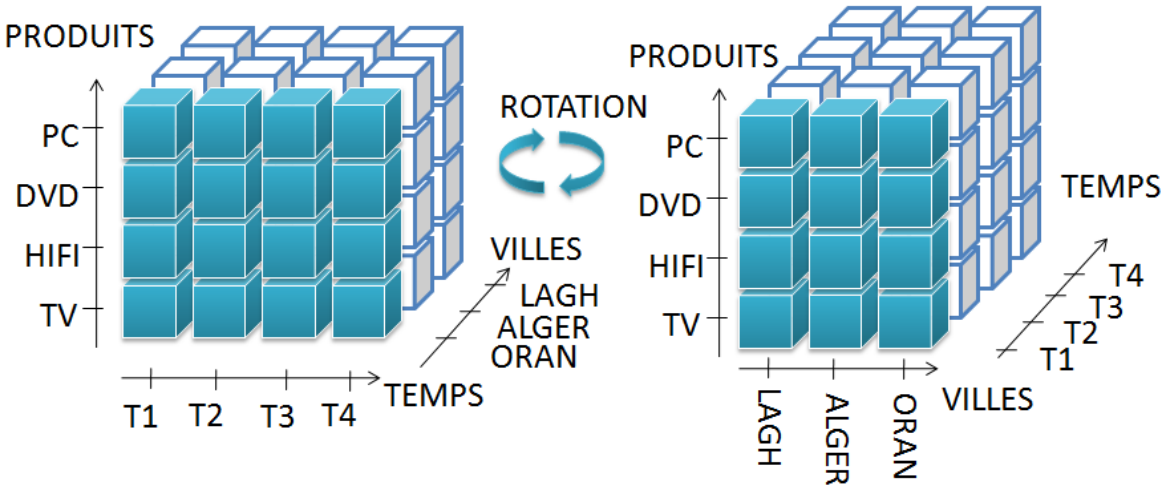

FIGURE 2.7 – Principe de rotation

maîtrisées. Une large gamme d'outils puissants d'analyse et de synthèse est disponible au profit des décideurs [Sullivan 2001] [Tseng 2006] [Campos 1]. Cependant, ces outils manipulent seulement 20% des données contenus dans les systèmes d'information des entreprises [Tseng 2006]. Les 80% restant des données sont des données complexes qui restent toujours hors de la portée des systèmes d'aide à la décision car elles contiennent des images, des vidéos et des documents textuels tels que les rapports, les notes et les articles.

Les documents représentent une source de connaissances importante au sein des entreprises. Il est donc naturel de pouvoir exploiter le contenu qu'ils portent par l'adoption et l'adaptation des techniques et des méthodes spécifiques à leurs types de données. Ces techniques sont inspirées des domaines de la recherche d'informations, la fouille de données et de la théorie des graphes.

2.2.1 Les différents types de documents

Le document représente le support d'information le plus utilisé au sein des entreprises. Il inclut plusieurs sous-éléments de différents types (textes, nombres, figures). Les documents se présentent sous différentes formes, par exemple, des factures, des rapports, des articles scientifiques ou encore des ouvrages. En général deux catégories de documents se distinguent :

- Les documents issus d'une base de données : par exemple des données organisées sous forme de tableaux, de transactions d'une base de données relationnelle. Ce sont

des documents dont le contenu est fortement structuré avec des champs clairement séparés et bien identifiés.

- Les documents textuels : par exemple, les emails, les livres électroniques et les articles scientifiques. Ce genre de document se caractérise par le fait que leur contenu est principalement composé de texte et non de champs. Donc, à l'instar de [Fuhr 2001], nous distinguons deux types de documents : (1) Les documents semi structurés représentent le contenu d'un document XML ou HTML. La structure arborescente est employée pour décrire bien le schéma des données. (2) Les documents non structurés qui sont principalement composés de textes tels que les versions électroniques des documents papiers. Ces documents ont une structure plus hétérogène et contiennent différents types de données tels que les images, et les tableaux. Dans ce type de document, on retrouve les articles scientifiques, les articles d'information, le contenu textuel d'un mail et les livres numériques (e-books).

Le traitement associé à ces deux types de documents est différent. En effet, un document semi structuré étant beaucoup plus facile à traiter qu'un document non structuré, les traitements à appliquer sur un document semi structuré sont relativement similaires aux traitements applicables sur des données très structurées telles que des bases de données. Cependant, un document non structuré nécessite des traitements plus complexes pour extraire les connaissances qu'il véhicule. Dans notre thèse nous nous intéressons à l'analyse des documents textuels non structurés dans le contexte OLAP.

2.2.2 Les entrepôts de documents

Vu le volume important des documents, plusieurs travaux de recherche ont été proposés pour intégrer les documents dans des entrepôts dits de documents et le terme «document warehousing» émergea dans [Sullivan 2001], [Abiteboul 2003], [Abiteboul 2006]. Les entrepôts de documents sont des entrepôts de contenus qui archivent des informations qualitatives alors que les entrepôts de données sont plus orientés vers des données quantitatives.

Dans [Sullivan 2001], [Tournier 2008], [Aknouche 2014], les auteurs ont envisagé l'intégration de documents dans un environnement OLAP pour permettre la conception d'outils d'analyse de textes. En allant au-delà de ces travaux, notre problématique s'articule autour de l'analyse multidimensionnelle de données textuelles et leur domaine d'application. Nous nous intéressons plus particulièrement à l'agrégation des données textuelles.

2.3 Les systèmes de recherche d'information

Un système de recherche d'informations (SRI) est un intermédiaire entre une collection de documents sous forme d'un corpus et des utilisateurs cherchant via des requêtes, des informations susceptibles de se trouver dans ces documents. Les SRI contiennent une série de méthodes permettant d'extraire ces informations. Plusieurs techniques, modèles et mécanismes ont été proposés pour améliorer les performances des SRI [Baeza-Yates 1999], [Grossman 2012]. Pour ce dernier point, il faut noter qu'il y a plusieurs études faites sur la RI qui ont donné naissance à un très grand nombre de publications. Nous restreignons cette première partie de chapitre aux concepts-clés de la RI et les techniques les plus populaires que l'on peut exploiter pour l'agrégation textuelle.

Définition 3 *La recherche d'informations est un domaine qui à pour objectif, la représentation, l'analyse, l'organisation, le stockage, et l'accès à l'information [Chowdhury 1984]*

Un système de recherche d'informations possède un ensemble de modèles pour la représentation des unités d'informations. Il intègre également un mécanisme de recherche/sélection. Ce dernier permet de sélectionner l'information pertinente en réponse aux besoins exprimés par l'utilisateur. Dans un SRI, plusieurs éléments clés y sont distingués [Grossman 2012] :

Le document : le document constitue l'information élémentaire d'une collection documentaire. L'information élémentaire peut représenter le tout ou une partie d'un document.

La collection de documents (corpus) : la collection de documents (ou corpus) constitue l'ensemble des informations accessibles et exploitables par l'utilisateur. Elle est constituée d'un ensemble de documents. Pour des raisons d'optimalité, la collection constitue des représentations simplifiées de ces documents. Ces représentations sont étudiées de telle sorte que la gestion et l'interrogation de la collection se fassent dans les meilleures conditions de coût.

Le besoin d'information : la notion de besoin en information où en recherche d'informations est souvent assimilée aux besoins de l'utilisateur. Il existe trois types de langage d'interrogation permettent de formuler les besoins [Bilhaut 2007] : **(1)** Interrogation en langage booléen où l'utilisateur exprime sa requête sous forme d'une suite de termes reliés entre eux par des opérateurs booléens (et, ou, non). **(2)**Interrogation en langage naturel ou quasi naturel où l'utilisateur exprime sa requête en langage libre

(langage naturel) sous forme de mots-clés et **(3)** Interrogation en langage graphique où une interface d'aide à la formulation de la requête est proposée à l'utilisateur. En effet, une vue d'ensemble de la base d'information et en particulier une vue de termes représentant le contenu sémantique des documents est donnée à l'utilisateur pour l'assister à formuler sa requête.

2.3.1 La représentation des documents

La représentation des documents est supportée par un ensemble de règles et de notations permettant la traduction d'un document d'une description brute vers une description structurée. Ce processus de conversion est appelé l'indexation. L'indexation est une opération permettant d'extraire d'un document une représentation paramétrée qui couvre au mieux son contenu sémantique.

Le résultat de l'indexation constitue le descripteur du document. Ce dernier est souvent une liste de termes ou groupe de termes significatifs pour l'unité textuelle correspondante, généralement assortis de poids représentant leur degré de représentativité du contenu sémantique de l'unité qu'ils décrivent. Les descripteurs des documents (mots, groupe de mots) forment le langage d'indexation. L'indexation est une étape primordiale dans la recherche d'informations. De sa qualité dépend en partie la qualité des réponses du système. Conscients de son importance, et soucieux de bien la réaliser, les développeurs des SRI ont proposé plusieurs manières de procéder. Les principales sont l'indexation manuelle et l'indexation automatique. Elles sont définies comme suit [Zargayouna 2004] :

Indexation manuelle : dans le cas de l'indexation manuelle, chaque document est analysé par un spécialiste du domaine ou par un expert documentaliste. En fonction de ses connaissances, cet expert détermine les mots-clés qui lui semblent les plus significatifs pour représenter le document. L'indexation humaine est une activité fondée sur le jugement d'un être humain. Elle se caractérise par sa profondeur, sa cohérence (ce qui est fondamental pour la cohérence du fond et des fichiers) et sa qualité (exhaustivité - spécificité). Elle est cependant trop dépendante de l'état des connaissances des indexeurs. Cela induit une subjectivité de ses résultats. Elle nécessite la lecture de l'intégralité des documents. Son application est de ce fait inadaptée à des collections de taille importante. L'indexation automatique permet de pallier ce problème.

Indexation automatique : L'indexation automatique reconnaît des chaînes constituées de mots non vides de sens. Elle détecte automatiquement les termes les plus représentatifs du contenu d'un document. Ce type d'indexation est actuellement

le plus répandu. Elle comprend deux étapes fondamentales : l'identification des termes d'indexation et l'évaluation de leurs poids.

L'identification des termes d'indexation consiste à analyser le texte du document mot à mot. Son objectif est d'extraire les mots-clés et éliminer les mots vides de sens qui ne jouent qu'un rôle syntaxique. Les mots vides sont identifiés puis éliminés grâce à un anti-dictionnaire (stop list en Anglais). Les mots apparaissant trop souvent n'ont aucun intérêt, ils sont également éliminés. Seuls les mots significatifs représentant les concepts du document sont retenus. Les opérations de l'indexation automatique se déroulent comme suit [Daille 2000] :

a- Analyse lexicale : l'analyse lexicale est le processus qui permet de convertir le texte d'un document en un ensemble de termes. Un terme est un mot considéré dans sa valeur de désignation, en particulier dans un vocabulaire spécialisé. L'analyse lexicale permet de reconnaître les espaces de séparation des mots, les chiffres, les ponctuations, etc.

b- L'élimination des mots vides : un des problèmes majeurs de l'indexation consiste à éliminer les mots vides (pronoms personnels, prépositions, ...). Les mots vides sont des mots peu significatifs augmentant ainsi la taille de l'index et rendant la recherche plus lente. De ce fait l'élimination est une étape indispensable. On distingue deux techniques pour éliminer les mots vides de sens :

- L'utilisation d'une liste de mots vide de sens (aussi appelée anti-dictionnaire)
- L'élimination des mots dépassant un certain nombre d'occurrences dans la collection.

c- Lemmatisation : Un mot donné peut avoir différentes formes dans un texte. On peut citer par exemple : économie, économiquement, économétrie, économétrique, etc. Il n'est pas forcément nécessaire d'indexer tous ces mots et un seul suffirait à représenter le concept véhiculé. Pour résoudre le problème, une substitution des termes par leur racine ou lemme est utilisée. Plusieurs méthodes de lemmatisation ont été proposées dans la littérature, parmi lesquelles : la troncature et la méthode des n-grammes [Jalam 2002] [Longrée 2016].

Afin d'évaluer les mots indexés et pour augmenter la qualité de la recherche, la pondération des termes extraits est primordiale. La plupart des techniques de pondération des termes sont basées sur les facteurs Tf et Idf [Ramos 2003] :

Tf (term frequency) mesure l'importance d'un terme dans un document. Cette mesure est calculée en fonction de la fréquence d'un terme dans un document.

Idf (Inverse of Document Frequency) mesure l'importance d'un terme dans

toute la collection. Un terme trop fréquent dans la collection ne doit pas avoir le même impact sur la collection qu'un terme moins fréquent.

La mesure Tf-Idf est une bonne approximation de l'importance d'un terme dans un document, elle l'est particulièrement dans les corpus de documents de tailles homogènes, tels que les corpus contenant des résumés. Cette mesure a eu un succès limité dans les corpus de tailles très variables [Donna 2000].

2.3.2 Les modèles de recherche d'information

Le résultat de l'opération de l'indexation est donc un ensemble de termes (mots-clés) obtenu par la procédure du prétraitement citée précédemment. Par ailleurs, l'un des principaux objectifs du SRI est d'établir une forte correspondance entre le document du corpus et les requêtes des utilisateurs. Pour cela plusieurs modèles ont été proposés. Ces derniers sont classés en trois catégories [Sauvagnat 2005] :

Les modèles ensemblistes : considèrent les opérations de recherche d'information comme une série d'opérations logiques à effectuer sur des ensembles des termes-contenus dans un corpus de documents.

Les modèles algébriques : dans ce modèle, les documents sont représentés dans l'espace vectoriel engendré par les termes d'indexation. L'espace est de dimension N (N étant le nombre de termes d'indexation du corpus de documents).

Les modèles probabilistes : ce modèle utilise un modèle mathématique fondé sur la théorie de la probabilité. Le processus de recherche se traduit par calcul de proche en proche, du degré ou de la probabilité de pertinence d'un document au sein d'un corpus de documents.

2.3.3 Évaluation des performances des systèmes de RI

L'évaluation des systèmes de recherche d'information constitue une étape primordiale dans l'élaboration d'un modèle de recherche d'informations. En effet, elle permet de caractériser l'approche proposée et de fournir des éléments de comparaison entre approches. Les résultats obtenus par les SRI sont évalués par les mesures de rappel et de précision.

2.3.3.1 Les mesures de Rappel/ Précision

Le rappel et la précision sont deux mesures de base pour évaluer les performances des systèmes de recherche d'information. Le principe de ces deux mesures est basé

sur la connaissance à priori des documents pertinents de la collection d'une part, et d'autre part la partition de l'ensemble des documents restitués par le SRI en deux catégories : documents pertinents et documents non pertinents [Tamine 2000]. Ces deux mesures se définissent comme suit :

Le rappel : Le rappel mesure la proportion de documents pertinents restitués parmi tous les documents pertinents disponibles. Si le rappel vaut 1 c'est que les documents pertinents disponibles ont tous été restitués par le système, inversement si le rappel vaut 0 c'est qu'aucun document pertinent n'a été restitué.

La précision : mesure la proportion de documents pertinents restitués parmi tous les documents restitués. Elle mesure la capacité du système à trouver exclusivement des documents pertinents. La précision vaut 1 quand tous les documents restitués sont pertinents. Elle vaut 0 si aucun des documents restitués n'est pertinent.

La figure 2.8 illustre la partition de la collection de documents pour une requête d'une façon générale, et les taux de précision et de rappel sont donnés par les formulations suivantes :

$$Rappel = \frac{Dp, r}{Dp, r + Dp, nr} \tag{2.1}$$

$$Prcision = \frac{Dp, r}{Dp, r + Dnp, r} \tag{2.2}$$

Où :

Dnp,r : représente les documents non pertinents restitués.

Dp,r : représente les documents pertinents restitués.

Dp,nr : représente les documents pertinents non restitués.

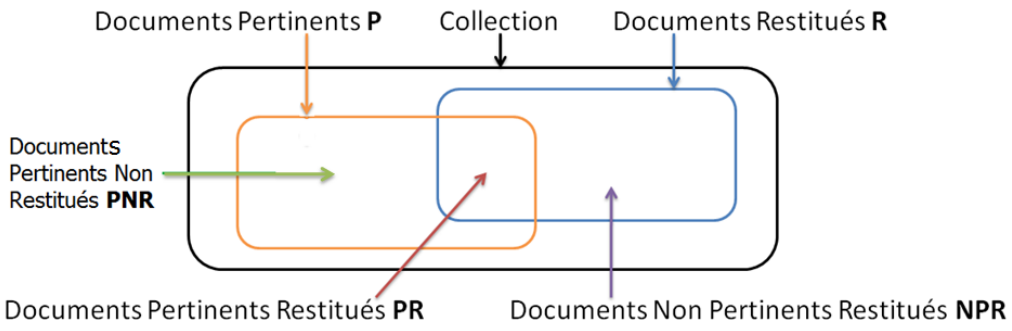

FIGURE 2.8 – Partition de la collection pour une requête [Tamine 2000]

2.3.3.2 Les mesures précision exacte et la précision moyenne

Deux mesures communément utilisées, notamment dans le cadre de TREC [1] (Text REtrieval Conference), sont la précision à x mots-clés (x=3, 4, 5,...,10 etc.) et la précision moyenne.

- La précision à x mots-clés est souvent liée à ce que l'on appelle la précision exacte ou la R-précision. La précision exacte est la précision calculée à partir des n premiers mots-clés de la liste ordonnée des mots-clés restitués selon leurs fréquences.

- La précision moyenne est la moyenne des valeurs de précision à chaque mot-clé pertinent de la liste ordonnée dans chaque document de la collection. Elle tient compte à la fois de la précision et du rappel. Elle est mesurée comme la moyenne des précisions (non interpolées) calculée pour chaque mot-clé pertinent à trouver.

2.3.3.3 Autres mesures de performance

Il existe aussi d'autres mesures de performance des SRI telles que [Campos 1] :
- Le temps de réponse acceptable : un SRI doit pouvoir fournir à l'utilisateur les réponses correspondantes à sa demande dans des temps très courts.
- Le nombre total de documents pertinents retournés : cette mesure permet d'évaluer la performance globale du système au final, en fonction ou non du nombre de documents pertinents total.
- Le rang du premier document pertinent : cette mesure a été proposée pour prendre en compte la satisfaction de l'utilisateur (comme c'est éventuellement le cas pour les décideurs des entreprises).
- La longueur de recherche : elle est égale au nombre de documents non pertinents que doit lire l'utilisateur pour avoir un certain nombre n de documents pertinents.

D'autres mesures qui combinent les scores de précision et de rappel, appelées aussi les mesures composites d'évaluation ont été définies. Rijsbergen et al. [Rijsbergen 2000] proposent une mesure générale d'efficacité, appelée F-mesure, qui est le rapport entre le rappel et la précision, qui contrôle l'importance qu'il donne à la précision et au rappel.

$$F - mesure = \frac{2 * Rappel * Prcision}{Rappel + Prcision} \tag{2.3}$$

1. http ://trec.nist.gov/

2.4 Conclusion

Nous avons présenté dans ce chapitre les principales notions et concepts des systèmes d'aide à la décision et l'analyse en ligne OLAP. Nous y avons développé les principales étapes d'un processus de recherche d'information, à savoir, la représentation ou l'indexation de l'information. Les principaux modèles existant dans la littérature ont été également présentés, ainsi que les différentes méthodes et cadres connus d'évaluation des performances des systèmes de recherche d'information.
Des approches d'agrégation pour l'analyse en ligne, sont présentées par des chercheurs comme solution afin d'adapter l'OLAP aux données complexes plus particulièrement aux données textuelles. Dans le chapitre suivant, nous dressons un état de l'art sur les approches d'agrégation textuelle dans le contexte OLAP.

Chapitre 3

Les approches d'agrégation textuelles : état de l'art

Sommaire

3.1 Introduction

Pour pouvoir faire face aux grandes quantités de données textuelles disponibles dans les entrepôts de documents, il est indispensable d'offrir aux décideurs de nouveaux outils qui vont lui permettre d'appréhender les éléments significatifs contenus dans leurs documents. Parmi ces approches, on trouve les fonctions d'agrégation qui sont un élément important de la génération des rapports.

Les fonctions d'agrégations ont rencontré des évolutions progressives et ont été adaptées et adoptées par plusieurs systèmes. La spécification de ce genre de fonctions dans les bases de données relationnelles a été une problématique active depuis la définition de l'algèbre et du calcul relationnel dans [Codd 1972]. La notion "fonctions d'agrégation" est un principe qui était encore mal compris à l'époque. Pour l'éclaircir, plusieurs travaux [Klug 1982], [Ozsoyoglu 1987], [Shoshani 1997] ont été proposés pour la spécification d'une fonction d'agrégation, au sein des bases de données statistiques basée sur la notion de "summary tables".

Plus complexe, l'agrégation dans un environnement multidimensionnel, a commencé par les propositions de [Ozsoyoglu 1987]. Cette problématique fut reprise avec

l'apparition des premiers modèles cubes OLAP [Gyssens 1997], [Agrawal 2000]. Par la suite, l'apparition de notion de hiérarchisation dans les données représentant les axes d'analyse a aussi donné lieu à de nouvelles propositions [Malinowski 2004], [Mansmann 2006], [Pedersen 2006]. Les systèmes de base de données relationnels sont accompagnés d'un ensemble de fonctions d'agrégation classiques. Il s'agit de fonctions simples qui regroupent un ensemble de valeurs statistiques en une valeur unique. Parmi ces dernières, on trouve généralement les suivantes : (1) Comptage (COUNT) : cette fonction compte le nombre d'instances dans un agrégat. (2) Maximum (MAX) : cette fonction retourne la plus grande valeur d'un agrégat (3) Minimum (MIN) : cette fonction retourne la plus petite valeur d'un agrégat, (4) Somme (SUM) : cette fonction retourne la somme numérique de l'agrégat et (5) Moyenne (AVERAGE) : cette fonction retourne la valeur moyenne d'un agrégat.

Ce jeu de fonctions a été renforcé par des fonctions statistiques pour permettre à la génération de rapports plus complets dans l'environnement des bases de données statistiques |[Shoshani 2003][Torlone 2003]. Il s'agit essentiellement de fonctions statistiques tels que le calcul de moyennes mobiles, de barycentres ou encore l'écart type. De plus, les SGBD récents (par exemple Oracle12c) offrent une interface de programmation permettant à un utilisateur de définir ses propres fonctions d'agrégation statistiques.
Diverses fonctions d'agrégation évoluées ont récemment vu le jour dans différents domaines :

- les systèmes décisionnels multidimensionnels.
- les systèmes d'information géographiques (SIG).
- les réseaux inter-véhiculaires (VANETs : Vehicle Ad-hoc Networks)

Le système décisionnel a vu la création d'un opérateur de regroupement CUBE [Boukraâ 2010][Etcheverry 2012] qui emploie de manière intensive les fonctions d'agrégation dans un environnement décisionnel. Il s'agit d'un opérateur calculant les totaux généralisés d'une sélection de données. Dans le cadre des SIG, un domaine décisionnel apparut notamment avec le SOLAP (SpatialOLAP) qui permet aux décideurs d'explorer et d'analyser une grande quantité de données géo-référencées [Viswanathan 2011]. Des fonctions spécifiques adaptées aux données géographiques virent le jour [Silva 2012]. Les données géographiques étant stockées sous la forme de données numériques telles que la surface, la température, la pressions et l'altitude ; les fonctions adaptées se chargent de permettre un regroupement de ces types de données

(avec par exemple la température moyenne, la surface moyenne,...).

Des fonctions d'agrégation issue du réseaux inter-véhiculaires (VANETs) [Dietzel 2011], [Mohanty 2012], [Dietzel 2014] ont été aussi proposées. Ces fonctions cherchent à générer dynamiquement à partir des données collectées par les véhicules au cours de leur trajet un résumé (ou agrégat) qui fournit des informations aux conducteurs (par exemple : les accidents, les embouteillages, les travaux, les freinages d'urgence, les places de stationnement disponibles, la présence de radar de police, les véhicules d'intervention d'urgence). Elles permettent à un véhicule de résumer l'ensemble des événements observés pour améliorer sa base de connaissances locales par échange avec ses voisins.

Néanmoins, ces fonctions ne sont pas utiles dans l'environnement OLAP pour les données complexes. Les analyses restent simples et se limitent aux capacités des fonctions d'agrégation disponibles. La caractéristique primordiale commune à ces propositions est qu'elles traitent seulement les données statistiques et que ne prennent pas en considération l'analyse du contenu de documents textuels. Dans le cadre de cette thèse, les fonctions d'agrégations textuelles semblent les plus intéressantes car elles donneraient une solution à notre problématique, à savoir l'analyse des 80% des données issues du système d'informations des entreprises qui reste hors de portée des systèmes OLAP traditionnels [Tournier 2008].

3.2 Approches d'agrégation textuelles

La prise en considération de documents pour une analyse en ligne OLAP a un impact sur la qualité des résultats obtenus. Les fonctions d'agrégation classique s'opèrent sur des données numériques. Toutefois, ceci est loin d'être le cas pour les données textuelles. Ainsi, des propositions ont vu le jour dans le cadre d'agrégation de données textuelles qui peuvent être regroupées en deux catégories :

- Agrégation basée sur la structure du cube de données.
- Agrégation basée sur le contenu textuel.

La première approche regroupe les spécifications des opérateurs permettant d'agréger des données en exploitant la structure multidimensionnelle du cube de données. Dans la seconde catégorie, les auteurs des travaux proposés dans cette catégorie emploient les techniques de la recherche d'informations pour l'analyse du contenu des documents.

3.2.1 Approche d'agrégation basée sur le cube de données

DocCube : Dans [Mothe 2003], les auteurs proposent une approche baptisée DocCube qui offre une nouvelle façon d'accéder au contenu textuel et qui vise à aider l'utilisateur dans les deux tâches difficiles que sont l'expression du besoin et la navigation dans l'espace d'information. Un des éléments essentiels de DocCube est le concept de l'hiérarchie qui structure l'espace d'informations. DocCube repose sur une modélisation multidimensionnelle pour proposer à l'utilisateur des visualisations globales d'informations. En d'autres termes, l'information est représentée et organisée selon différentes dimensions hiérarchiques et des faits peuvent être analysés de façon interactive.
Comme dans les systèmes OLAP, les valeurs des mesures textuelles sont recalculées en fonction du niveau d'agrégation (i.e. de généralité) auquel l'utilisateur s'intéresse. Cela permet un retour aux contenus lorsque leurs représentations synthétiques ne sont pas suffisantes.

OPAC : Dans [BenMessaoud 2004] Les auteurs ont proposé une approche pour la structuration et la classification des données multidimensionnelles baptisée OPAC (Operator for Aggregation by Clustering). Ils ont agrégé les faits d'un cube de données selon leur ordre de proximité et non plus selon l'ordre d'appartenance hiérarchique de leurs modalités dans les dimensions. Pour cela, ils ont utilisé la classification ascendante hiérarchique (CAH) en vue de construire des classes qui correspondent à de nouveaux agrégats dans le cube. Ainsi, la classification est perçue comme une technique d'agrégation dans les cubes de données.

R-Cube : Dans [Pérez 2007], les auteurs ont présenté une approche appelée R-Cube (Relevance cube). Cette dernière permet aux décideurs d'obtenir des informations agrégées pertinentes en combinant toutes les sources et les types des documents disponibles. Le R-cube se base sur la notion de la contextualisation des faits stockés avec des documents décrivant leurs circonstances. Les auteurs ont défini un contexte comme un sous-ensemble des documents textuels qui contient des informations agrégées et des connaissances pertinentes pour le processus de la prise de décision. L'utilisateur fournit d'abord les mots-clés afin de définir le contexte souhaité. Le R-cube est en suite matérialisé par des faits liés au contexte spécifié. L'approche R-Cube nécessite que l'utilisateur soit un expert du domaine de la RI pour être capable de définir le contexte de l'analyse.

TUBE : Dans [Lauw 2007] Les auteurs ont présenté une approche appelée TUBE (Text cUBE). ils ont proposé un nouveau cube de données textuelles pour faciliter

des associations qui peuvent exister entre les différentes entités présentées dans les documents textuels. Les entités peuvent être des personnes, des organisations ou des concepts. Ils ont proposé deux représentations, la première est la représentation d'un TUBE (Tex-cUBE) sous forme d'un tableau multidimensionnel où les dimensions représentent les entités où les cellules contiennent les associations entre ces entités. Ils ont proposé aussi quelques opérations pour la manipulation d'un TUBE. De plus, ils ont introduit une autre représentation sous forme d'un réseau intuitif qui recense les différents chemins qui peuvent être découverts entre les entités.

Text Cube : Dans [Lin 2008] les auteurs ont présenté une approche appelée Text Cube basée sur le modèle multidimensionnel. L'idée de base de cette approche est de faire une extension du modèle de cube de données classique afin d'agréger les documents textuels. Ils ont introduit la notion de la hiérarchie des termes qui représente les liens sémantiques entre les termes extrait du texte. Pour l'agrégation du texte, ils ont adopté une mesure de RI qui s'appelle TF*IDF. Les cellules du cube contiennent les agrégats d'un ensemble de documents associés aux dimensions. Son objectif est de trouver les K premières cellules pertinentes dans un cube de texte.

Topic Cube : Dans [Zhang 2009] les auteurs ont présenté une approche appelée Topic Cube (cube de thèmes). Leur objectif est de combiner l'OLAP avec la modélisation probabiliste des thèmes. Ils ont adopté l'approche PLSA (probabilistic latent semantics analysis) [Hofmann 1999] pour extraire les thèmes représentatifs des documents. La construction de TopicCube est basée sur le modèle multidimensionnel et utilise un composant qui s'appelle arbre des thèmes (Topic-tree). Cet arbre est défini comme un ensemble de thèmes utiles pour l'utilisateur et qui sont organisés de façon hiérarchique. Les nœuds ancêtres dans l'arbre représentent les thèmes (les agrégats), et les fils représentent les sous-thèmes. L'idée principale est d'utiliser l'arbre comme une hiérarchie dans chaque dimension du cube afin de permettre à l'utilisateur d'explorer le contenu textuel et d'appliquer les opérations de l'OLAP (drill-down, roll-up) le long des ces hiérarchies.

iNextCube : Dans [Yu 2009] Les auteurs ont présenté une approche appelée iNext Cube, comme une extension de Text Cube et Topic Cube, leur objectif est de coupler l'OLAP avec l'analyse des réseaux d'information. Cette approche a pour objectifs : le regroupement (clustring), le classement automatique et la génération des hiérarchies.

BienCube : Dans [Bringay 2010], les auteurs ont présenté un modèle de données pour représenter les cubes de textes. Ce modèle repose sur un schéma en étoile, où chaque dimension est définie sur un domaine partitionné en un ensemble de catégories

(ou niveaux de granularité). Chaque catégorie représente les valeurs associées à un niveau de granularité. L'ordre partiel défini sur les domaines des dimensions correspond à l'inclusion ensembliste des mots-clés associés aux valeurs de dimension considérées. Les auteurs ont utilisé sur une hiérarchie adaptée aux données textuelles. Dans cette hiérarchie, les nœuds sont les éléments qu'ils souhaitent agréger et les feuilles sont les descripteurs (mots-clés) de ces éléments. Pour chaque agrégation, le but est de sélectionner les descripteurs pertinents. Cette sélection dépendra du niveau et des nœuds qu'ils désirent agréger. Ils ont proposé une mesure fondée sur la mesure TF-IDF bien connue en recherche d'information.

CXTCube : Dans [Asfari 2013] Les auteurs ont présenté une approche d'agrégation basée sur un modèle vectoriel adapté, nommé CXT-cube. Le principe de leur approche est d'extraire des concepts à partir d'une ontologie de domaine. Néanmoins, les auteurs ne décrivent pas la façon dont les concepts utiles sont extraits à partir des documents d'une part et à partir de l'ontologie d'une autre part. Leur proposition s'appuie sur un modèle vectoriel et une mesure de distance et de propagation de pertinence. Ils ont représenté chaque document dans le corpus par un vecteur de concepts pondérés, puis ils ont associé ces vecteurs aux dimensions adéquates. Pour agréger les données textuelles, ils ont procédé au classement des documents par rapport aux espaces vectoriels qui les représentent.

3.2.2 Approche d'agrégation basée sur le contenu textuel

Après avoir défini les différentes approches basées sur la structure de données qui peuvent être employées afin d'agréger les données textuelles d'un corpus, nous nous intéressons dans cette section aux approches basées sur le contenu textuel pour faire l'agrégation. En effet, bien que les mots clés dans une collection de documents soient généralement pertinents, il semble nécessaire de ne sélectionner que les "meilleurs" qui représentent les agrégats d'un corpus. Les approches basées sur le contenu textuel que nous avons recensées dans la littératures peuvent être classées en trios catégories : (1) Les approches basées sur des connaissances linguistiques, (2) Les approches qui utilisent des connaissances externes et (3) Les approches basées sur des techniques statistiques.

3.2.2.1 Les approches basées sur des connaissances linguistiques

D'une manière générale, un corpus est décrit par les mots clés de ses documents. Cependant, un certain nombre de ces mots clés peut apporter du bruit et dégrader la qualité de la représentation d'un corpus. Nous présentons dans cette section des approches d'agrégation des mots-clés basées sur des informations morphosyntaxiques, nous classifions les travaux proposés en deux sous-catégories, (a) les approches basées sur les connaissances lexicales et (b) les approches basées sur les connaissances syntaxiques.

a- Les connaissances lexicales : L'utilisation de connaissances lexicales basées sur la famille d'un mot permet une sélection linguistique en fonction de l'importance des champs lexicaux d'un corpus. Par exemple, les noms sont des mots qui ont plus de crédibilité et susceptibles de décrire au mieux un concept spécifique [Poudat 2006][Lefrançois 2014].
Les auteurs des ces travaux [Poudat 2006] [Lefrançois 2014] présentent dans leurs études une méthode de classification de documents textuels de discours scientifique par l'agrégation des mots-clés. Ces derniers se fondent sur l'association entre les noms et un certain nombre de mots-clés techniques qu'ils qualifient de "caractéristiques" du discours scientifique comme des abréviations ou encore des acronymes. Dans leurs expérimentations ont montré la qualité des noms comme des agrégats, pour décrire un corpus de documents. D'autres travaux de la littérature comme ceux de [Kohomban 2007] montrent également que le choix des noms comme mots-clés est un choix pertinent. Citons également [Benamara 2007] qui montrent que les adjectifs et parfois les adverbes sont assez adaptés aux textes exprimant des opinions. Notons qu'un certain nombre de combinaisons et pondérations peuvent être utilisé avec ce type de mots-clés afin d'agréger de façon pertinente les données textuelles d'un corpus.

b- Les connaissances syntaxiques : Outre les connaissances lexicales, il est possible de sélectionner les mots-clés qui représentent mieux un corpus, par des approches utilisant les informations syntaxiques [Béchet 2009][Bechikh 2013]. Le principe est assez similaire à l'approche précédente en ne conservant uniquement pas les mots clés comme les syntagmes. Nous définissons ci-dessous ce type des mots clés. Notons que l'obtention de ce type de mots-clés nécessite le plus souvent une analyse syntaxique.
- Le syntagme : en grammaire moderne [Baten 2014], on appelle syntagme, un groupe d'éléments formant une unité par son sens et sa fonction (l'intermédiaire entre un mot est une phrase). On distingue : syntagme nominal, syntagme verbal, syntagme

prépositionnel et syntagme adjectival. La sélection de syntagmes est une extension logique de la sélection de catégories lexicales pouvant être des noms, des verbes, des adverbes. En effet, un syntagme peut se définir comme un groupe de mots formant une unité lexicale par son sens et par sa fonction. Un syntagme est formé d'un noyau qui va définir la catégorie lexicale du syntagme. Par exemple, dans le syntagme "une belle voiture", le noyau est le nom "voiture". Nous parlons alors du syntagme nominal. Notons que chaque mot constituant un syntagme est dissociable. Un groupe de mots non dissociable est appelé un mot composé, formant ainsi un mot-clés à part entière (comme par exemple "après-midi"). Un type particulier de mot composé, appelé une locution, est défini comme un mot composé contenant au moins un espace. Il s'agit souvent d'un syntagme qui est figé dont les mots ne sont plus dissociables comme la locution "base de données". Dans cet exemple, le sens du syntagme ne peut être déduit du sens de "base" et de "données" pris séparément. Une description plus complète de la notion de syntagme peut être trouvée dans [Porhiel 2013].

L'agrégation des données textuelles d'un corpus se définit comme l'ensemble des mots clés "techniques" décrivant le plus significativement le domaine du corpus. Plusieurs méthodes permettant d'extraire et d'agréger les mots-clés et qui sont basées sur des approches linguistiques ont été proposées. Citons par exemple les travaux de [Xu 2012][Rodriguez 2015] qui utilisent des méthodes linguistiques pour l'agrégation des données textuelles tel que l'approche "2-Tuple" proposé par [Rodriguez 2015]. Les syntagmes qui agrègent un corpus sont très utilisés dans le domaine de la classification de textes comme dans [Sun 2012] ou encore dans [Szymanski 2015]. Ces derniers proposent de construire des agrégats à base de syntagmes afin de classifier des sentiments.
- Les relations syntaxiques : la syntaxe peut se définir comme un ensemble de règles régissant les relations entre les mots clés d'un corpus. Ces relations de dépendance sont appelées des relations syntaxiques. Il existe plusieurs types de relations syntaxiques comme les relations "verbe-objet" ou "sujet-verbe". Une description détaillée des relations syntaxiques peut être trouvée dans [Kister 2012]. Les agrégats de type relations syntaxiques ne sont pas employés en tant que tels dans la littérature. Ils sont cependant utilisés de manière connexe à d'autres approches dans différents domaines. Shen et al dans [Shen 2015] présentent une approche construisant des agrégats fondés sur des relations syntaxiques afin de produire un système de réponse automatique aux questions.

Il existe de nombreuses autres approches permettant l'agrégation des mots clés. Notre objectif fut de présenter dans cette section des approches issues de différents

domaines sans être nécessairement exhaustif. D'autres types de méthodes permettent également l'agrégation des données textuelles comme l'utilisation des connaissances externes.

3.2.2.2 Les approches basées sur des connaissances externes

Le principe des approches d'agrégation textuelle utilisant des modèles de connaissances externes est de ne sélectionner que certains mots-clés qui représentent le mieux un domaine. Ce type d'approche utilise souvent des ressources supplémentaires afin d'effectuer une tâche sur des données textuelles. Nous ne nous intéressons pas ici à la construction de telles ressources mais à leur utilisation. Nous distinguons deux principaux types de ressources de connaissance : les thésaurus et les ontologies.

- Les thésaurus : un thésaurus est défini comme "un ensemble de termes normalisés, qui sont utilisés pour représenter le contenu thématique des documents à des fins de classement et de recherche" [Shen 2015]. En d'autres termes, un thésaurus contient un ensemble de termes d'une langue de spécialité. Ces termes sont décrits par un ensemble de relations sémantiques avec les autres synonymes. Ainsi, le thésaurus est lié à l'étude terminologique d'un domaine général ou spécialisé [Béchet 2009].
Le principe de thésaurus est d'organiser les descripteurs (terme simple) de manière conceptuelle, ils sont regroupés par affinité sémantique et sont complétés d'indication de relations [Shen 2015]. L'un des thésaurus les plus utilisés dans la littérature est Wordnet qui vise à décrire de manière générale la langue anglaise [Redkar 2015].

- Les ontologies : Ce terme issu du domaine de la philosophie désigne un "discours sur l'être en tant qu'être" [Aimé 2015]. Il a été repris en informatique pour décrire des concepts pouvant être des représentations mentales ou encore des catégories issues de la philosophie de la connaissance [Malhotra 2015]. Les concepts d'une ontologie sont définis au-delà des langues et caractérisent davantage un domaine de spécialité, et ils sont organisés hiérarchiquement [Torres 2014].
Un exemple d'application est donné dans [Ravat 2007]. Les auteurs ont proposé une fonction d'agrégation AVG-KW qui est conçue pour synthétiser un ensemble de mots-clés issus d'un vocabulaire contrôlé en un ensemble plus petit de mots-clés plus généraux. La fonction prend en entrée un ensemble de mots-clés et génère un nouvel ensemble de mots-clés agrégés. Le processus d'agrégation se base sur l'ontologie de domaine définie précédemment. Les mots-clés sont tous issus du vocabulaire contrôlé représenté par l'ontologie dont le domaine est proche de celui des documents à analyser. Pour chaque paire de mots-clés, la fonction trouve le plus petit ancêtre commun

et l'utilise pour agréger la paire des mots-clés. Cependant, lors de l'agrégation de mots clés très éloignés dans l'ontologie, il y a une très forte probabilité de retourner systématiquement le mot clés représenté par le nœud racine de l'ontologie. Afin d'éviter ce phénomène, une limite dans le processus d'agrégation doit être imposée. En effet, plus les mots-clés sont éloignés les uns des autres, plus l'agrégation se traduit par une perte de sens. Pour surmonter ce problème, la fonction emploie une distance maximale autorisée lors de l'agrégation de mots-clés.
Une autre application est donnée dans [Verma 2007]. Les auteurs proposent une méthode d'agréger de façon automatique les mots-clés de documents médicaux en se basant sur une ontologie afin de sélectionner les termes discriminants. Ils ont supposé que l'ontologie et le corpus de documents appartiennent à un même domaine. Notons pour finir qu'il existe un grand nombre de modèles de connaissances pouvant permettre d'agréger les mots-clés comme la ressource terminologique européenne IATE (InterActive Terminology for Europe)[1] et les ontologies biomédicales [Rubin 2008].

3.2.2.3 Les approches basées sur des techniques statistiques

L'agrégation des mots clés basée sur des techniques statistiques est le type d'approche le plus répandu. Elle consiste à employer des mesures et fonctions statistiques afin de donner un score de qualité à un mot -clé. Ainsi, seuls les k premiers mots-clés sont conservés afin de décrire le corpus. Les approches de cette catégorie se fondent entre autres sur la notion d'occurrence en proposant des paramètres précis pour l'agrégation des mots-clés.

Dans [Ravat 2008], les auteurs ont proposé une fonction d'agrégation permettant l'agrégation de données textuelles au sein d'un environnement OLAP, au même titre que les fonctions d'agrégation arithmétiques traditionnelles. La fonction Top-Keywords (ou TOP-KW) résume un ensemble de documents par leurs termes les plus significatifs, en employant une fonction de pondération issue de la recherche d'information : $Tf - Idf$. Le principe de la fonction d'agrégation TOP-KW est simple : il s'agit, à l'instar de la fonction $MAX - K$ qui retourne les k plus grands nombres d'un ensemble de nombres à agréger, de fournir les k mots les plus représentatifs d'un fragment de texte. Pour ce faire, elle ordonne les mots qui composent un document en fonction de leur représentativité vis-à-vis d'un ensemble de documents en utilisant la fonction $Tf - Idf$.

Dans [Wartena 2008], les auteurs ont introduit une nouvelle astuce pour faire

1. http ://iate.europa.eu/

l'agrégation des mots-clés qui représentent un corpus de documents. Leur idée se base sur le regroupement des mots-clés via l'algorithme "k-bisecting clustering" qui est un algorithme de regroupement des mots-clés à l'aide d'une hiérarchie. Le principe de cet algorithme est de regrouper tous les mots-clés dans un seul cluster, puis divise ce cluster en deux sous clusters selon la distance entre les mots-clés. Les noyaux de ces deux nouveaux clusters sont les mots les plus élongés et qui ont une distance plus élevée. Cette distance est calculée par la mesure "Jensen-Shannon divergence" [Fuglede 2004] qui mesure la similarité de la probabilité des distributions des mots clés dans un corpus. Cette procédure se répète jusqu'à ce que la distance se stabilise ou le nombre des clusters atteigne le nombre k cluster défini par l'utilisateur.

Dans [Kou 2015] les auteurs ont opté pour la méthode LSA "Latent Semantic Analysis" d'agrégation des mots clés. Cette méthode s'appuie sur l'hypothèse distributionnelle, émise par [Dumais 2004], qui est basée sur le fait que des mots qui apparaissent dans un même contexte sont sémantiquement proches. Pour appliquer LSA, le corpus est représenté sous forme matricielle. Les lignes sont relatives aux mots et les colonnes représentent les documents. Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chaque document. Deux mots proches du niveau sémantique sont représentés par des vecteurs proches. La mesure de proximité est généralement définie par le cosinus de l'angle entre les deux vecteurs.

3.3 Synthèse et comparaison des travaux

Nous proposons dans cette section une synthèse des approches d'agrégation textuelle présentée dans ce chapitre. Le tableau 3.1 récapitule les travaux que nous avons exposés et les présentent selon différents critères.
Pour synthétiser les approches recensées sur l'agrégation des données textuelles, nous avons considéré les paramètres de comparaison suivants :
Structure : l'approche se base sur la structure cube de données ou non.
Linguistique : l'approche se base sur des connaissances linguistiques (lexical ou syntaxique).
Statistique : l'approche se base sur les informations statistiques extraites du texte (tf-idf, probabilité, Association, etc).
Datamining : l'approche se base sur une des techniques de la fouille de données (Classification Hiérarchique, k-means).
Connaissance Externe : l'approche utilise des connaissances externes (ontologie,

thésaurus).

Corpus : la collection des documents textuels utilisée pour tester l'approche.

Après avoir fait le tour des approches proposées, il s'avère que les auteurs partagent, dans leur majorité, l'idée de l'intégration des méthodes et technique issu du domaine de la recherche d'informations dans les systèmes OLAP pour bien traiter et manipuler les bases multidimensionnelles de textes afin de mieux agréger le contenu textuel pour des raisons d'aide à la prise de décision.

Le choix des techniques est variable d'une proposition à l'autre. D'après l'exposé des travaux, il a été constaté que la mesure commune, la plus utilisée, est $Tf - Idf$ et ses variantes. Bien qu'elle demeure ambigüe, elle reste une mesure qui facilite l'extraction des informations les plus fréquentes dans un corpus de documents. Les approches d'agrégation linguistique ne peuvent être appliquées de façon générale. Il n'est pas possible en effet d'agréger des textes sans aucune connaissance de la langue naturelle utilisée et ses règles linguistiques.

Les fonctions d'agrégation se fondent sur des ressources externes imposant un respect de format. La ressource utilisée doit en effet respecter le même format que celui du corpus étudié. Les ressources externes sont souvent construites à partir de concepts. Nous pouvons cependant extraire les mots clés du corpus. Alors, les ressources externes, décrites par des concepts, pourront être utilisées, nous pouvons utiliser ces ressources en sélectionnant par exemple les ancêtres de mots clés pertinents extraits du corpus.

Nonobstant ces propositions, il n'en demeure pas moins que la difficulté d'agrégation sémantique des données textuelles constitue jusqu'à nos jours un verrou scientifique. Eu égard à ce qui précède, la solution de cette problématique est assujettie aux développement de nouvelles fonctions d'agrégation OLAP pour des mesures textuelles. De ce point de vue, nous proposons de fusionner les techniques issues du demain de RI, fouille de données et de la théorie des graphes afin de les coupler avec les systèmes OLAP pour constituer une piste intéressante pour traiter et manipuler les données textuelles dans un environnement multidimensionnel.

TABLE 3.1 – Comparaison entre les approches d'agrégation textuelle

connai Ref	Approche	Structure	Statistique	DataMining	Linguistique	C.Externe	Corpus
[Mothe 2003]	DocCube	Cube		Classif. H			MeSH
[BenMessaoud 2004]	OpAC	Cube		Classif. H			
[Poudat 2006]	SVM				Lexical		Ling-corpus
[Benamara 2007]	AACs		Pearson correl		Lexical		Blog posts
[Pérez 2007]	R-Cube	Cube	Relevance value				News papers
[Ravat 2007]	AVG-Kw					Ontology	
[Lauw 2007]	TUBE		Association				TKB's files
[Verma 2007]	TextSum		Frequency			Ontology	WordNet
[Lin 2008]	TextCube	Cube	Term hierarchy				Inspec/DUC
[Wartena 2008]	Topic		Jensen-Shan	K-means			Wikipedia
[Ravat 2008]	Top-Kw		Tf-Idf				
[Zhang 2009]	TopicCube	Cube	Probabilité				
[Yu 2009]	iNextCube	Cube	Probabilité				
[Béchet 2009]	SelDe				syntax.		PerformanSe
[Bringay 2010]	Biencube	Cube	Tf-Idf				
[Asfari 2013]	CXT-Cube	Cube				Ontology	
[Shen 2015]	CQA		Probabilité		Lexical		CQA Site
[Rodriguez 2015]	2-Tuple				syntax.		

3.4 Conclusion

Au cours de ce chapitre, nous avons présenté une vue d'ensemble des fonctions d'agrégation des données textuelles dans le contexte OLAP proposées dans la littérature, on a proposé une nouvelle classification de ces approches. Puis nous avons présenté un état de l'art sur ces approches où on a exposé leurs motivations, leurs idées principales et leurs résultats. Enfin, on a terminé ce chapitre par une synthèse globale qui regroupe ces approches selon leurs paramètres de fonctionnement. Au cours de notre étude bibliographique, nous avons vu que les approches qui se basent sur la combinaison entre les techniques issues du demain de RI, fouille de données et de la théorie de graphe n'ont pas beaucoup été implantés dans le contexte OLAP. Pour cela, nous allons présenter nos approches dans le chapitre suivant, où nous essayons de prendre en considération les lacunes des approches précédentes.

Chapitre 4

GOTA et TAG : deux approches pour l'agrégation textuelle

Sommaire

4.1 Introduction

La gamme des techniques disponibles dans l'OLAP traditionnel offre aux décideurs une capacité d'analyse puissante lors des manipulations des données numériques, néanmoins cette capacité est devenue inutile lorsqu'on parle des données textuelles. L'objectif majeur de nos propositions est d'aller au-delà des capacités d'analyse fournies par les opérateurs classiques et de fournir un environnement OLAP plus riche en associant l'analyse des données textuelles aux opérateurs actuels de l'OLAP. Pour

ce faire, nous proposons des fonctions d'agrégation textuelle dans le contexte OLAP pour augmenter la capacité d'analyse au profit des décideurs. Nous commençons par la définition formelle de nos approches textuelles proposées, puis nous présentons une description générale pour les deux approches qui se diffèrent selon la façon de génération des mots-clés agrégés, automatique ou non, et selon la qualité des résultats obtenus par rapport aux autres approches proposées dans la littérature.

4.2 Formalisation du problème

La modélisation conceptuelle multidimensionnelle vise à représenter les besoins utilisateurs en terme d'analyse. Cette modélisation, indépendante de toute contrainte d'implantation logique ou physique, permet d'obtenir une vision orientée décideur [Golfarelli 2002] et facilite la compréhension de l'ensemble des données mises à disposition de l'analyste [Rizzi 2006]. Notre objectif est de permettre l'analyse en ligne (OLAP) des documents par de nouvelles approches qui prennent en considération le contenu textuel des documents. Afin de refléter les besoins décisionnels, nous utilisons le modèle schéma en étoile qui permet l'analyse du contenu des documents pour faciliter les prises de décisions.

Un schéma en étoile textuel E est défini par $E = (TF, TD)$ où TF est la table des faits, et $TD = \{D_1, ..., D_n\}$ est un ensemble de dimensions. Un fait F est défini par $F = (CF, MF)$ où $CF = \{CF_1, ..., CF_q\}$ est un ensemble de clés étrangères qui associe la table du fait F aux dimensions D_i. Une dimension D est définie par $D = (AD)$ où $AD = \{AD_1, ..., AD_u\}$ est un ensemble d'attributs (paramètres et attributs faibles). Une mesure $MF = \{M_1, ..., M_n\}$ est un ensemble de mesures, elle est définie par $M = (m, FAGG)$ où m est la mesure et $FAGG = \{f_1, ..., f_x\}$ est un ensemble de fonctions d'agrégation compatible avec le contenu textuel.

L'analyse des mesures textuelles requiert des moyens d'agrégation spécifiques. Les fonctions d'agrégations actuelles n'ont pas la capacité de prendre en entrée des données textuelles. Dans l'environnement standard OLAP, seules les fonctions d'agrégation *Count* et *List* peuvent être employées sur des données non numériques et non additives. En plus des deux fonctions d'agrégation génériques (*Count* et *List*), nous proposons les fonctions d'agrégation suivantes :

GOTA : une fonction qui agrège un ensemble de mots-clés en utilisant la technique de fouille de données qui s'appelle $K - means$ avec une nouvelle distance adoptée appelée Google similarity distance [Bouakkaz 2015].

TAG : une fonction qui retourne les n mots-clés agrégés d'une mesure textuelle par l'utilisation de la théorie des graphes plus précisément l'exploitation des cycles dans un graphe pour avoir les résultats attendus [Bouakkaz 2014].

Nous emploierons nos deux fonctions d'agrégation textuelle sur un corpus de documents. Ces fonctions prennent en entrée un ensemble des mots-clés qui représentent le corpus et retournent les n mots-clés agrégés. Nous donnons à l'utilisateur la possibilité de spécifier le nombre des mots-clefs agrégés, k, pour la première approche GOTA, mais pour la deuxième approche le nombre des mots-clés agrégés est obtenu automatiquement.

En guise d'exemple, afin d'analyser les activités d'un laboratoire de recherche, un utilisateur analyse le contenu d'une collection de documents composée d'articles scientifiques. Ces articles ont un entête avec une structure générique commune. De plus, ces articles contiennent un certain nombre de méta-données : noms des auteurs, leurs affiliations (institut, pays), la date de publication, une liste de mots-clés... Le schéma en étoile textuel correspondant est présenté dans la Figure 4.1. Le modèle adopté pour analyser les documents est le même avec celui des propositions de [Aknouche 2013], [Azabou 2015]. Une fonction d'agrégation textuelle est nécessaire pour faire l'agrégation des mots-clés du corpus.

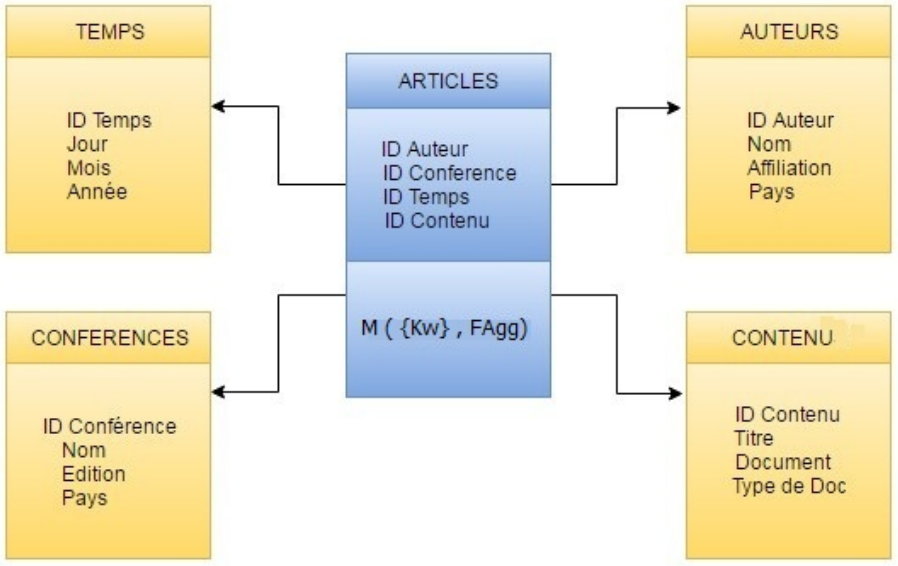

FIGURE 4.1 – Schéma en étoile textuel

Prenons l'exemple de l'analyse d'une collection d'articles scientifiques portant sur les publications des auteurs Au1, Au2 et Au3 durant la période entre 2013 - 2015.

L'analyse sélectionne les principaux mots-clefs de chaque ensemble de documents. Ainsi, les fonctions listent les principaux mots-clefs agrégés de la collection.

Dans la Figure 4.2, un utilisateur analyse les publications de l'auteur Au1 durant l'année 2015. Il affiche les résultats de l'analyse par mois (partie supérieure de la Figure 4.2). Par l'application de l'opération Roll-up et à l'aide d'une fonction d'agrégation, les mots-clés des deux publications du mois de Juin et Septembre sont agrégés dans une seule cellule pour l'année 2015 (partie inférieure de la Figure 4.2). La vision par mois dans une table dimensionnelle génère deux cellules, une correspondante aux publications du mois de Juin et une autre correspondante aux publications du mois de Septembre et la fonction d'agrégation permet de trouver le mot-clé le plus représentatif des travaux de l'auteur Au3 durant l'année 2015.

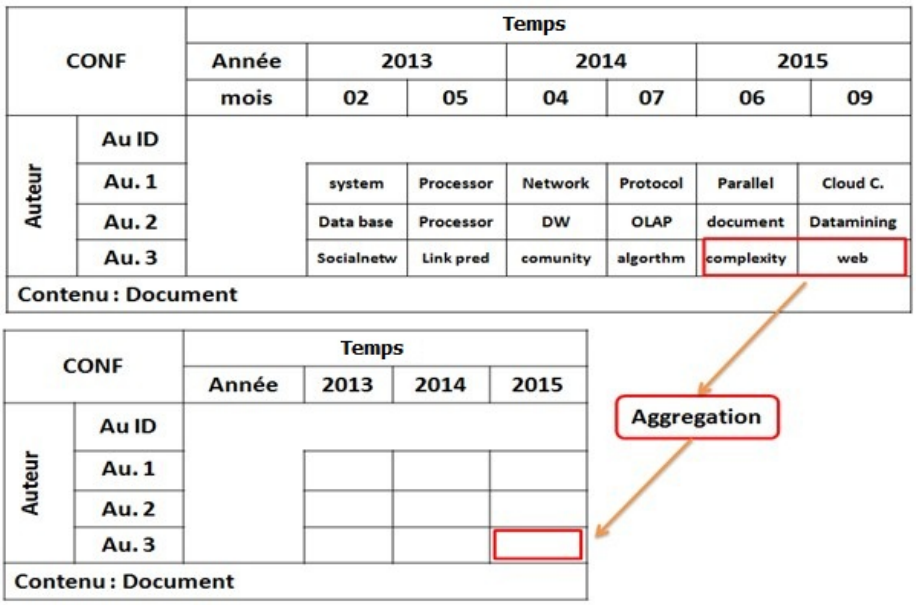

FIGURE 4.2 – Exemple d'analyse textuelle

Les fonctions d'agrégations textuelles que nous proposons sont des fonctions opérant sur des mots-clés. Notre idée est de fournir des fonctions qui s'inspirent de la moyenne mathématique pour permettre d'agréger les mots clés en d'autres mots-clés plus généraux. Par exemple, les mots clés protocole, adresse et liens regroupés en réseau.

4.3 Le couplage entre la fouille de donnée et l'OLAP

Étant donné que la taille des entrepôts de données augmente de façon exponentielle, le besoin d'approches et d'outils qui permettent d'automatiser le processus d'extraction des connaissances, devient très crucial. La technologie OLAP traditionnelle se limitant à des tâches d'exploitations de données issues de l'entrepôt, c'est donc au décideur de trouver manuellement les connaissances potentiellement contenues dans les données d'un cube en se basant sur ses expériences. À partir de ce constat, l'idée est de pousser l'OLAP vers d'autres possibilités d'analyse, et de fournir à l'utilisateur des outils automatiques pour l'aider à interpréter et utiliser les résultats des opérateurs OLAP. La puissance du domaine de la fouille de données peut apporter des réponses à ces lacunes. En effet, la fouille de données couvre une gamme de techniques et méthodes destinées à extraire des informations pertinentes (non triviales, implicites, non connues précédemment et potentiellement utiles), des associations, des contraintes ou des motifs, à partir de gros volumes de données. Une fois validée, l'information extraite devient une connaissance.

Divers travaux de recherche [BenMessaoud 2006][Jadav 2012][Fangbo 2013] se sont intéressés au couplage de la fouille de données et de l'OLAP. Le concept de fouille de données multidimensionnelles, où OLAP Mining, a été introduit par Han dans [Han 1997]. Han définit la notion de l'OLAP Mining comme un mécanisme qui intègre des tâches de fouille de données dans des systèmes décisionnels. Ce mécanisme peut s'appliquer à différents niveaux de granularité des données et à différentes parties d'un entrepôt de données. Parmi les travaux qui ont abordé le couplage de l'analyse en ligne et de la fouille de données on trouve entre autres, Palpanas [Palpanas 2000] qui a affirmé que ce couplage ouvre des directions de recherche prometteuses par l'intégration des techniques de la fouille pour manipuler les structures multidimensionnelles des données. Ceci les rendra possible de générer des connaissances à différents niveaux de granularité de l'information soit par le fourrage ou soit par l'agrégation. Par la suite, divers travaux se sont intéressés à l'exploration des cubes de données avec des algorithmes d'extraction de règles d'association [Lim 2010][Jadav 2012][Dehkordi 2013]. Cependant, la majorité de ces travaux est prévue pour les données tabulaires et numériques et n'est pas adaptées aux contenus complexes tels que les données textuelles.

Dans cette section, nous proposons une approche pour l'agrégation des données textuelles. Nous agrégeons l'ensemble des mots clés d'un corpus de documents en nous basant sur une technique de fouille de données baptisée K-means et une mesure de distance adoptée appelée "Google similarity distance".

4.4 L'utilisation de K-means pour l'agrégation textuelle

Un des algorithmes les plus connus et les plus utilisés pour le partitionnement des données est le K-means [Nashi 2010]. C'est un algorithme de classification non supervisée (clustering en anglais). Il sert à séparer un ensemble de points en k classes. Le principe de notre proposition est d'appliquer une nouvelle variante de l'algorithme K-means pour classifier les mots-clés dans des classes afin de les agréger. Notre variante de K-means ne se base pas sur la distance classique, qui est la distance euclidienne entre les individus statistiques, car cette dernière ne prend pas en considération l'aspect sémantique entre mots-clés. La nouvelle variante de k-means adopte une nouvelle distance appelée "Google similarity Distance" [Cilibrasi 2007], cette dernière est une distance sémantique proposée par Google Lab® afin de mesurer la distance entre deux termes.

Après les prétraitements des documents et l'extraction des mots-clés, la modélisation des faits permet la représentation des collections de textes par deux matrices : la matrice des fréquences $FREMAT$ (Frequency Matrix) et la matrice de distance $GoogleMAT$ (Google Distance Matrix). Cette dernière représente la distance sémantique entre les mots-clés qui est calculée à l'aide de la distance "Google Similarity Distance". Nous avons a baptisée notre approche GOTA "Google distance for Olap Textual Aggregation".

Soit D=$\{d_1, d_2, ..., d_n\}$ un ensemble des faits d'un corpus C, Kw=$\{kw_1, kw_2, ..., kw_m\}$ un ensemble des mots-clés extraits à partir de l'ensemble des documents D. On a $n = |D|$ et $m = |Kw|$. La matrice de fréquences de C, notée $FreMat$ est une matrice de n lignes et m colonnes. La ligne i de la matrice correspond à un document D et la colonne j de la matrice correspond à un mot-clé de Kw. Les éléments de cette matrice notés Fkw_jD_i correspondent aux fréquences du mot-clé Kw_j dans un document D_i. Plus formellement, $FreMat$ est telle que :

$$FreMat(i, j) = Fkw_jD_i \ \ ; i = 1 \cdots n, j = 1 \cdots m \tag{4.1}$$

Après le remplissage de la matrice des fréquences, on fait appel à une mesure de distance pour calculer la distance sémantique entre les différents mots clés obtenus. Cette mesure est appelée "Google Similarity distance". C'est un concept donné par Rudi Cilibrasi [Cilibrasi 2007] qui calcule la similarité entre un certain nombre de mots clés dans un corpus donné. Cette fonction est utilisée par le moteur de recherche

Google® pour découvrir la métrique de similitude entre les mots clés qui figurent dans les pages web. La similarité est basée sur le principe de la fréquence d'apparition d'un mot-clé dans les résultats de recherche du moteur de recherche sur Internet. Dans notre contexte chaque page web représente un document. En utilisant la formule 4.2, on peut trouver la similitude entre les mots-clés des faits.

$$GoogleDistance(x,y) = \frac{Max(logH(x), logH(y)) - logD(x,y)}{logN - min(logH(x), logH(y))} \tag{4.2}$$

Où $H(x)$ et $H(y)$ est le nombre des faits contenant les mots clés x et y respectivement. $D(x,y)$ est le nombre des documents où les deux mots clés x et y apparaissent ensemble. N est le nombre total des documents dans le corpus.

Les valeurs des distances sémantiques entre les mots clés calculées par cette formule sont stockée dans une matrice de distance appelée *GoogleMat*, cette dernière est utilisée comme paramètre d'entrée, ainsi que le nombre de classes introduites par l'utilisateur pour la fonction K-means. L'application de la méthode K-means sert à partitionner l'ensemble des mots clés sur les k classes selon les distances sémantiques obtenues entre eux. Une fois tous les mots-clés affectés chacun à une seule classe, on sélectionne les mots-clés les plus fréquents dans chaque classe. Ils seront considérés comme des agrégats qui représentent les faits. La structure de notre approche est résumée dans la Figure 4.3.

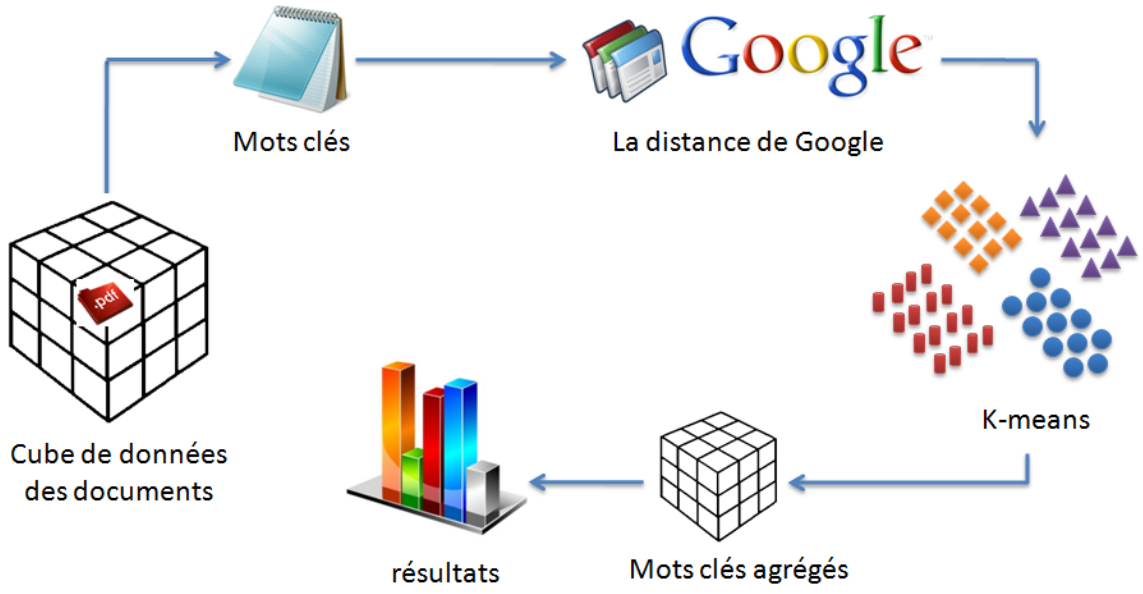

FIGURE 4.3 – La structure de l'approche GOTA

4.4.1 Algorithme GOTA

Cette section décrit l'algorithme proposé Alg 1. Tout d'abord, nous considérons que les documents initiaux sont pré traités. Les mots clés de la collection sont extraits via la fonction $ExtractKeywords(D)$. La matrice $FreMat$ correspond aux fréquences des mots-clés dans les faits. en utilisant deux fonctions, $SomFreqKw$ qui calcule la somme des fréquences de chaque mot-clé dans le corpus et $FreqKwInD$ qui calcule la fréquence de chaque mot-clé dans chaque document. La procédure K-means est appliquée pour construire les différentes classes en se basant sur la distance "Google Similarity Distance". Enfin nous sélectionnons dans chaque classe le mot-clé qui a la fréquence la plus élevée. Ces mots-clés représentent les agrégats du corpus.

4.4.2 Exemple d'application

Considérons un ensemble de documents composé de 13 articles scientifiques. L'ensemble des mots-clés des documents après les prétraitements est présenté dans le tableau 4.1 :

TABLE 4.1 – La liste des mots clés - GOTA

T1	T2	T3	T4	T5
OLAP	XML	Dmining	Query	DWHouse
T6	T7	T8	T9	T10
Document	System	Function	Cube	Network

Dans cet exemple, nous illustrons la construction des matrices FreMat et GoogleMAT.

La 1 ère étape consiste à construire la matrice des fréquences $FreMat$ (Table 4.2) :

La 2 ème étape consiste à construire la matrice des distances entre les mots-clés en se basant sur la mesure "Google similarity distance" (Table 4.3).

Puis nous appliquons l'algorithme K-means pour construire un nombre de classes données k. Dans notre exemple on va prendre le nombre de classes k=3, et les classes obtenues sont C1=T1, T4, T6, C2=T3, T8, T9 et C3=T2, T5, T7, T10, puis nous sélectionnons les mots-clés les plus fréquents dans chaque classe. Le résultat obtenu donne une liste des agrégats qui représente le corpus de test, tel que {T6=Document, T9=Cube, T10=Network}

Algorithme 1 GOTA

Entrées
Un corpus de document $D = \{D1, D2, ..., Dn\}$
Nombre des clusters NbrClus
Sorties
Une Matrice de fréquence FreMat
Une Matrice de distance GoogleMat
Ensemble des clusters $C = \{C1, C2, ..., Cn\}$
Liste des agrégats ListAgg
Début
// Extraction de l'ensemble des mots clés Kw
Kw=ExtractKeywords(D) ;
NombreDesMotsClés= |Kw| ;
// Calcul des fréquences des mots clés par documents FreMat
Pour chaque Di à Dn faire
Pour chaque Kwj à Kwm faire
FreMat(i,j) = FreqKwInD(Di,Kwj) ;
// Calcul de la somme des fréquences de chaque mot clé
Pour chaque Kwj à Kwm faire
Pour i=1 ; i <= |D|, i++ faire
SomFreqKw(kwj)
MettreàJour(VFreqKw(j)) ;
// Construction de la matrice de distance GoogleMat de Kw
GoogleMat est vide ;
Pour chaque (Kwi,Kwj) à VKw do
Calculer GoogleSimilarity (Kwi,Kwj) ;
MettreàJour (GoogleMat (i,j)) ;
// L'application de k-means
Tant que (le cluster n'est pas stable) faire
Affecter chaque mot clé kwi au cluster le plus proche ;
Calculer les nouveaux le "leader" ;
// Sélection des agrégats
Pour chaque Cluster faire
ListAgg.add(SélectionnerLeader(cluster)) ;
Fin.

TABLE 4.2 – La matrice des fréquences FreMat - approche GOTA

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
D1	10	9	22	15	9	20	15	9	28	39
D2	15	22	26	0	9	16	11	0	25	0
D3	5	15	0	15	22	0	15	0	0	0
D4	0	16	0	0	15	10	0	0	0	0
D5	16	12	2	13	16	12	0	12	2	0
D6	21	0	19	21	17	9	0	0	10	0
D7	13	0	14	0	0	15	1	0	17	0
D8	17	0	8	0	0	8	0	18	20	0
D9	22	14	0	0	14	21	0	17	0	0
D10	0	7	0	0	7	0	15	18	20	0
D11	5	18	10	5	15	15	15	18	20	0
D12	20	4	7	17	4	7	0	5	3	105
D13	1	10	11	1	10	17	0	16	10	0

4.5 L'utilisation des graphes pour l'agrégation textuelle

Au cours de la section précédente (Section 4.3), nous avons présenté notre proposition basée sur une technique de la fouille de données. Nous avons également vu un exemple d'agrégation des données textuelles par l'approche GOTA. Dans cette section, nous proposons une nouvelle approche d'agrégation automatique des données textuelles basée sur la notion des circuits dans un graphe. Cette dernière reprend la notion de co-occurrence tout en l'étendant et sur une nouvelle notion qui s'appelle l'affinité entre les mots clés. Ainsi, contrairement aux approches présentées dans l'état de l'art, nous ne pénalisons pas les termes non discriminants s'ils peuvent servir à enrichir la connaissance via leurs co-occurrences qui existent dans les documents, nous exploiterons ensuite cette connaissance pour représenter les contenus textuels en introduisant le concept "cycle d'agrégats" ou "circuit d'agrégats".
Nous verrons au cours de cette section des notions de base sur les graphes qui nous vont nous servir à la formalisation de notre approche puis on présentera les différents traitements qui sont effectués à partir d'un ensemble de textes initiaux afin d'obtenir un cycle représentant les connaissances du corpus.

TABLE 4.3 – La Matrice GoogleMat - GOTA

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	0									
T2	1,2	0								
T3	0,5	1,6	0							
T4	0,7	0,8	0,7	0						
T5	1,2	0	1,4	0,8	0					
T6	0	1,2	0,5	1	1,2	0				
T7	0,8	1,4	0,8	0,9	1	1,1	0			
T8	1	0,6	0,9	0,5	0,6	0,6	1,3	0		
T9	0,4	1,4	0,3	0,8	1,4	0,4	1	0,8	0	
T10	0,9	0,9	0,8	0,7	0,9	0,9	0,5	0,7	0,9	0

4.5.1 La conceptualisation par un graphe

La modélisation de données par un graphe est une idée assez ancienne en sciences et couvre aujourd'hui l'ensemble du champ scientifique. Les graphes ont été appliqués dans plusieurs domaines afin de bien présenter leurs données, telle que la chimie et la biologie. Ils sont d'usage courant dans les productions scientifiques. Un graphe est caractérisé par un ensemble de sommets et un ensemble de liens où chaque lien est défini par un couple de sommets [Bollobás 2013].

Les exemples de la Figure 4.4 montrent les différentes formes de liaisons entre entités, qu'il s'agisse des équipements informatiques, des emails entre des étudiants ou des routes entre les villes Algériennes.

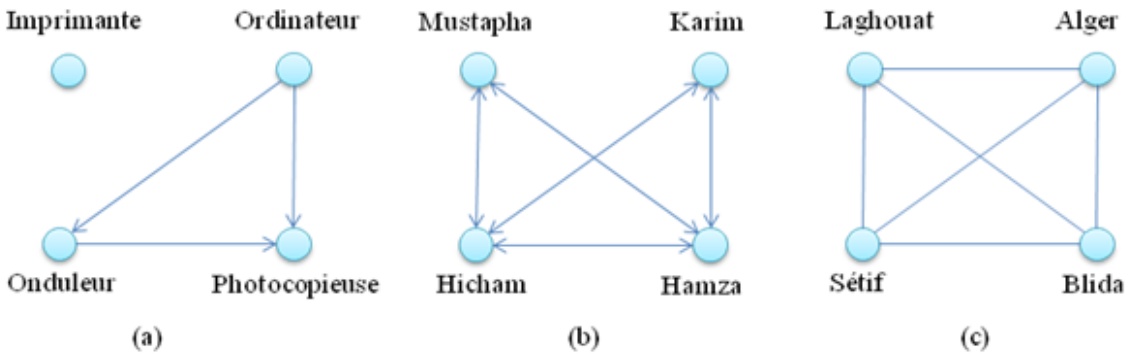

FIGURE 4.4 – Graphes orientés ou non, connexes ou non.

Le graphe (a) des équipements informatiques est orienté non connexe car le sommet imprimante est isolé. Le graphe (b) représentant les correspondances électroniques entre les étudiants est orienté et connexe. Le graphe (c) qui représente les villes est à la fois non orienté et connexe.
Dans la suite de cette thèse, nous utiliserons indifféremment les termes de réseau et de graphe.

4.5.2 Formalisation par graphe

- **Arrêtes et sommets** : la structure du graphe permet de représenter la relation et la dépendance entre plusieurs éléments, même si les éléments sont très nombreux. Les éléments qui composent un graphe sont généralement appelés sommets ou nœuds (vertices). Les liens entre ces éléments se divisent en deux types, liens orientés et liens non orientés selon les propriétés structurelles des graphes.
Dans la modélisation du graphe, les liens orientés sont appelés "arêtes" pour les distinguer de ceux non orientés qui sont appelés "arc". Un graphe est dit non orienté si tous les arcs sont symétriques, c'est-à-dire : il existe une relation binaire entre ces sommets [Bollobás 2013].

Théoriquement un graphe $G = (V, E)$ est défini par deux ensembles, l'ensemble fini $V = \{v_1, v_2, ..., v_n\}$ dont les éléments sont appelés sommets (Vertices en anglais), et par l'ensemble fini $E = \{e_1, e_2, ..., e_m\}$ dont les éléments sont appelés arêtes (Edges en anglais). Une arête e de l'ensemble E est définie par une paire non ordonnée de sommets, appelés les extrémités de e. Si l'arête e relie les sommets a et b, on dira que ces sommets sont adjacents, ou incidents avec e, ou bien que l'arête e est incidente avec les sommets a et b. On appelle ordre d'un graphe le nombre de sommets n de ce graphe.

Pour la représentation graphique, les graphes sont des modélisations topologiques sans géométrie particulière. Ils tirent leur nom du fait qu'on peut les représenter par des dessins. À chaque sommet de G, on fait correspondre un point distinct du plan et on relie les points correspondant aux extrémités de chaque arête. Il existe donc une infinité de représentations d'un graphe. Les arêtes ne sont pas forcément rectilignes. Si on peut dessiner un graphe G dans le plan sans qu'aucune arête n'en coupe une autre (les arêtes ne sont pas forcément rectilignes), on dit que G est planaire.

La représentation (a) dans la Figure 4.5 du graphe G est non planaire, par contre la représentation (b) dans la Figure 4.5 du graphe G est planaire.

On peut aussi représenter un graphe par une matrice d'adjacences. Une matrice

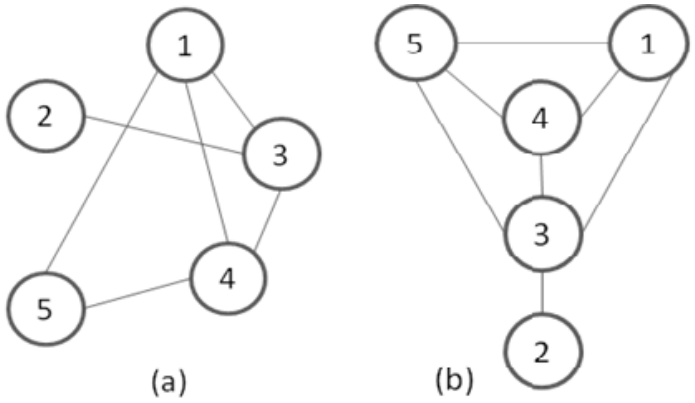

FIGURE 4.5 – Une représentation d'un graphe

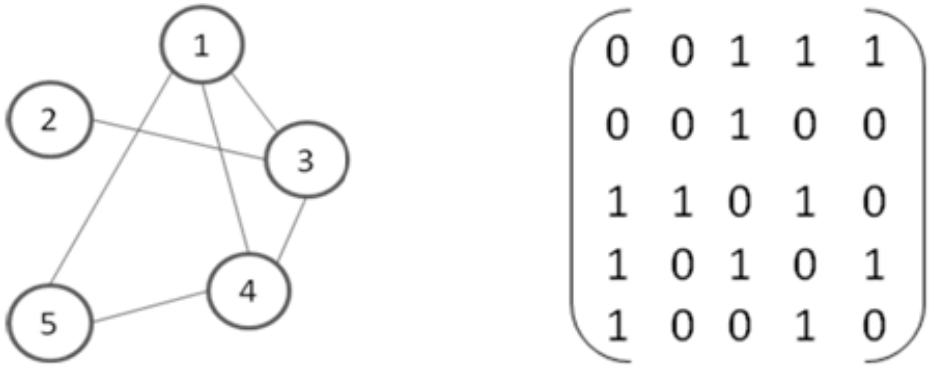

FIGURE 4.6 – La matrice d'adjacences du graphe G

$(n \times m)$ est un tableau de n lignes et m colonnes. (i, j) désigne l'intersection de la ligne i et de la colonne j. Dans une matrice d'adjacences, les lignes et les colonnes représentent les sommets du graphe. Un "1" à la position (i, j) signifie qu'une arrête part de i pour rejoindre j.

Cette matrice a plusieurs caractéristiques, elle est carrée, il n'y a que des zéros sur la diagonale. Un "1" sur la diagonale indiquerait une boucle, et une fois que l'on fixe l'ordre des sommets, il existe une matrice d'adjacence unique pour chaque graphe.

- **Connexité et parcours** : Un graphe non orienté est donc un ensemble de nœuds reliés par des arcs, on le considère comme un graphe simple s'il y a au plus un arc entre deux nœuds d'une part et s'il n'y a pas de boucle réflexive d'un nœud sur lui-même. Un graphe se caractérise par son ordre et sa taille : le nombre des ses nœuds représente l'ordre du graphe et le nombre de ses arcs en est la taille. Un graphe est connexe si à partir d'un nœud x on peut atteindre un autre nœud y par quelques sauts

consécutifs. Un nœud se caractérise par son degré qui représente le nombre des arcs qui l'ont pour extrémité. Si un nœud n'est le départ aucune arrête/arc on l'appelle un nœud isolé, c'est à dire il n'a aucun voisin [Deo 2016].

Dans un graphe non orienté on utilise le terme chaîne (chemin dans un graphe orienté). Une chaîne est la séquence alternée des nœuds et des arcs qui démarre à partir d'un nœud donnée et se clôture dans un autre nœud. La longueur d'une chaine est le nombre d'arcs parcourues. Un circuit est un chemin dont le sommet de départ et de fin est le même, on parle parfois aussi de cycle (par exemple dans l'expression graphe acyclique orienté) [Bollobás 2013]

4.6 TAG : L'agrégation textuelle par graphe

L'objectif de notre contribution est de proposer une approche d'agrégation automatique des mots-clés décrivant des faits en utilisant les graphes.

La démarche se décompose en quatre phases principales. Tout d'abord, à partir des documents initiaux, un prétraitement similaire à celui de l'approche GOTA est nécessaire pour pouvoir manipuler les contenus textuels et réduire les données à conserver lors des étapes suivantes. Dans la deuxième phase, nous modélisons les documents à l'aide de différentes matrices. Un algorithme de construction du graphe est ensuite appliqué. Enfin, l'extraction des agrégats est faite par la sélection de circuit le plus pertinent dans le graphe.

4.6.1 Modélisation du corpus de documents

Après les prétraitements des documents et l'extraction des mots clés, la modélisation de corpus de documents, se charge de la représentation des collections de textes par des matrices. Ces matrices sont la matrice des fréquences $FreMat$ (Frequency Matrix) présentée dans la section 4.4.1 qui permet de construire la matrice de co-occurrences , et la matrice d'affinité $AffiMat$, cette matrice représente le degré d'affinité entre les mots clés qui sera calculé à partir la matrice $FreMat$.

Matrice d'affinité : Soit Kw l'ensemble des mots-clés des faits, noté Kw=$\{Kw_1,..., Kw_2,...,Kw_m\}$ où m= $|Kw|$, i.e. le nombre de mots-clés. La matrice d'affinité notée $AffiMat$ est une matrice carrée de dimension m. Cette matrice est symétrique, les éléments de la matrice correspondent aux degrés d'affinité entre les mots-clés de Kw. Plus formellement, $AffiMat$ est défini comme suit :

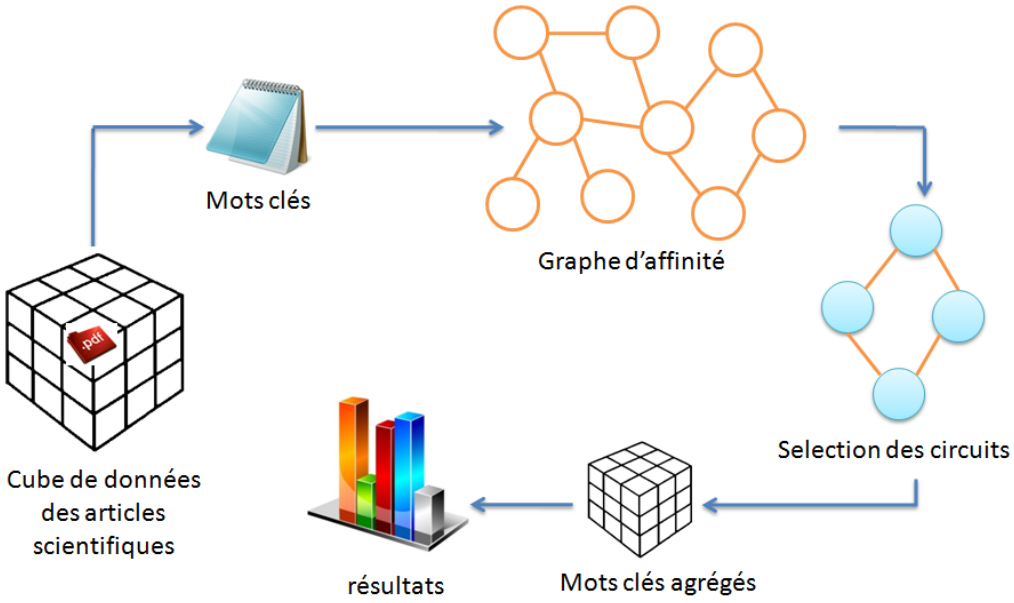

FIGURE 4.7 – La structure de l'approche TAG

$$AffiMat_{ij} = \begin{cases} \sum_{k=1}^{n} FreMat_{kj} & \texttt{if } i = j \\ \sum_{k=1}^{n} (FreMat_{ki} + TF_{kj}) & \texttt{else} \end{cases} \tag{4.3}$$

À partir de la matrice d'affinité, on peut créer le graphe d'affinité correspondant afin de trouver les circuits qui représentent les agrégats des mots-clés les plus représentatifs du corpus.

Graphe d'affinité : un graphe d'affinité (AffiGraph) est un graphe pondéré où les arrêtes sont étiquetées et dont toutes les étiquettes sont des nombres réels positifs ou nuls. Ces nombres sont les poids des liaisons entre les sommets qui représentent les mots clés. Le poids d'un circuit dans un graphe d'affinité est la somme des poids des arêtes qui constituent le circuit. La structure de notre approche est illustrée dans la figure 4.7.

4.6.2 Algorithme TAG

L'approche TAG prend en entrée les mêmes paramètres utilisés par l'approche $GOTA$ présentée dans la section 4.4.1. il utilise les fonctions $ExtractKeywords(D)$ pour extraire les mots-clés des faits qui sont regroupés dans une matrice $FreMat$ qui correspond aux fréquences des mots clés dans les documents du corpus, puis à l'aide

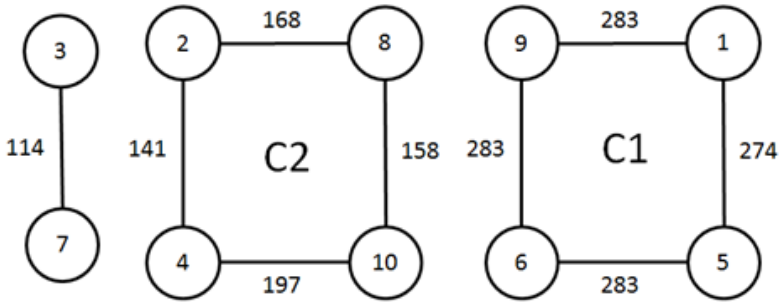

FIGURE 4.8 – Le graphe d'affinité

d'une fonction qui calcule l'affinité entre ces mots clés, une matrice carrée $AffiMat$ est obtenue. Enfin nous construisons le graphe d'affinité $AffiGraphe$ afin de trouver les mots clés les plus représentatifs à partir du circuit le plus pertinent. Ce dernier nous permet d'obtenir les agrégats représentatifs du corpus. Le pseudo-code de cette approche est présenté dans l'algorithme 2.

4.6.3 Exemple d'application

Pour appliquer l'approche TAG on va prendre le même corpus utilisé que pour l'approche GOTA, avec 13 articles scientifiques et 10 mots-clés (Table 4.4), et la même matrice des fréquences FreMat (Table 4.5). Dans cet exemple nous illustrons la construction de la matrice d'affinité $AffiMat$, puis, nous construisons le graphe d'affinité de mots-clés qui permet de trouver des groupements des mots-clés pertinents dans le corpus de documents.

TABLE 4.4 – La liste des mots clés - TAG

T1	T2	T3	T4	T5
OLAP	XML	Dmining	Query	DWHouse
T6	T7	T8	T9	T10
Document	System	Function	Cube	Network

Une fois le graphe d'affinité est construit on calcule le poids de chaque circuit pour qu'on puisse sectionner le circuit le plus pertinent.
$Moyenne$ (C1) = (283 + 274 + 283 + 283) / 3 = 280.75
$Moyenne$(C2) = (168 + 158 + 197 + 141) / 3 = 166
Dans cet exemple on va sélectionner le circuit C1 qui a un poids de 280.75 et qui regroupe les mots-clés suivants {T1=OLAP, Data warehouse, Document, Cube}

Algorithme 2 TAG

Entrées
    Un corpus de document $D = \{D_1, D_2, ..., D_n\}$
Sorties
    Une Matrice de fréquence FreMat
    Une Matrice d'affinité AffMat
    Un graphe d'affinité AffiGraph
    Un circuit d'agrégats
    Liste des agrégats ListAgg
Début
// Extraction de l'ensemble des mots clés et calcul de leurs fréquences
    Kw=ExtractKeywords(D) ;
    NombreDesMotsClés= |Kw| ;
    Pour chaque Di à Dn faire
    Pour chaque Kwj à Kwm faire
    FreMat(i,j) = FreqKwInD(Di,Kwj) ;
// Calcul de la somme des fréquences de chaque mot clé
    Pour chaque Kwj à Kwm faire
    Pour i=1 ; i <= |D|, i++ faire
    SomFreqKw(kwj) ; MettreàJour(VFreqKw(j)) ;
// Construction de la matrice d'affinité AffiMat de Kw
    Pour chaque (Kwi,Kwj) à VKw faire
    Calculer Affinité (Kwi,Kwj) ;
    MettreàJour (AffiMat (i,j)) ;
// Construction de graphe d'affinité AffiGraph
    i= randomise(NombreDesMotsClés) ; count =0 ;
    Tant que count < NombreDesMotsClés faire
        Indice-Sommet = MaxAffiMat(Kwi) ;
        Valeur-de-lien = AffiMat (kwi, Indice-Sommet) ;
        MettreàJour-AffiGraph (kwl, Indice-Sommet, Valeur-de-lien) ;
        Si (il existe un circuit) alors
        MettreàJour-Liste-des-circuits(circuit, Poids)
    Count ++ ;
// Sélection de circuit pertinent
    circuit-pertinent= MaxPoids(Liste-des-circuits) ;
    ListAgg.add(circuit-pertinent) ;
Fin.

TABLE 4.5 – La Matrice des fréquences - TAG

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
D1	10	9	22	15	9	20	15	9	28	39
D2	15	22	26	0	9	16	11	0	25	0
D3	5	15	0	15	22	0	15	0	0	0
D4	0	16	0	0	15	10	0	0	0	0
D5	16	12	2	13	16	12	0	12	2	0
D6	21	0	19	21	17	9	0	0	10	0
D7	13	0	14	0	0	15	1	0	17	0
D8	17	0	8	0	0	8	0	18	20	0
D9	22	14	0	0	14	21	0	17	0	0
D10	0	7	0	0	7	0	15	18	20	0
D11	5	18	10	5	15	15	15	18	20	0
D12	20	4	7	17	4	7	0	5	3	105
D13	1	10	11	1	10	17	0	16	10	0

TABLE 4.6 – La Matrice d'affinité - TAG

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
T1	145	198	259	180	274	200	119	125	283	175
T2	198	127	135	141	200	280	120	168	197	157
T3	259	135	119	143	155	238	114	112	242	173
T4	180	141	143	87	136	152	80	111	145	197
T5	274	200	155	136	129	283	117	171	199	157
T6	200	280	238	152	283	150	108	170	283	171
T7	119	120	114	80	117	108	67	57	132	54
T8	125	168	112	111	171	170	57	95	161	158
T9	283	197	242	145	199	283	132	161	143	175
T10	175	157	173	197	157	171	54	158	175	144

4.7 Conclusion

Dans ce chapitre, nous avons proposé deux approches originales pour l'agrégation des données textuelles dans le contexte OLAP. Notre apport concerne principalement la résolution d'un verrou scientifique à savoir : l'agrégation des données textuelles issues des faits. Pour atteindre cet objectif, nous avons proposé deux approches appelées $GOTA$ et TAG, reposant sur deux étapes principales : (1) le prétraitement du corpus et (2) l'agrégation des données textuelles. L'originalité de nos approches réside dans le fait que pour la première approche $GOTA$ on a utilisé une technique de fouille de données appelé $K-means$ avec une nouvelle distance qui s'appelle "Google Similarity Distance". Pour la deuxième approche on a exploité la richesse de la théorie des graphes par l'utilisation de la notion des circuits pour l'agrégation des données textuelles. La similarité entre nos approches est qu'elles sont des approches non supervisées et qui se lancent sans aucune préconnaissance de domaines. Pour la différence, l'approche $GOTA$ nécessite l'intervention d'un expert pour définir le nombre k des classes, en revanche ce n'est pas le cas pour TAG qui donne des agrégats comme résultats sans l'intervention d'un expert. Dans le chapitre suivant nous présentons l'étude expérimentale faite pour évaluer nos approches et mesurer leurs performances par rapport à des approches comparables.

Chapitre 5

Implémentation et validation

Sommaire

5.1 Introduction

Dans le chapitre précédent, nous avons présenté deux approches d'agrégation textuelle dans le contexte OLAP, ces approches permettent l'agrégation des mots-clés à partir des faits. Ce chapitre présente les expérimentations que nous avons réalisées. Ces expérimentations ont pour objectif d'évaluer nos approches. Dans ce sens, nous appliquons nos approches d'agrégation textuelle à deux corpus de données réels : la collection ITINNOVATION et la collection médicale OHSU-MED. Ensuite, nous comparons les résultats obtenus par nos propositions avec ceux de 5 autres approches les plus citées dans le domaine. Nous passerons ensuite à la description de ces expérimentations, nous commençons à ce niveau par la présentation des données, l'im-

plémentation et l'environnement d'évaluation, puis nous présentons et discutons les résultats.

5.2 Les corpus de données

Nous expérimentons nos approches d'agrégation textuelle dans un cadre réel. Il s'agit de la collection des articles scientifiques publiés dans les actes de la conférence ITINOVATION, entre 2007 et 2014, et la collection médicale OHSU-MED. L'utilisation de données réelles permet d'un côté d'évaluer nos propositions et d'un autre côté de positionner nos approches par rapport aux approches existantes.

5.2.1 Le corpus ITINNOVATION

La conférence annuelle ITINNOVATION est organisée chaque année sous l'égide de l'université des Emirats Arabs Unis et sous le patronage de l'organisme IEEE. Cette conférence offre un espace de débat à la communauté scientifique pour évaluer les nouvelles innovations dans le domaine de la technologie d'information et de la communication.
Un ensemble de contributions est proposé par les différents participants. Certaines contributions font appel à des approches issues du domaine de réseaux informatique et de l'Internet des objets, d'autres approches proposent des solutions à d'autres problématiques d'actualité dans les domaines de systèmes d'information, du Big Data, de l'intelligence artificielle et du traitement automatique des langues.
Le corpus ITINNOVATION est un ensemble d'articles scientifiques publiés dans les actes de la conférence entre les années 2007 et 2014, en langue anglaise ; il est disponible sur le site Web IEEE Explorer. Ce corpus contient 600 articles scientifiques. Ces articles sont écrits sous la forme standard de IEEE (deux colonnes) entre 8 à 12 pages, incluant les tables et les figures, enregistrées sous forme PDF.
Les caractéristiques globales de ce corpus sont résumées dans le tableau 5.1, et les mots-clés sont extraits et comptés à l'aide de notre outils d'extraction des mots-clés présenté dans la section 5.3 :

TABLE 5.1 – Les caractéristiques du corpus ITInnovation

Nombre des documents	Nombre des termes
600	800.000

5.2.2 Le corpus OHSUMED

Ce corpus est composé de la collection OHSUMED qui est un ensemble extrait de la base MEDLINE. Il se compose d'articles scientifiques médicaux. En général, ces textes comportent un titre et un résumé. Ces textes contiennent, en plus, des annotations manuelles qui correspondent à des catégories manuelles appelées Medical Subject Headings ou MeSH. Dans tous les textes du corpus figurent : la référence de la revue, les annotations manuelles, le titre de l'article, le type de la revue, le résumé et les noms des auteurs. Le corpus OHSUMED est constitué de 13000 documents. Les caractéristiques globales du corpus sont résumées dans la table suivante :

TABLE 5.2 – Les caractéristiques du corpus OHSUMED

Nombre des documents	Nombre des termes
13000	1.317

5.3 L'outil d'extraction des mots-clés

L'extraction des mots-clés cherche à trouver les termes représentant les principaux aspects abordés dans un document de manière précise. Il s'agit donc d'une tâche ardue qui peut être accomplie sans connaissance du domaine étudié dans le document. Dans cette dernière décennie, les méthodes et les outils d'extraction des mots-clés ont été évalués grâce aux propositions des chercheurs dans le domaine du traitement naturel du langage [Bougouin 2013].
La majorité des méthodes existantes se basent sur les techniques de recherche d'information pour extraire les mots-clés afin de réduire le nombre des mots présents dans un corpus. L'idée principale est de supprimer les mots dont on sait a priori qu'ils ne seront pas utiles. Cette étape est critique, car les mots supprimés lors de cette étape sont : soit des mots qui apparaissent le plus souvent dans un corpus, les mots grammaticaux ou les mots de liaisons, ils sont communément appelés mots vides de sens (ou stop words en anglais), soit les mots rares dans un corpus. Le problème qui se pose dans ces méthodes est que la suppression de ces mots n'est pas nécessairement justifiée : certains mots peuvent être très fréquents ou rares, mais très informatifs.

Pour surmonter ces limites, nous présentons dans le paragraphe suivant une méthode d'extraction des mots-clés à l'aide du site web Microsoft Academic Search. L'avantage de cette méthode est qu'elle est très rapide et sure, car elle repose sur

l'index de Microsoft Academic Search qui donne une crédibilité aux résultats obtenus.

5.3.1 L'index Microsoft Academic Search

Microsoft Academic Search (MAS) est un service expérimental de recherche développé par Microsoft Research pour être exploré par les chercheurs, les scientifiques et les étudiants à trouver du contenu académique, des chercheurs, des institutions et des activités. Microsoft Academic Search indexe des millions de publications académiques, il affiche également les relations clés entre les domaines de recherche en soulignant les liens essentiels qui aident à définir les axes de recherche scientifique. L'interface de site web MAS illustré dans la Figure 5.1.

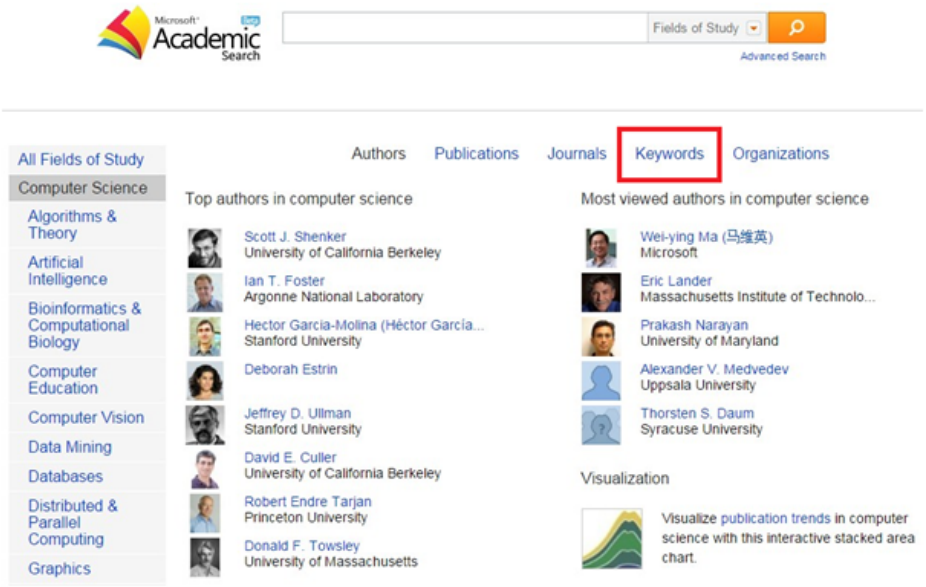

FIGURE 5.1 – L'interface du site Web : Microsoft Academic Search

5.3.2 Implémentation de l'outil d'extraction des mots-clés

Nous avons développé l'outil d'extraction des mots-clés pour qu'il sera intégré comme un module dans notre environnement d'agrégation textuelle OLAP-TAS (OLAP Textual Aggregation System). Lorsque l'on dispose d'un ensemble de textes pertinents pour un corpus donné, on cherche à trouver les mots-clés spécifiques à ce corpus. Cet outil se décompose en trois modules : le premier module extrait les termes présents dans le corpus, la deuxième étape filtre ces termes pour ne garder que ceux qui sont contenus dans l'index des mots-clés proposé par Microsoft Academic Search. La

dernière étape calcule la fréquence des mots-clés trouvés dans chaque document dans le corpus afin de créer la matrice de fréquence Document x mots-clés. La structure de notre outil est illustrée dans la Figure 5.2.

Les modules de notre outil sont implémentés en JAVA sous forme graphique pour faciliter son exploitation par d'autres utilisateurs. Notre outil est exécuté sur PC doté d'un processeur CPU i7 @ 2.20 Ghz, une RAM de 4.0 Go et un système d'exploitation Windows 7. L'interface de notre outil est illustrée dans la Figure 5.3.

Après l'application de notre outil d'extraction des mots-clés nous avons obtenu le nombre des mots-clés suivant :

TABLE 5.3 – Les caractéristiques des corpus de Test

Le corpus	Nombre des mots-clés
ITINNOVATION	2.182
OHSUMED	1.317

Les résultats obtenus par notre outil d'extraction des mots-clés sont exploités et utilisés comme paramètres d'entrée pour notre application d'agrégation textuelle afin de simuler nos propositions et les comparer aux autres approches proposées dans la littérature.

5.4 OLAP-TAS : un environnement d'agrégation textuelle

Cette section présente notre environnement d'agrégation textuelle dans le contexte OLAP (que nous avons développé) appelé OLAP-TAS (OLAP Textual aggregation System) fondé sur l'occurrence des mots-clés présents dans les documents de corpus. Nous utiliserons trois modules pour atteindre les résultats attendus : le premier module est celui du nettoyage et de préparation de corpus, le deuxième module concerne l'implémentation de nos propositions et les différentes approches qui existent dans la littérature, et le troisième module sert à assurer l'évaluation des résultats obtenus par les différentes approches implémentées.

- **Le module de nettoyage et de préparation de corpus** :fait appel à notre outil d'extraction des mots-clés et assure la lecture de la matrice obtenue des fréquences Document par mots-clés, la liste des documents du corpus et la liste des auteurs de chaque document.

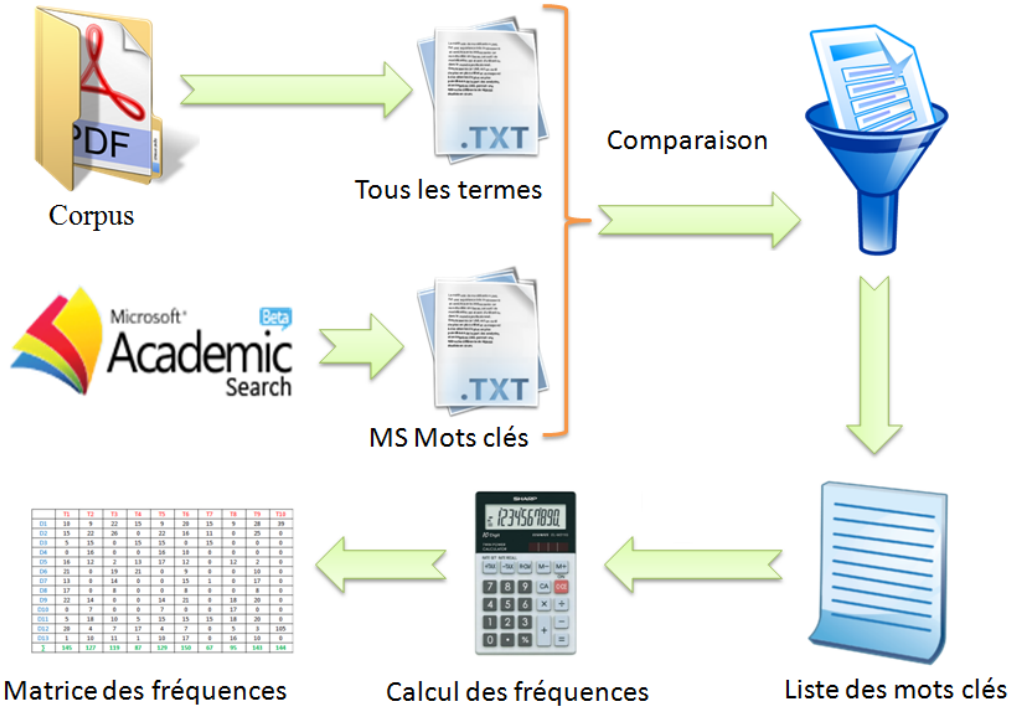

FIGURE 5.2 – La structure de l'outil d'extraction des mots-clés

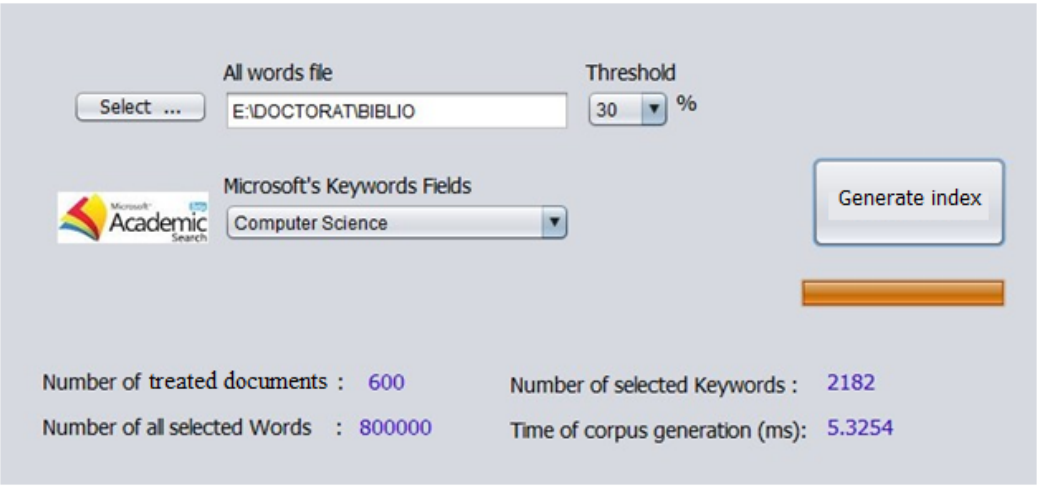

FIGURE 5.3 – L'interface de l'outil d'extraction des mots-clés

- **Le module développement** : comporte nos propositions ainsi que les différentes approches que l'on veut comparer avec les nôtres.
- **Le module d'évaluation** : dans ce module on a implémenté les différentes mesures d'évaluation des approches implémentées.

La structure de notre environnement et son interface sont illustrées dans les Figure 5.4 et la Figure 5.5 respectivement.

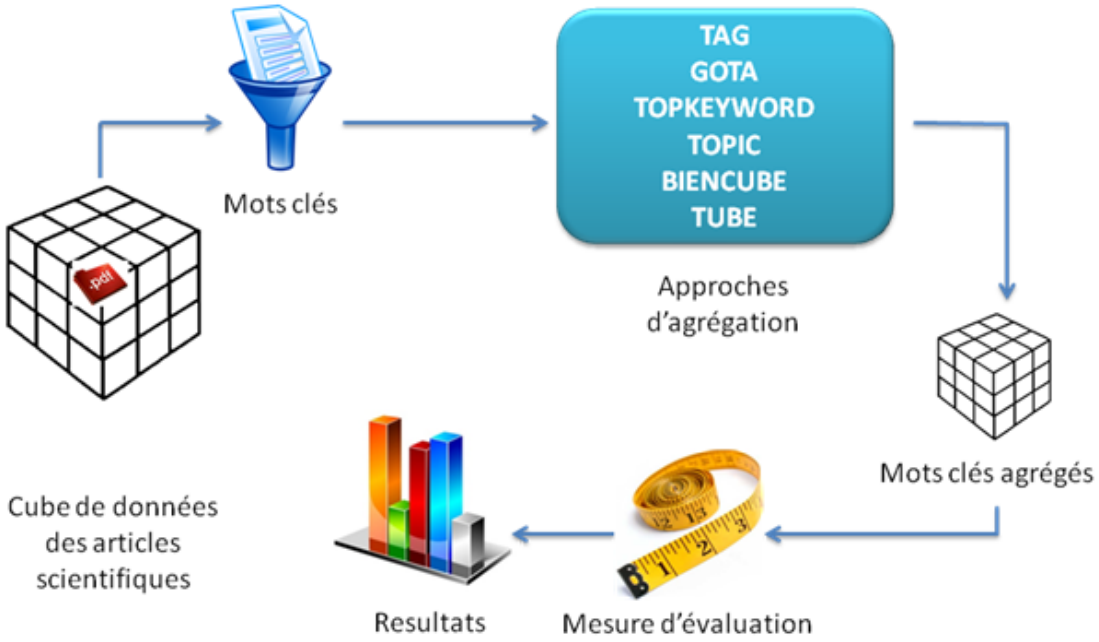

FIGURE 5.4 – l'architecture de l'environnement OLAP-TAS

Il est possible, dans notre environnement OLAP-TAS, de comparer les performances des approches implémentées. Cependant, pour pouvoir faire ces comparaisons, il est nécessaire de remplir plusieurs conditions : d'une part, le corpus servant à l'évaluation des performances doit être identique pour toutes les approches, et d'autre part, les performances doivent être évaluées avec les mêmes mesures. Nous présenterons dans la section suivante les mesures d'évaluation adoptées dans notre environnement.

5.5 Mesures d'évaluation

Pour comparer les différentes approches d'agrégation, il faut définir une mesure pour évaluer leurs performances. Malheureusement, différentes mesures sont utilisées dans la littérature, cela rend les comparaisons souvent difficiles. Le but de ce paragraphe n'est pas d'en faire une présentation exhaustive, mais d'introduire uniquement

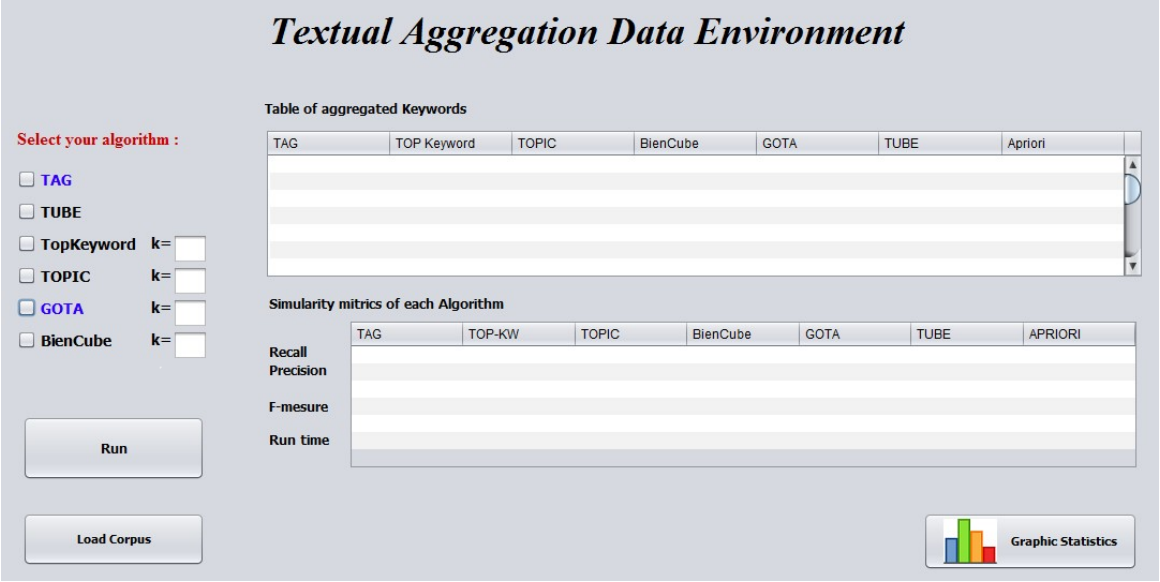

FIGURE 5.5 – l'interface de l'environnement OLAP-TAS

les plus importantes d'entre elles, d'en choisir certaines, et de mettre en évidence la difficulté de l'évaluation des performances.
Les mesures d'évaluation qui existent actuellement sont divisées en deux catégories [Chaudiron 2004].

La première catégorie regroupe des mesures d'évaluation humaine qui sont liées au niveau d'expertise de l'utilisateur dans l'analyse des résultats des approches d'agrégation. Par contre la deuxième catégorie regroupe des mesures d'évaluation formelle qui reposent sur des mesures statistiques. Entre ces deux approches, une multitude de méthodologies d'évaluation ont été proposées. Ces propositions portent essentiellement sur les caractéristiques du corpus de documents, sur les mesures d'évaluation et sur le jugement de pertinence.

5.5.1 Mesures d'évaluation humaine

L'utilisateur joue un rôle primordial dans la sélection de la meilleure approche d'agrégation textuelle selon les résultats obtenus. Dans ce genre d'évaluation, l'utilisateur a la possibilité d'exprimer sa satisfaction par rapport aux résultats. Cependant, ce jugement humain reste subjectif. En effet, deux utilisateurs différents ne jugent pas nécessairement de la même manière les agrégats fournis par les différentes approches, parce que l'interprétation de chaque utilisateur dépend en partie de ses connaissances personnelles et de ses expériences [Kompaoré 2008].

5.5.2 Mesures d'évaluation formelle

Tout l'enjeu du processus d'évaluation des résultats des approches d'agrégation textuelle est de définir la meilleure approche qui donne de bonnes valeurs afin d'assurer la satisfaction de l'utilisateur. Plusieurs mesures standard en RI ont été proposées pour évaluer les performances des approches d'agrégation. Nous nous basons sur les mesures standards d'un système de recherche d'information qui sont présentées dans le Chapitre 2. Ces mesures sont le rappel, la précision et la F-mesure qui est le rapport harmonique entre le rappel et la précision. Maximiser la F-mesure revient à trouver le meilleur compromis entre le rappel et la précision [Boubekeur 2008].

5.6 Résultats et discussion

Nous présenterons dans cette section l'ensemble des expérimentations que nous avons menées durant toute cette thèse. L'objectif de ces expérimentations est double : d'abord, de prouver l'intérêt des approches TAG et GOTA, puis de les comparer avec d'autres approches proposées dans la littérature telles que TOPKEYWORD [Ravat 2008], TOPIC [Wartena 2008], BIENCUBE [Bringay 2010] et TUBE [Lauw 2007], ces approches sont présentées dans le chapitre 3. Nous avons fait plusieurs expériences sur les deux corpus réels afin de garantir la crédibilité des résultats obtenus en variant le nombre des mots-clés agrégés par chaque approche implémentée dans notre environnement. Nous avons utilisé pour l'expérimentation notre plateforme dans laquelle nous avons implémenté six approches, dont deux font partie de la catégorie des approches qui se basent sur les mesures statistiques telles que Topkeywords et Topic. Les deux autres appartiennent à la catégorie des approches basées sur le cube de données tel que BienCube et TUBE. Les deux approches qui restent sont nos propositions GOTA et TAG, la première se base sur une variante d'une technique de fouille de données appelée k-means avec une nouvelle mesure de distance qui s'appelle "Google Similarity Distance". La deuxième proposition se base sur les fondements de la théorie des graphes pour faire l'agrégation textuelle. Les résultats obtenus par les déférentes expériences sont évalués, et les pourcentages trouvés par les mesures : le rappel, la précision, la f-mesure et le temps d'exécution sont résumés dans les figures 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 et 5.13 respectivement.

Ces expériences sont effectuées pour k (le nombre des agrégats) variant entre 3 jusqu'à 10. Notre première série des expériences menées sur le corpus ITINNOVATION permet d'obtenir des valeurs nettement meilleures pour nos propositions par

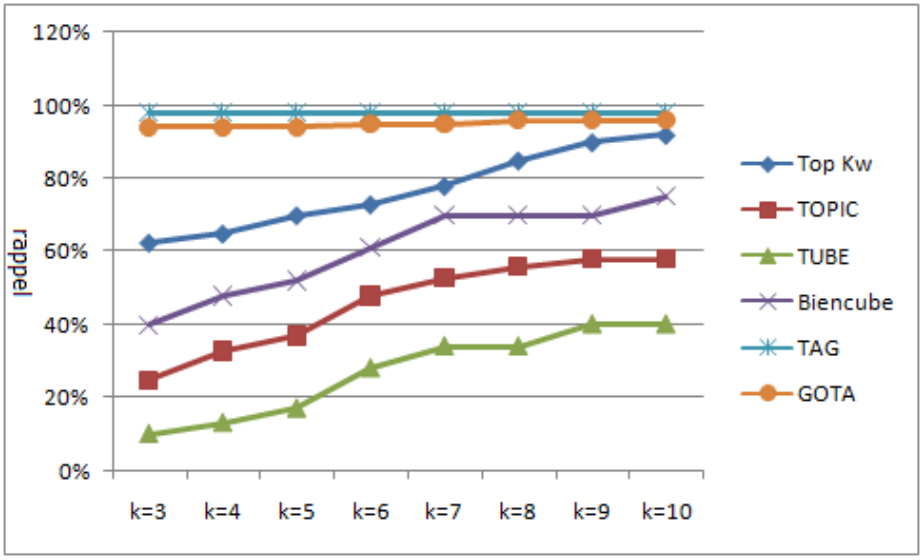

FIGURE 5.6 – Comparaison des rappels : corpus ITINNOVATION

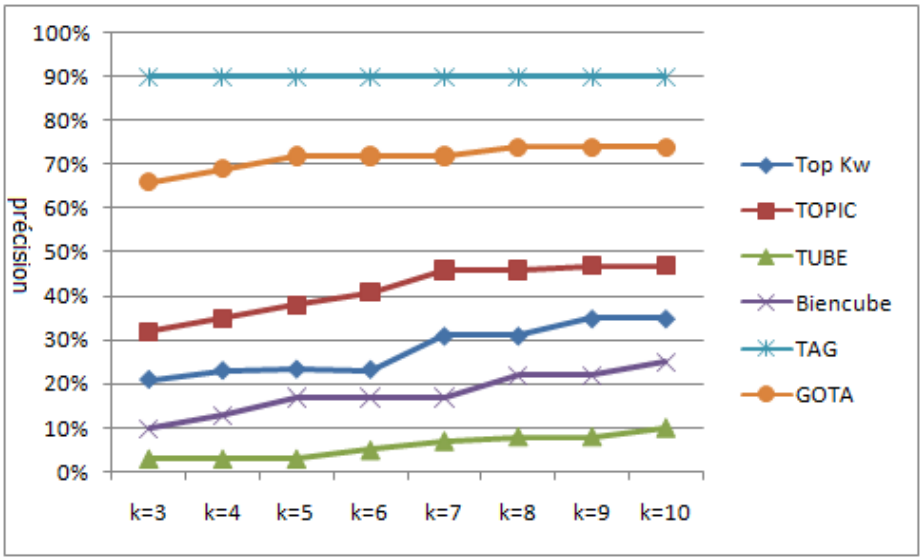

FIGURE 5.7 – Comparaison des précisions : corpus ITINNOVATION

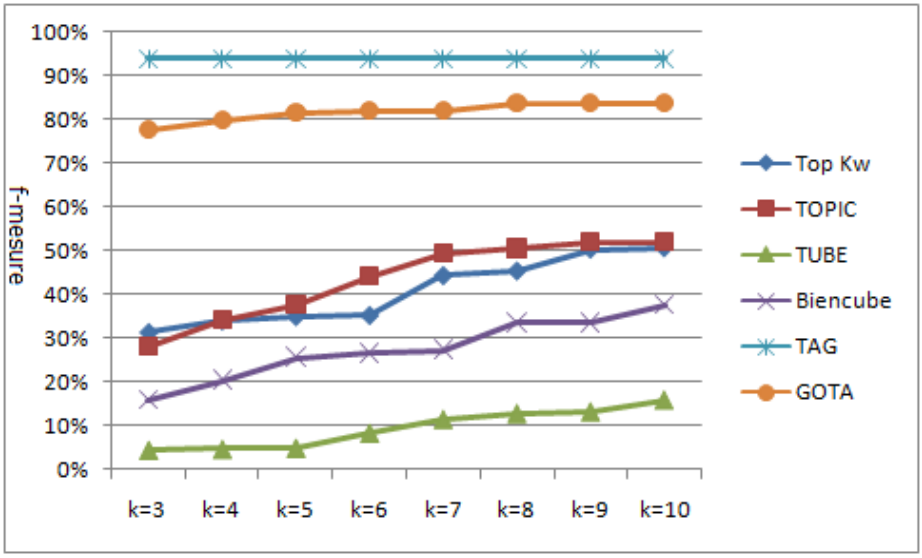

FIGURE 5.8 – Comparaison des F-mesures : corpus ITINNOVATION

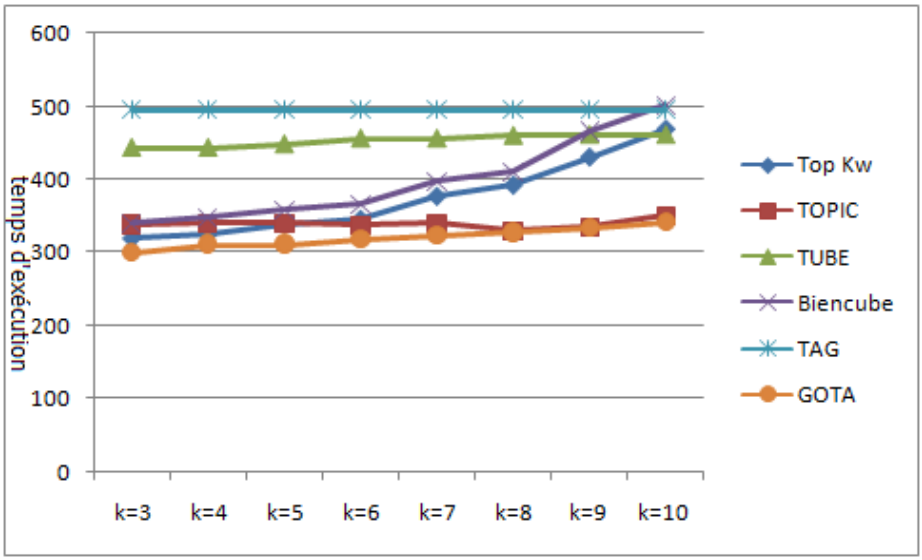

FIGURE 5.9 – Comparaison des temps d'exécution : corpus ITINNOVATION

rapport aux autres approches en matière de rappel, de précision, F-mesure et le temps d'exécution.

Pour le rappel présenté dans la Figure 5.6, notre proposition TAG donne un meilleur résultat avec 98%. Ce résultat est indépendant de k, car notre approche est automatique et elle s'exécute sans aucune intervention de la part de l'utilisateur pour déterminer le nombre des agrégats attendus. Cette indépendance est interprétée dans la figure sous forme d'une ligne droite. En revanche, les autres approches nécessitent la détermination du variable k pour sélectionner les agrégats demandés. Dans le cas où k=3, on obtient une valeur de 94% pour notre approche GOTA comparée à ceux de TopKeyword, BienCube , Topic et TUBE avec les valeurs de 63%, 40%, 25% et 10% respectivement. Pour k=10, notre proposition GOTA donne une valeur de 96% comparée à ceux de TopKeyword, BienCube , Topic et TUBE avec les valeurs de 92%, 75%, 58% et 40% respectivement.

Pour la précision présenté dans la Figure 5.7, notre approche TAG donne une valeur de 90% comparée à ceux de GOTA avec une précision de 66% et les autres approches avec les valeurs de 32%, 21%, 10% et 3% pour Topic, TopKeyword, BienCube et TUBE dans le cas où k=3, dans le cas où k=10, GOTA donne une valeur de 74% comparée à 47%, 35%, 25% et 10% pour Topic, TopKeyword, BienCube et TUBE respectivement.

La F-mesure est le rapport harmonique entre le rappel et la précision présenté dans la Figure 5.8, les résultats obtenus montrent la performance de nos propositions TAG et GOTA par rapport aux autres approches. Notre approche TAG donne une meilleure valeur de F-mesure avec 94% par rapport aux autres approches grâce aux valeurs de rappel et de précision les plus élevés. Pour notre approche GOTA la valeur de F-mesure est 84% comparée à ceux de Topic, TopKeyword, BienCube et TUBE pour les valeurs de 52%, 51%, 38% et 16% respectivement.

Les résultats obtenus par la deuxième série d'expériences sur le corpus OHSEMED présenté dans les Figures 5.10, 5.11et 5.12 montrent que nos propositions sont plus précises et confirment les résultats obtenus par la première série d'expériences sur le corpus ITINNOVATION. En effet, dans les Figures 5.10, 5.11et 5.12 notre approche TAG donne des résultats meilleurs par rapport aux autres approches en matière de rappel, de précision et de F-mesure avec les valeurs de 95%, 88% et 91% respectivement. Pour les autres approches dans le cas k=10 GOTA donne 91%, 65% et 76% en matière de rappel, de précision et de F-mesure comparés à ceux de Topic avec 49%, 41% et 45%, pour TopKeyword on a obtenu 86%, 21%, 34%, BienCube donne

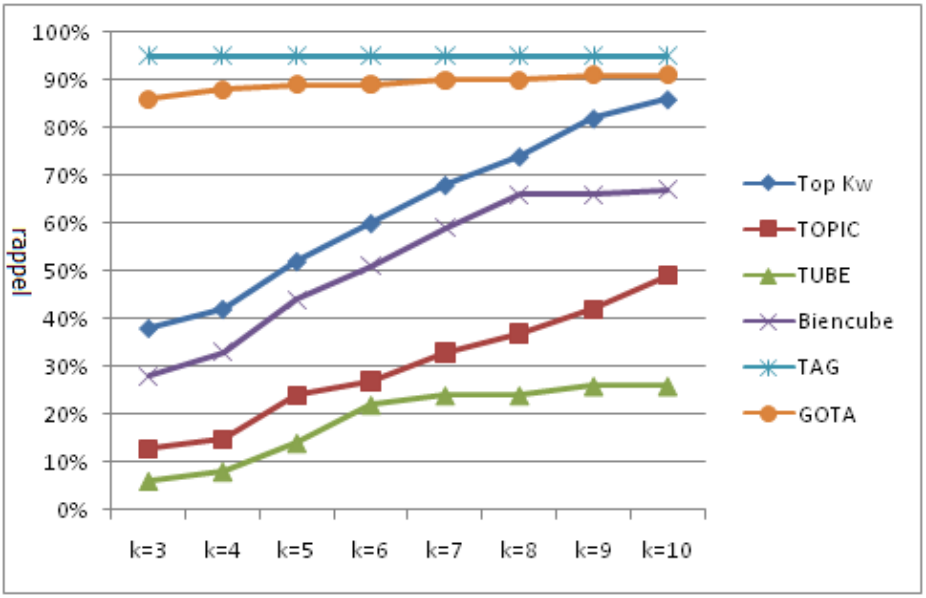

FIGURE 5.10 – Comparaison des rappels : corpus OHSUMED

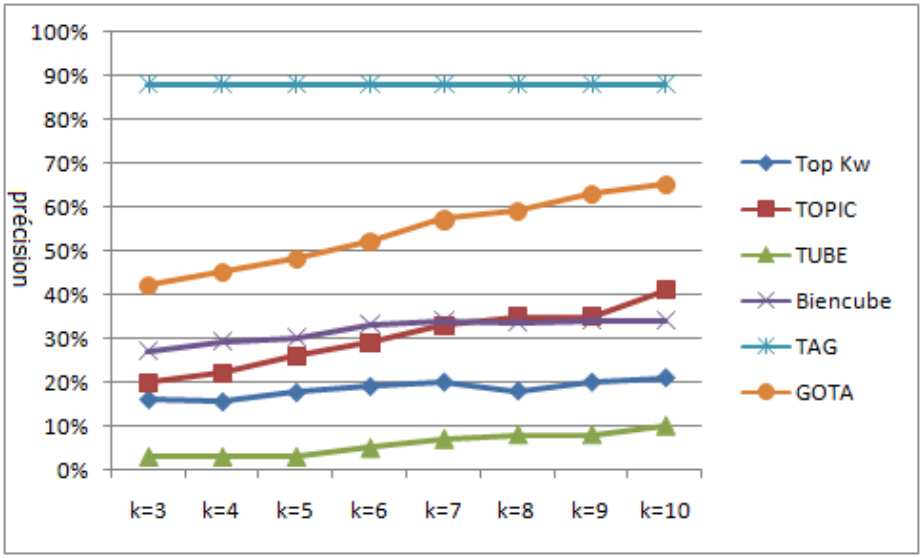

FIGURE 5.11 – Comparaison des précisions : corpus OHSUMED

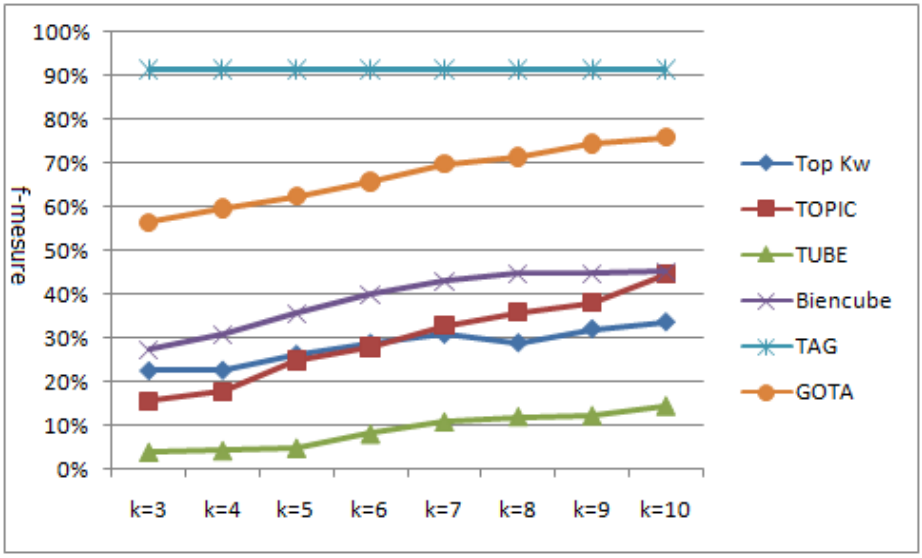

FIGURE 5.12 – Comparaison des F-mesures : corpus OHSUMED

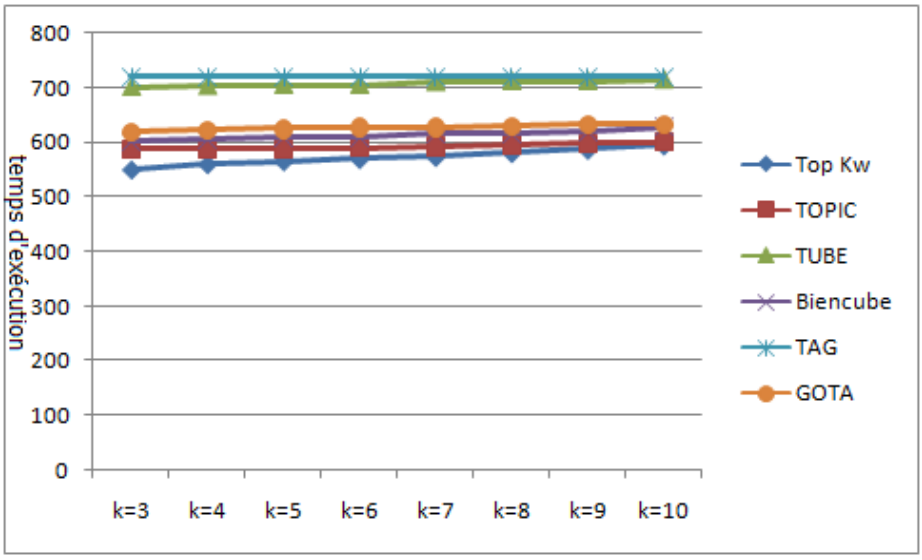

FIGURE 5.13 – Comparaison des temps d'exécution : corpus OHSUMED

67%, 34% et 45%, pour TUBE on a obtenu 26%,10% et 14% en termes de rappel, de précision et de F-mesure.

En termes de temps d'exécution les différentes approches sont proches et donnent des valeurs entre 300 à 500 ms présenté dans le Figure 5.9 pour le premier corpus et entre 550 à 750 ms présenté dans le Figure 5.13 pour le deuxième corpus. Ce temps d'exécution est négligeable par rapport à la puissance des équipements informatiques disponibles. Mais la variation du temps d'exécutions est liée à la complexité de chaque approche implémentée dans de notre plateforme d'agrégation textuelle. Pour notre proposition GOTA la complexité est égale à $O(N)$ qui est aussi celle de Top-Keywords [Ravat 2008] et Biencube [Bringay 2010]. La complixité de TAG est similaire à la complexité de TUBE qui égale à égale à $O(N_2)$ [Lauw 2007], pour Topic la complexité est $O((K\text{-}1)^*KN)$ où K est le nombre des classes générées par cette approche [Wartena 2008].

À la lecture des résultats obtenus pour chaque corpus de documents par rapport au nombre des mots-clés agrégés il ressort ce qui suit :

- Le rappel, la précision et F-mesure de nos propositions TAG et GOTA sont nettement meilleurs par rapport aux autres approches au fur et à mesure qu'on augmente le nombre des mots agrégés K pour les deux corpus.
- La taille du corpus favorise nos propositions par rapport aux autres approches, chaque fois qu'on augmente la taille du corpus la différence entre les valeurs de performance de TAG et GOTA s'améliorent de façon remarquable et l'intervalle de divergence par rapport aux valeurs des autres approches augmente de plus en plus.
- Le temps d'exécution lié à la taille du corpus, et les tests appliqués dans notre environnement confirme la performance de nos propositions. On prend en considération la complexité de chaque approche.
- Pour l'approche Topkeyword, elle donne un rappel arrivé jusqu'à 86%, mais la précision ne dépasse pas 21% dans le meilleur des cas. Ces valeurs de performance signifient que Topkeywords génère beaucoup d'agrégats bruits qui influent négativement sur la qualité des résultats obtenus. - L'avantage d'exploiter les fonctions d'agrégation est de réduire le coût d'exploration des données au profit de l'administrateur. Ce dernier préfère que les approches d'agrégation soient automatiques et qu'elles ne nécessitent aucune configuration préliminaire ou une intervention de sa part pour définir les points d'arrêt de l'opération d'agrégation. Cet avantage est remarquable dans notre approche d'agrégation textuelle TAG qui s'exécute et génère des agrégats optimums de façon automatique.

- En se fondant sur les résultats obtenus et les valeurs de performance illustrées dans les différentes Figures de 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 et 5.13, on peut constater que les approches fondées sur les données textuelles offrent une performance nettement meilleure que les approches fondées sur les propriétés de la structure de données représentée sous forme d'un cube de données.

5.7 Conclusion

Dans ce chapitre, nous avons présenté une étude expérimentale. Nous avons donné une description des corpus utilisés pour valider nos propositions et nous avons développé un outil qui permet l'extraction des mots-clés à partir de corpus de documents et de calculer leurs fréquences. Le résultat obtenu sera utilisé comme paramètre d'entré dans notre environnement d'agrégation textuel OLAP-TAS où on a implémenté 06 techniques d'agrégation, dont deux font partie de la catégorie des approches qui se basent sur les mesures statistiques. Les deux autres appartiennent à la catégorie des approches basées sur le cube de données. Les deux approches qui restent sont nos propositions GOTA et TAG. On a implémenté aussi dans notre environnement OLAP-TAS des mesures de performance formelle afin de comparer les résultats obtenus par chaque approches. Les résultats obtenus sont encourageants et nous permettent d'appliquer nos propositions à un domaine d'application pour voir leur impact réel. Nous avons choisi comme domaine d'application l'analyse des réseaux sociaux. Nous allons présenter dans le chapitre suivant les concepts de base utilisés dans l'analyse des réseaux sociaux en focalisant notre attention sur la prédiction des liens entre les membres de ces réseaux.

Chapitre 6

Domaine d'application : Les réseaux sociaux

Sommaire

6.1 Les réseaux sociaux

L'analyse des réseaux sociaux est un domaine multidisciplinaire qui a connu ces dernières années une croissance rapide notamment l'étude des interactions humaines sur internet. La définition des réseaux sociaux n'a pas encore atteint un consensus général ; Au cours de la dernière décennie, un grand public a largement adopté l'Internet. En 2004, l'organisation des Nations Unies (ONU) a compté 900 millions d'internautes, on a recensé 3,025 milliards d'internautes, soit 42% de la population mondiale fin 2015. Une des conséquences de cette vaste adoption, est la croissance de la communication humaine via Internet. Plusieurs enquêtes ont été réalisées pour étudier l'impact de

cette progression sur les relations sociales. Par exemple, "Pew Research Institute[1]", un groupe de recherche privé américain, estime que 73% des adultes au niveau mondial utilisent au moins un des médias sociaux [Madden 2013]. Une autre entreprise qui classe les sites Web selon leurs trafics appeler Alexa[2], estime que les sites de médias sociaux tels que Facebook, Twitter, et Linkedin sont tous parmi les quinze sites les plus visités, en 2015 [Alexa 2015]. La caractéristique la plus commune entre ces réseaux sociaux est qu'ils contiennent un volume d'information énorme partagé et échangé entre les utilisateurs.

Ces mines d'informations sont considérées comme des sources brutes qui nécessitent des études intensives afin de comprendre les interactions humaines. Ces études sont abordées par les sciences sociales et les sciences informatiques. La masse d'informations qui circule dans les réseaux sociaux peut être classée en deux catégories : (a) le contenu créé par les utilisateurs (document, messages, images) ; (b) les traces de leur interaction (qu'est ami avec qui ; qui a partagé le contenu de qui).

L'analyse des réseaux sociaux cherche à extraire des connaissances à partir des données publiées par les utilisateurs. Ces connaissances devraient être efficaces et pertinentes pour permettre la compréhension des comportements des internautes afin de résoudre des problèmes comme, la prédiction des liens, la détection des communautés, la sélection des dirigeants des communautés et la prédiction d'événements dans le temps.

Ce chapitre se répartira en deux sections. La première section permettra un survole des définitions des réseaux sociaux et décrira les récentes évolutions de ce genre de médias. La deuxième section de ce chapitre sera le lieu d'une esquisse succincte de l'état de l'art des travaux qui traitent la prédiction des liens dans les réseaux sociaux.

6.1.1 Définition des réseaux sociaux

Selon Michele Boyd [Boyd 2008] le terme "réseau social" est un terme générique, synonyme de la communication entre les individus par ordinateur, qui désigne l'ensemble des outils, des services et des applications qui permettent aux gens d'interagir entre eux via les technologies de l'information et de la communication. D'autres définitions sont proposées par Larry Ellison qui définit le réseau social de manière unifiée [Ellison 2013]. Ces définitions s'adaptent à tous les types de réseaux sociaux soit généraux comme Facebook et Twitter ou spécialisés comme Linkedin et ResearchGate.

1. www.pewresearch.org
2. www.alexa.com

Cette définition est la suivante : un site web de réseau social est un système tel que : **(a)** " les utilisateurs ont des profils identifiables de manière unique qui se composent de contenu fourni par l'utilisateur lui-même ou fournies par d'autres utilisateurs" ; **(b)** "Les utilisateurs peuvent produire et/ou interagir avec les flux de contenu généré par les membres inscrits dans le site" et **(c)** " Les utilisateurs peuvent établir des connexions qui peuvent être visualisées par les autres membres du site".

De manière générale les réseaux sociaux sont des sites web qui permettent l'élaboration des liens sociaux entre les membres inscrits et leur permettre de publier, d'échanger et de partager leurs informations privés, semi publique et publique. Ce genre de site web "réseau social" se base sur la technologie de Web 2.0. Cette dernière est inventée pour la première fois en 1999, est définie par Easley et Kleinberg [Easley 2010] comme un ensemble de technologies qui permettent de :

1. Rendre le Web comme une plate-forme publique.
2. La création collective de contenu.
3. Rendre les données comme des connaissances implicites.

Un des exemples des réseaux sociaux qui se basent sur la technologie du Web 2.0 est celui de ResearchGate. C'est un réseau social qui permet aux chercheurs, enseignants et étudiants de partager leurs productions scientifiques telles que les articles scientifiques, les rapports, les thèses, de poser des questions scientifiques et même de commenter les contenus des autres membres du site.

La structure et la philosophie utilisées par les réseaux sociaux leur permettent de prendre une place de plus en plus importante dans les habitudes quotidiennes des internautes. Ce genre de site Web est modélisé par un graphe et il est constitué par des entités qui représentent des individus, des groupes ou des organisations d'une part et un ensemble des relations et des liens entre ces différentes entités. Les liens dans les réseaux sociaux peuvent être interprétés selon la nature du service fourni, comme par exemple : les relations d'amitié, les collaborations scientifiques ou les relations d'affaires. La structure adoptée par les réseaux sociaux permet la progression exponentielle et permanente des réseaux : le nombre croissant des nouveaux utilisateurs inscrits nécessite une augmentation du nombre des liens créés.

Les réseaux sociaux offrent une large gamme d'avantages aux utilisateurs. Ces avantages peuvent être résumés en trois points essentiels :

1- Les réseaux sociaux offrent aux utilisateurs un espace pour s'exprimer de façon libre : la majorité des réseaux sociaux donnent aux utilisateurs l'opportunité de créer

et publier des contenus privés, semi-publiques ou publiques, d'organiser et de partager leurs profils et leurs contenus avec les autres membres du réseau social. Ces services sont devenus très populaires, ce qui a poussé les responsables de ces réseaux sociaux à chercher toujours à améliorer la qualité de leurs services par le développement de nouvelles approches. Ces approches doivent prendre en considération la croissance exponentielle du nombre d'utilisateurs. Les statistiques de l'ONU montrent que sur les 3,025 milliards d'internautes à travers le monde, 2,060 milliards sont actifs sur les réseaux sociaux, soient 68% d'internautes. Le site web Facebook par exemple a plus de 1.49 milliards d'utilisateurs, Twitter a plus de 304 millions d'utilisateurs, et ResearchGate a plus de 8 millions d'utilisateurs.

2- Les réseaux sociaux encouragent le partage de connaissances, et d'apprentissage collectif : les réseaux sociaux améliorent l'interactivité entre les individus, ce qui est influé sur la conscience collective par la diffusion des connaissances et la collaboration entre les membres. Ce service permet de réunir dans un espace virtuel des membres qui partagent entre eux des intérêts personnels ou organisationnels. LinkedIn par exemple est un réseau social composé de plus de 200 millions de professionnels qui partagent leurs compétences, CVs et leurs expériences.

3- Les réseaux sociaux offrent aux utilisateurs un espace de communication virtuel synchrone comme, par exemple le chat et la vidéoconférence afin de garantir une communication permanente entre les membres. Ils autorisent aussi les connexions sociales hétérogènes et complexes.

6.1.2 L'analyse des réseaux sociaux

L'analyse des réseaux sociaux (ARS) est l'étude approfondie de développement de ces réseaux. L'ARS permet l'analyse des comportements des individus, des groupes, d'organisations et de nombreuses autres entités connectées qui fournissent certaines connaissances. Cette analyse se base sur un ensemble de mesures et d'approches pour comprendre et surmonter les défis rencontrés dans les réseaux sociaux, tels que les techniques issues de la théorie des graphes et la fouille de graphes (graph mining) [**?**].

Les sommets ou les nœuds d'un réseau social représente des personnes et des groupes tandis que les liens reflètent les types des relations ou les flux de données entre les nœuds. On applique l'ARS pour faire une analyse visuelle et formelle afin de comprendre l'évolution dynamique de réseau social, de découvrir les schémas de communication entres les entités de réseau, et de comprendre les relations humaines en dehors des espaces virtuels.

6.1.2.1 Les tâches de l'analyse des réseaux sociaux

Les réseaux sociaux sont des réseaux très dynamiques, ils changent leurs structures de façon très rapide et permanente. Plusieurs nœuds s'ajoutent périodiquement et cela nécessite la création/suppression de plusieurs liens dans le réseau. Ces relations et liens changent selon les comportements des nœuds et selon l'évolution du réseau.

Les principales tâches de l'ARS sont la prédiction des liens entre les nœuds qui représentent soit des personnes, des groupes ou des organisations, le marketing viral, la détection des communautés, la détermination des responsables des communautés, la détermination des hiérarchies sociales implicites, et la prédiction des événements dans le temps. Il existe de nombreuses mesures pour établir des liens entre les nœuds comme, par exemple, le degré de centralité, le degré de proximité, le coefficient de clustring et le degré de cohésion. La prédiction des liens consiste à chercher les liaisons réelles ou sémantiques non connues ou qui ont une probabilité d'apparition dans un futur proche. Le Marketing viral est l'exploitation des connectivités sociales des utilisateurs pour propager des connaissances de la consommation de tel produit. La détection des communautés permet de trouver les agrégats sociaux qui émergent lorsqu'un nombre suffisant d'entités actives dans le réseau mènent des interactions. Ces interactions peuvent déboucher sur des interactions physiques (événements, manifestations) ou de forts investissements émotionnels.

L'ARS a de nombreuses applications, elle donne aux responsables la possibilité de comprendre la façon dont les informations se propagent, de comprendre et de suivre le comportement d'un utilisateur ou une communauté dans un réseau social comme par exemple dans le e-commerce, L'ARS donne la possibilité aux annonceurs de cibler leurs clients, dans la sécurité, elle aide à déterminer les entités influentes, à suivre les terroristes, à trouver leurs localisations et prédire leurs actions. Elle permet aussi aux responsables d'avoir une idée sur les tendances politiques et sociales pour un utilisateur ou une communauté d'utilisateurs du réseau social. Dans notre contexte nous nous intéressons à la résolution de problème de la prédiction des liens dans les réseaux sociaux.

6.2 La prédiction des liens dans les réseaux sociaux

Les réseaux sociaux se caractérisent par les différents liens qui relient les nœuds. Ces liens sont définis par l'intermédiaire des informations disponibles au niveaux des entités. Ces informations ne sont exploitées que partiellement ce qui influe négati-

vement sur la compréhension de l'évolution du réseau social. Pour comprendre cette évolution, les méthodes de prédiction sont plus que nécessaires. La prédiction des liens consiste à prédire un lien future entre deux entités.

6.2.1 Définition du problème

Étant donné une capture de réseau social (des nœuds et de liens) au temps t, nous avons besoin de prévoir avec précision les liens qui seront ajoutés au réseau au cours de l'intervalle de temps t à un moment futur $t + 1$. En effet, la prédiction des liens se concentre sur le problème de savoir comment découvrir des liens inconnus afin de comprendre et à anticiper l'évolution d'un réseau social. Par exemple, considérons un réseau social des co-auteurs des publications scientifiques, il y a différentes raisons pour que deux chercheurs qui travail sur des thèmes similaires et qui n'ont jamais écrit un article ensemble va le faire dans les prochaines années à travers une coopération. Ces interactions sont très difficiles à prédire, mais en étudiant les caractéristiques du réseau, nous pouvons prédire les liens possibles qu'ils vont se former. Notre objectif est de rendre cette idée intuitive réalisable, et de comprendre quelles mesures de proximité permettent de concrétiser ces prédictions de façon précise.

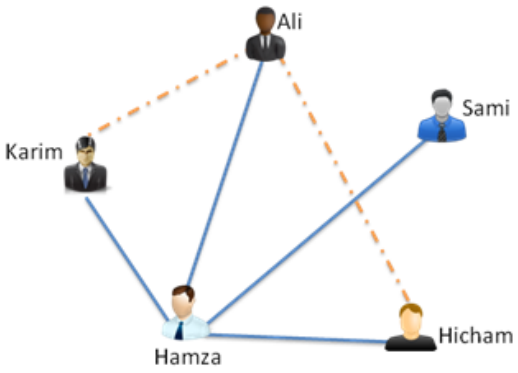

FIGURE 6.1 – Un exemple pour la prédiction des liens

La prédiction des liens traite également le problème de l'obtention de liens manquants à partir d'un réseau connu, dans un certain domaine. Il implique la prédiction des liens supplémentaires qui ne sont pas directement visibles à l'heure actuelle, ou susceptibles d'exister dans le réseau basé sur les données observables. Ce problème s'explique par un simple réseau social entre cinq personnes illustré dans la Figure 6.1, dans lequel des liens présentés par des lignes continues indiquent les interactions déjà

existantes au moment t, et les lignes pointillées indiquent les futures interactions qui vont apparaitre au cours de l'intervalle de temps $[t, t']$.

Pour résoudre le problème de la prédiction des liens, il faut déterminer et comprendre la façon d'élaboration des liens entre toutes les paires de nœuds. En effet, les probabilités d'apparition des liens manquants peuvent être mesurées par les similarités entre les nœuds ou les liens communs partagés entre les nœuds. Le schéma présenté dans la figure 6.2 montre une vue générale sur les deux familles des solutions de prédiction des liens.

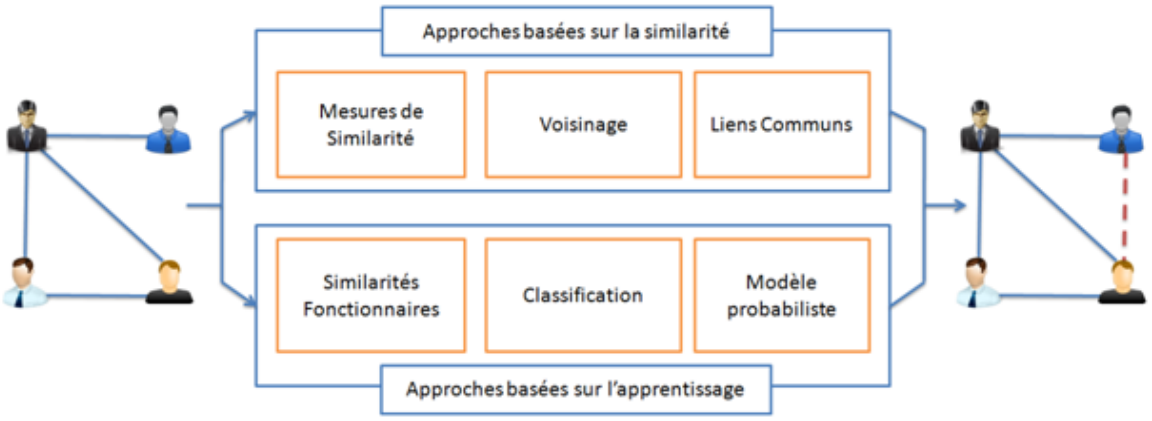

FIGURE 6.2 – Catégorie des solutions de problème de la prédiction des liens

Le principe des approches fondées sur les similarités est de calculer les similarités entre les paires de nœuds non connectés dans un réseau social. A chaque paire de nœuds (x, y) est attribué un score. Un score élevé signifie que la probabilité d'avoir un lien entre x et y dans le futur est très élevée. Les nouveaux liens obtenus seront classés dans une liste par ordre décroissant selon leur score. Les liens en haut de la liste sont donc les plus susceptibles d'apparaître.

L'idée des approches fondées sur l'apprentissage automatique est de traiter le problème de la prédiction des liens comme un problème de classification binaire [Lee-Hoon 2006]. Chaque paire de nœuds non connectés correspond à une fonction décrivant les nœuds et étiquette les liens. S'il existe un lien potentiel reliant une paire de nœuds, cette paire est étiquetée comme étant positive, sinon elle est négative. La Figure 6.3 présente une autre vision plus détaillée des techniques de prédiction des liens.

Toutes les techniques de prédiction de lien sont génériques et peuvent être utilisées pour résoudre des problèmes de prédiction de lien dans les réseaux sociaux.

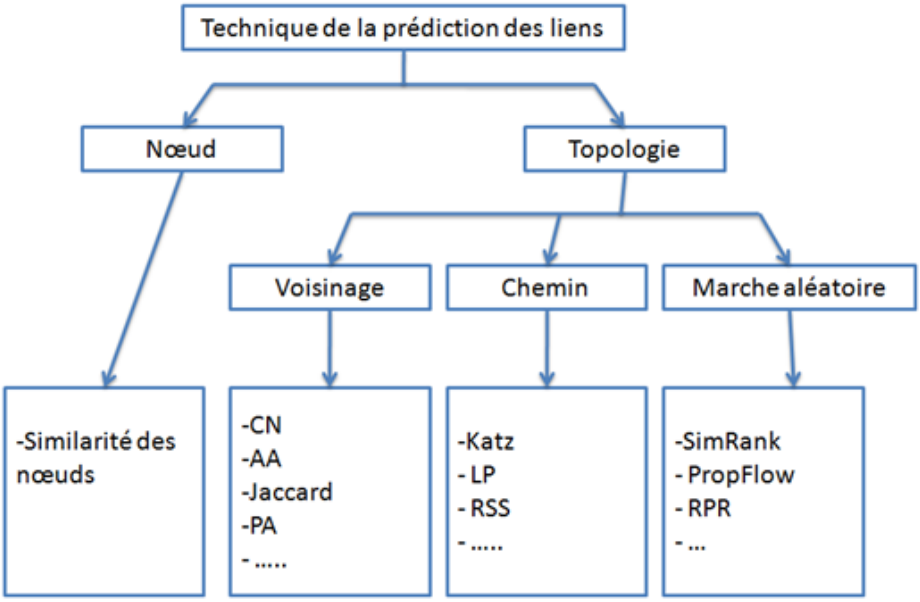

FIGURE 6.3 – Les catégories des techniques de prédiction des liens

6.2.2 État de l'art sur les méthodes de la prédiction des liens

Soit un réseau social $G<V, E>$, où V représente l'ensemble des nœuds du réseau et E représente l'ensemble des liens entre ces nœuds dans un instant t. supposant qu'on a une capture de réseau social dans un temps donné. Nous choisissions 3 temps $t_0< t_1< t_2$. Nous appliquons les approches de la prédiction des liens sur la capture du réseau social prise dans l'intervalle $[t_0,t_1]$. Les résultats obtenus de la prédiction des nouveaux liens sont comparés avec les liens réels apparus dans la nouvelle capture de réseau social prise entre $[t_1,t_2]$. Nous appelons l'intervalle $[t_0,t_1]$ l'intervalle d'apprentissage et l'autre intervalle $[t_1,t_2]$ l'intervalle du test.

La majorité des approches de base de prédiction des liens sont basées sur la similarité entre les nœuds par le calcul de la distance entre eux [Lü 2011]. Cette distance représente le nombre des liens entre deux nœuds x, y dans le réseau. Autrement, deux nœuds sont similaires si et seulement si, ils ont le chemin le plus court entre eux dans le réseau. Les approches de la prédiction des liens sont classées en deux catégories principales : (1) les approches basées sur le voisinage des nœuds et (2) les approches basées sur les chemins.

6.2.2.1 Approches basées sur le voisinage des nœuds

Le voisinage des nœuds signifie que deux nœuds sont liés entre eux. C'est une technique simple qui consiste à traverser seulement les chemins de longueur de taille 2. Pour chaque nœud A, on vérifie les voisins du voisin direct, puis on compte leurs similarités avec A. on ne prend pas en considération que les caractéristiques locales du réseau, en se concentrant principalement sur la structure des nœuds (c'est-à-dire il est fondé sur le nombre d'amis communs partagés entre deux nœuds). Parmi ces approches on peut citer :

- **Les voisins communs (VC)** : L'approche des voisins communs (Common Neighbors) fournit une mesure de similarité entre les nœuds, en calculant l'intersection entre plusieurs nœuds voisins pour prédire des liens futurs [Kumar 2008]. La formule utilisée par l'approche des voisins communs (VC) est définie comme suit :

$$VC(x, y) = \Gamma(x) \cap \Gamma(y) \tag{6.1}$$

Cette mesure est fondée sur l'idée que deux nœuds x et y ont une forte probabilité de se connecter s'ils partagent un autre nœud voisin z. Plus le nombre de voisins partagés augmente, plus la probabilité devient élevée.

- **L'indice de Sorensen (SI)** : Cette mesure est basée le nombre des voisins communs, il mesure la présence ou l'absence des voisins communs. L'indice de Sorensen est défini comme suit :

$$SI(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|} \tag{6.2}$$

- **Coefficient de Jaccard (CJ)** : Le Coefficient de Jaccard est une variante normalisée de l'approche des voisins communs [Liben-Nowell 2007], [Gupta 2015]. ce coefficient définit la probabilité qu'un voisin commun entre deux nœuds x et y soit sélectionné au hasard à partir de l'union des ensembles des voisins de x et de y. Le coefficient Jaccard est défini comme suit :

$$CJ(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{6.3}$$

- **La similarité cosinus (SC)** : Cette mesure est fréquemment utilisée en tant que mesure de ressemblance entre deux nœuds [Leicht 2006]. Il pourra s'agir de comparer les caractéristiques issues d'un nœud dans une optique de classification. La similarité cosinus est définie comme suit :

$$SC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| * |\Gamma(y)|}} \tag{6.4}$$

- **Adamic / Adar (AA)** : C'est une mesure qui compare le nombre des attributs communs entre deux nœuds [Adamic 2003]. Elle est proposée la première fois pour comparer deux pages web selon un ensemble des caractéristiques (z). elle est définit sous la forme suivante :

$$AA(x,y) = \frac{1}{log(frequence(z))} \tag{6.5}$$

Pour la prédiction des liens, cette mesure est adaptée par le fait de considérer les voisins communs comme des caractéristiques partagées entre deux nœuds [Liben-Nowell 2007]. La formule Adamic /adar sera calculée de la forme suivante :

$$AdamicAdar(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|} \tag{6.6}$$

- **L'attachement préférentiel (AP)** : L'approche d'attachement préférentiel est basée sur l'hypothèse qu'un nœud x aura de nouveaux voisins plus vite qu'un nœud y, si seulement si, y a moins de voisins que x, donc, la probabilité qu'un nœud forme un nouveau lien avec d'autre nœud varie selon le nombre de ses voisins [Capocci 2006]. La probabilité d'avoir un lien entre deux nœuds, fondé sur l'approche d'attachement préférentiel est mesurée en multipliant le nombre de leurs voisins [Papadimitriou 2011]. La formule utilisée par l'approche d'attachement préférentiel est définie comme suit :

$$PA(x,y) = \Gamma(x).\Gamma(y) \tag{6.7}$$

- **Allocation des ressources (AR)** : Cette métrique est proposée par Zhou et al. [Zhou 2009], elle est motivée par le processus d'allocation physique des ressources. La métrique AR est similaire à la mesure Adamic/Adar. En effet, les résultats obtenus par AA et AR ont très proche, mais la métrique AR fonctionne mieux pour les réseaux à fort degré de connectivité. En outre, AR utilise non seulement les voisins directs, mais aussi envisage les voisins de voisins.

$$AR(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \tag{6.8}$$

La table 6.1 [Wang 2015], représente une comparaison entre les approches populaires basées sur le voisinage des nœuds. Les critères de comparaison sont : la complexité et leurs principales caractéristiques.

TABLE 6.1 – Comparaison entre les approches basées sur le voisinage des nœuds

Approche	Complexité	Caractéristique
Voisins communs	$O(n^2)$	Simple et intuitive
Coefficient de Jaccard	$O(2n^2)$	La proposition des nœuds communs est relative au nombre total des voisins
L'indice de Sorensen	$O(n^2)$	Les nœuds ayant les degrés les plus faibles auraient plus grande probabilité d'avoir un lien entre eux.
La similarité cosinus	$O(n^2)$	Métrique de cosinus
Adamic/Adar	$O(2n^2)$	Les nœuds ayant moins de caractéristiques communs sont pondérées plus lourdement.
L'attachement préférentiel	O(2n)	Les nœuds ayant des degrés élevés seront plus susceptibles d'avoir des nouveaux liens entre eux.
Allocation des ressources	$O(2n^2)$	Similaire à Adamic/Adar mais pénalise les nœuds ayant les degrés les plus élevés

La complexité des approches est un facteur important, en particulier pour les réseaux sociaux à grande échelle. Supposons que n est le nombre moyen de voisins dans un réseau, la complexité de trouver tous les voisins d'un nœuds x est en $O(n)$, et la complexité de calculer l'intersection ou l'union de deux ensembles est en $O(n^2)$. Les approches VC, IS et SC ont une complexité de $O(n^2)$ parce qu'elles ont besoin une intersection entre deux ensembles pour avoir des résultats. La complexité de l'approche CJ est en $O(2n^2)$ parce qu'elle calcule une intersection et une union de deux ensembles. La complexité de AA et AR est en $O(2n^2)$ parce qu'elles ont besoin de faire une intersection entre deux ensembles et de trouver les voisins des voisins communs. Pour l'approche AP, elle doit trouver seulement les voisins communs entre deux nœuds, et sa complexité est de l'ordre de $O(2n)$.

6.2.2.2 Approches basées sur le chemin

Il existe dans la littérature d'autres approches de prédiction des liens qui se basent sur les chemins qui existent entre les nœuds pour prédire des liens possibles. Ce genre d'algorithme est très coûteux, car ils prennent en considération la topologie du graphe complet pour parcourir tous les chemins qui existent entre les nœuds. Ils se basent sur l'hypothèse que dans le cas où il existe plusieurs relations de longueurs variables entre deux nœuds, cela peut conduire à l'apparition d'une relation entre ces deux nœuds dans le futur. Parmi ces approches on peux citer :

- **Chemin local** : L'approche CL [Linyuan 2009] fait appel aux informations disponibles sur les chemins locaux de longueur 2 et de longueur 3. Contrairement aux approches qui utilisent uniquement les informations de voisinage, elle exploite certaines informations complémentaires des voisins loin du nœud actuel. Dans cette approche les chemins de longueur 2 sont les plus favorisés que les chemins de longueur 3. La métrique utilisée par cette approche est définie par la formule suivante :

$$CL(x,y) = A^2 + \alpha A^3 \tag{6.9}$$

Où A^2 et A^3 indiquent les matrices d'adjacences des nœuds ayant une distance entre eux de d'ordre 2 et 3. Par conséquent, CL représente aussi une matrice d'adjacence qui décrit les paires des nœuds ayant une distance d'ordre 2 et 3 entre eux.

- **L'approche Katz** : Cette approche prend tous les chemins entre deux nœuds en considération puis, Elle sélectionne les chemins les plus courts qui sont les plus favorisés. Elle attribue à chaque chemin un poids, puis elle prend les chemins qui ont le poids le plus réduit. Elle utilise un facteur de β^l où l est la longueur du chemin d'accès

entre deux nœuds [Linyuan 2009], [Chelmis 2013]. La formule utilisée par l'approche Katz est définie comme suit :

$$Katz(x,y) = \sum_{l=1}^{n} \beta^l |path_{x,y}^{<l>}| \tag{6.10}$$

Où $path_{x,y}^{<l>}$ est le nombre de tous les chemins entre x et y qui une longueur $=$ l. Si $path_{x,y}^{<l>}$=1 signifiait que x et y ont un chemin partagé, 0 sinon. Pour $path_{x,y}^{<l>}$ >1 signifiait qu'il existe l chemins entre x et y. Le variable β peut être utilisé pour contrôler la longueur des chemins qui doivent être considérés. Une très petite valeur de β, signifié que l'algorithme Katz converge vers les approches basées sur le voisinage.

- **Lien d'amitié (LA)** : L'approche LA [Papadimitriou 2012] est basée sur la similarité entre deux nœuds x et y. Elle parcourt tous les trajets d'une longueur limitée entre eux. Elle se base sur l'hypothèse que les utilisateurs d'un réseau social peuvent utiliser tous les chemins entre eux. Le nombre des chemins de longueur L entre x et y représente la similarité entre eux. La formule utilisée par l'approche LA est définie comme suit :

$$LA(x,y) = \sum_{i=2}^{(} n-1) \frac{1}{i-1} . \frac{|path_{x,y}^{i}|}{\prod_{j=2}^{i}(n-j)} \tag{6.11}$$

- **SimRank (SR)** : Cette approche se base sur l'hypothèse suivante, si deux nœuds sont référencés par plusieurs objets similaires, alors ces deux nœuds ont une valeur de similarité très élève. Cette similarité augmente la possibilité d'avoir un lien entre ces deux nœuds. Elle affecte à chaque nœud un score de similarité initialisé à 1. Les deux nœuds x et y ont similaire si ils ont les mêmes scores et qui partagent les mêmes voisins [Jeh 2002]. La formule utilisée par l'approche SimRank est définie comme suit :

$$SR(x,x) = 1 \tag{6.12}$$

$$SR(x,y) = \delta . \frac{\sum_{a\in\Gamma(x)} \sum_{b\in\Gamma(y)} SR(x,y)}{\Gamma(x).\Gamma(y)} \tag{6.13}$$

Où δ est un constant avec $\delta \in$[0,1], il est considéré comme une mesure de confiance.

- **L'approche basée sur la marche aléatoire** : L'idée de cette approche est de commencer les sauts à partir d'un nœud x jusqu'à un autre nœud y de façon aléatoire. Le nombre des sauts aléatoires entre x et y est défini par la fonction $Hitting - Time$ (x,y). Une petite valeur obtenue par la fonction $Hitting - Time(x,y)$ signifie qu'il

y a une forte similarité entre x et y, et qui implique une possibilité d'avoir un lien dans le future entre ces deux nœuds. Cette fonction n'est pas symétrique, pour la rendre symétrique une autre variante a été proposée appelé Commute-time(x,y), cette fonction considère que le nombre des sauts entre x et y est le même pour un saut aléatoire allant de y à x. La formule Commute-time est définie comme suit :

$$Com - time(x, y) = Hitting - Time(x, y) + Hitting - Time(y, x) \tag{6.14}$$

- **Rooted PageRank** Cette approche est une variante de la mesure de similarité utilisée dans le web PageRing pour avoir les liens entre les pages web. L'approche Rooted PageRank (RPR) attribue une probabilité α à chaque saut de x vers y. Cette probabilité signifie la possibilité de revenir vers x (1-α pour aller a un voisin aléatoirement). Cette variante est asymétrique et peut être faite par inversement du rôle de x et y.
Le score RPR entre tous les paires de nœuds est calculé selon la formule suivante : Soit D la diagonale de la matrice d'adjacence définie comme suit :

$$D[i, i] = \sum_{j} A[i, j] \tag{6.15}$$

Soit N est la matrice normalisée de la matrice A définie comme suit :
$N = D^{-1}$
Puis le Rooted Page Rank est calculé comme suit :

$$RPR(x, y) = (1 - \beta)(1 - \beta N)^{-1} \tag{6.16}$$

- **PropFlow** C'est une approche non supervisé qui calcule la probabilité d'atteindre un nœud y après un nombre défini de saut aléatoire l a partir d'un x nœud. Elle utilise les pions des liens entres les nœuds comme une probabilité de transition. La procédure de saut aléatoire entre les nœuds s'arrête, soit une fois le nœud y est atteint soit le nœud x est revisité. Cette approche est plus tolérante aux bruits topologiques.

- **Le plus court chemin** Cette approche utilise une mesure directe pour calculer la similarité entre nœuds dans le réseau social. Elle est définie comme la plus petite distance qui sépare x et y. Plus précisément, on initialise $S = x$ et $D = y$, a chaque étape on élargit les deux ensembles en incluions les voisins directes pour chaque nœud de façon récursive. Cette procédure s'arrête lorsqu'on trouve un nœud appartient aux deux ensembles.

Cet état de l'art nous aide à positionner notre contribution par rapport aux autres approches. En effet, nous devrions attirer l'attention sur le fait que, bien qu'il existe

de nombreuses mesures basées sur le concept de voisinage, mais dans la pratique, on doit choisir des approches en fonction des caractéristiques des réseaux sociaux, car de nombreuses études expérimentales ont montré qu'il n'y a pas une solution optimale dominante pour résoudre le problème de prédiction des liens en se basant sur le voisinage. Les approches basées sur le chemin sont très coûteuses, car elles prennent en considération le graphe complet pour parcourir tous les chemins qui existent entre les nœuds.

6.3 Les problèmes résolus par la prédiction des liens

Dans ces dernières années plusieurs travaux ont été proposés pour la résolution des problèmes spéciaux de la prédiction de lien qui peut être divisé en six catégories : (1) la prédiction des liens temporels, (2) la prédiction des liens actifs ou inactifs, (3) la prédiction des liens dans les réseaux bipartis, (4) la prévision des liens dans des réseaux hétérogènes, (5) la prédiction de l'apparition et la disparition des liens et (6) la prédiction des liaisons évolutives.

6.3.1 La prédiction des liens temporels

Au cours des dernières années, les études menées sur la prédiction des liens sont évolué dans plusieurs aspects. Un de ces aspects consiste à prendre en considération le temps dans le modèle du réseau social, qui peut être nommé la prédiction des liens temporels [Dunlavy 2011] [O'Madadhain 2005]. On peut représenter un réseau social selon l'axe du temps comme une matrice multidimensionnelle Z de taille M x N x T et qui peut être défini comme suit :

$$Z(i,j,t) = \begin{cases} 1 & \texttt{Si} \quad i \text{ avoir un lien avec } j \text{ dans le moment } t \\ 0 & \texttt{Sinon} \end{cases} \tag{6.17}$$

6.3.2 La prédiction des liens actifs ou inactifs

Les auteurs Munasinghe et al. ont proposé une hypothèse que si une paire de nœuds interagit récemment, le lien entre eux devient actif [Munasinghe 2013]. L'horodatage de la dernière interaction est une information vitale qui nous permet de décider si le lien est actif ou non. Par conséquent, la période la plus récente des horodatages des interactions entre les nœuds est utilisée pour calculer les futures prévisions.

6.3.3 La prédiction des liens dans les réseaux bipartis

Plusieurs réseaux sociaux peuvent être modélisés sous forme un graphe biparti, comme par exemple les réseaux entreprises/clients/produits. Cependant, les méthodes de la résolution du problème de la prédiction des liens sont généralement définies sur des graphes non pas bipartis ou ordinaires. Ce qui influe négativement sur les résultats obtenus lorsqu'on les applique sur les réseaux modélisés par un graphe biparti, car ces méthodes ne prennent pas en considération les propriétés de ce genre des réseaux.

Récemment, Il y a quelques chercheurs qui ont pris en considération la résolution de ce problème. Kunegis et al. ont montré que la méthode attachement préférentiel est la seule méthode parmi les méthodes de la prédiction des liens simples qui peuvent être utilisés dans les réseaux bipartis [Kunegis 2010]. D'autres chercheurs ont proposé des variantes et des extensions pour les méthodes de prédiction des liens classiques telles que la méthode des voisins communs, le coefficient de Jaccard, Adamic Adar, et l'attachement préférentiel pour qu'ils seront adapté aux réseaux bipartis [Xia 2012] [Chang 2012]. L'idée clé est de remplacer les voisins directs par les voisins des voisins dans la méthode des voisins communs. Allia et al. ont transformé un graphe biparti en graphe ordinaire par projection, ensuite, ils appliquent les méthode de la prédiction des liens pour prévoir les liens manquants [Allali 2011].

6.3.4 La prédiction des liens dans des réseaux hétérogènes

Récemment, la plupart des auteurs qui ont proposé des solutions pour la prédiction des liens ont axé leurs idées sur les réseaux homogènes, dans lesquels un seul type de nœud et un seul type de lien existe dans le réseau. Cependant, en réalité il existe certains réseaux sociaux qui sont hétérogènes et qui contiennent différents types de liaisons et différents types de nœuds. Ce qui peut impliquer différentes typologies et mécanismes de création des liaisons. Par conséquent, la prédiction des liens dans les réseaux sociaux hétérogènes n'a pas eu l'attention nécessaire par les chercheurs.

6.3.5 La prédiction de l'apparition et la disparition des liens

L'apparition et la disparition des liens dans les réseaux sociaux sont deux procédures fondamentales qui assurent le changement et l'évolution dynamique du réseau. Par exemple, un utilisateur de Facebook peut créer un lien d'amitié avec un autre utilisateur par le clic sur le bouton « Ajouter » comme il peut supprimer ce lien par le clic sur le bouton "Retirer de la liste des amis". Le même principe utilisé pour Twitter, si

un utilisateur clic sur le bouton "Suivre" un nouveau lien sera créé, cependant s'il clic sur le bouton "ne plus suivre " le même lien sera disparu. Dans notre connaissance, de nombreux efforts ont été faits dans l'étude du problème de l'apparition des nouveaux liens dans l'avenir, mais seulement quelques tentatives ont été proposées pour étudier le problème de la disparition des liens dans l'avenir dans un réseau social.

Kwak et al. ont analysé la disparition des liens dans Twitter, par l'analyse des comportements des utilisateurs qui clic souvent sur le bouton « ne plus suivre » et les raisons pour lesquelles ils choisis cette option. Ils ont trouvé que l'un des facteurs principaux de la disparition des liens est l'absence de la réciprocité dans la relation [Chun 2012]. Ils ont également pris en considération d'autres facteurs [Kwak 2012], en particulier les informations personnelles et les intérêts mutuels. Les résultats de leur étude montrent que la relation "Suivre" mutuelle entre les utilisateurs réduire la probabilité de disparition des liens dans Twitter.

6.4 Conclusion

Dans ce chapitre nous avons présenté le domaine d'application de notre approche d'agrégation textuelle, qui est celui des réseaux sociaux, et plus particulièrement la prédiction des liens dans ces réseaux. Nous avons présenté aussi un état de l'art sur les approches de la prédiction des liens afin de positionner notre contribution, suivi par une présentation des problèmes résolus par la prédiction des liens. Dans le chapitre suivant, nous présentons le passage de l'OLAP vers le Social OLAP et nous exposons en détail notre contribution de la prédiction des liens sémantiques basée sur l'agrégation textuelle.

Chapitre 7

Diamant : une nouvelle approche pour la prédiction des liens basée sur l'agrégation textuelle

Sommaire

7.1 Introduction

Dans cette dernière décennie, les réseaux sociaux, et plus particulièrement les réseaux dédiés aux chercheurs, telles que Mendeley ou ResearchGate sont devenus omniprésents sur le web ce qui a ouvert les portes vers une nouvelle génération telle que le web social. Ce genre des plateformes offre l'opportunité aux chercheurs de partager leurs travaux scientifiques, de communiquer entre eux et de débattre les sujets scientifiques, ce qui rend ces réseaux sociaux une source vitale d'information.

La masse volumineuse d'information et les publications partagées dans les réseaux sociaux dédiés aux chercheurs nécessitent l'analyse et l'exploration, afin d'étudier le comportement scientifique des auteurs à travers leurs publications scientifiques.

Cette analyse fait appel à plusieurs disciplines telles que la sociologie et l'histoire des sciences, les sciences de l'information, la linguistique computationnelle, et la statistique. Elle prend en considération l'ensemble des tâches du processus de la production scientifique, comme la publication des articles, les pratiques des citations, les relations de collaboration entre les auteurs et l'extraction des thèmes sur lesquels portent les publications. Le processus d'analyse de ces réseaux permet de découvrir des connaissances comme la prédiction des liens cachés ou sémantiques, la détection de communautés et la prédiction des événements.

Plusieurs études sont entamées pour analyser des réseaux sociaux, ils les considèrent comme des graphes complexes et hétérogènes [Jensen 2004] [Getoor 2007] [Du 2010] [Rossetti 2011] [Bothorel 2015]. Une des questions posées lors de l'analyse des réseaux sociaux dédiée aux chercheurs, est comment obtenir une vue plus significative de ces entités en se basant sur les relations existantes et sur l'agrégation des productions textuelles des entités ? Ceci peut naturellement se baser sur l'OLAP car l'agrégation et la visualisation sont deux points centraux de ce type de méthode d'analyse. Pour analyser et visualiser les données à travers plusieurs axes et selon plusieurs niveaux hiérarchiques, le cube OLAP offre une vision multidimensionnelle à plusieurs dimensions. Cette projection fait appel à un nouveau paradigme de recherche, qui est d'étendre l'OLAP à l'analyse des réseaux d'informations hétérogènes ou le social OLAP [Loudcher 2015]. Plusieurs auteurs recommandent qu'il soit nécessaire de faire évoluer l'OLAP vers le social OLAP [Morfonios 2008] [Tian 2008] [Chen 2008] [Loudcher 2015]. Ce nouveau concept consiste à adapter et adopter les méthodes et les techniques utilisées dans les systèmes d'aide à la décision par le processus d'analyse en ligne (OLAP), comme les opérations de forages vers le haut ou vers le bas, l'agrégation et la visualisation des données, pour analyser les réseaux sociaux.

Ce chapitre se devise en trois parties, dans la première partie on va présenter l'adaptation de l'OLAP traditionnel aux réseaux d'information afin d'avoir le social OLAP, dans la deuxième partie on va présenter notre nouvelle approche DIAMANT qui permet de résoudre le problème de la prédiction des liens sémantiques dans les réseaux sociaux on se basant sur l'agrégation textuelle, ce chapitre se termine par une étude expérimentale et comparative qui montre l'efficacité de notre proposition.

7.2 De l'OLAP vers le Social OLAP

Dans ces dernières années, de nombreux travaux de recherche s'intéressent à faire évoluer OLAP pour explorer et analyser les données issues des réseaux sociaux. Cette évolution nécessite l'adaptation et l'élargissement de l'OLAP sur des graphes complexes et hétérogènes ce qui nous donne le «Graph OLAP». Nous considérons les réseaux sociaux comme de grandes bases de données qui contiennent des données complexes, images, vidéos, documents, et d'autres données structurées et semi-structurées, la différence réside dans le fait que les enregistrements stockées dans ces bases de données sont interconnectés et liés entre eux à travers des relations multiples, contrairement aux bases de données traditionnelles où les enregistrements stockées sont isolés.

Par exemple, PubMed est une base de données traditionnelle qui stocke des publications scientifiques. On peut lui appliquer l'OLAP pour analyser le contenu textuel en utilisant les nouvelles fonctions d'agrégation textuelles. Un autre niveau d'analyse sera entamé si on ajoute des liens entre les différentes publications stockées dans la base de données PubMed. Ces liens sont des informations supplémentaires permettant de lier les enregistrements via les auteurs, les citations, les institutions, les thèmes etc. C'est-à-dire on peut construire un réseau des co-auteurs pour visualiser les collaborations entre auteurs, le réseau des citations et le réseau des conférences. L'OLAP traditionnel ne prend pas en charge ces informations pertinentes qui nécessitent l'élargissement de l'OLAP vers le social OLAP. Le social OLAP prend en compte le contenu et la structure du graphe et fournit en sortie un graphe plus général pour extraire des connaissances et de résoudre des problèmes d'analyse des réseaux sociaux tels que la prédiction des liens. Dans le social OLAP on peut distinguer plusieurs types de dimensions : les dimensions informationnelles (comme dans l'OLAP traditionnel) et les dimensions topologiques (contient des informations sur la structure). Cette dernière constitue une vraie valeur ajoutée dans la modélisation car elles permettent de modéliser les relations qui existent entre les objets. Loudcher et al. a comparé dans [Loudcher 2015] les caractéristiques de l'OLAP traditionnel et le social OLAP.

Certains chercheurs sont intéressées à l'étude du social OLAP, et l'étude des graphes hétérogènes d'information. Zhao et al dans [Zhao 2011] ont utilisé les fondements de Graph OLAP défini par [Chen 2008] pour introduire un nouveau modèle de cube de données, appelé Graph Cube. Dans le même axe [Qiang 2011] proposent un opérateur qui se base sur la topologie du réseau appelé T-OLAP (Topologie OLAP), il sert à modifier la structure de réseaux lorsque l'utilisateur lance une requête du type T-OLAP.

TABLE 7.1 – Comparaison entre l'OLAP traditionnel et le Social OLAP

	OLAP traditionnel	**Social OLAP**
Données	Données relationnelles ou semi-structurées	Objets interconnectés et de types différents
Problèmes	Non prise en compte des liens entre les objets	Prise en compte des liens entre les objets
Entrée	Faits multidimensionnels	Un réseau et des attributs
Sortie	Données agrégées	Un réseau agrégé ou plus général
Dimensions	Uniquement informationnelles	Informationnelles et topologiques
Mesures	Généralement numériques	Fonctions d'agrégation spécifiques
Opérations	Opérations classiques	Opérations spécifiques

D'autres travaux se sont inspiré de Graph OLAP, Yin et al. dans [Yin 2012] proposent un nouveau concept baptisé HMGraph OLAP qui adapte la modélisation des réseaux d'informations homogènes par l'ajout d'une nouvelle dimension celle de la dimension "entité" plus les deux dimensions standard du social OLAP, la dimension informationnelle et topologique. Cette nouvelle dimension "entity dimension" sert à modéliser l'hétérogénéité des entités et des relations. Dans HMGraph OLAP la dimension topologique sert à créer de nouveaux types de nœuds, et la dimension-entité sert à changer la signification des arêtes et compte le nombre de nœuds dans le réseau. Un autre modèle proposé par Benatallah et al. baptisé GOLAP, un modèle qui prend en considération les nœuds d'un réseau et les liens entre eux [Benatallah 2012]. Ils adaptent les notions de l'OLAP traditionnel pour les projeter aux graphes. Ils proposent deux opérateurs UPDATE et UPSERT pour pouvoir réaliser efficacement des navigations multidimensionnelles sur les cubes de données.

La caractéristique commune entre les travaux présentés est dans le fait qu'il repose sur des réseaux homogènes issus d'un entrepôt de données (exemple : auteur x auteur). Cependant, ils ne prennent pas en considération les réseaux hétérogènes (exemple : auteur x auteur x contenu) d'une part, et que la majorité d'entre eux cherchent à agréger un réseau social en se basant sur des données numériques d'autre part. Pour surmonter ces limites, on a adopté une autre vision différente qui s'intéresse à l'intégration de contenu dans le réseau issu d'un entrepôt de données. En effet, par exemple

à partir d'un corpus de données des publications scientifiques, on peut extraire par exemple le réseau (auteur x auteur) où les nœuds sont les auteurs et les arêtes sont les liens de co-rédaction d'un papier, puis on utilise le contenu des papiers scientifiques publié par les auteurs pour prédire des liens sémantique entre les nœuds de réseau.

7.3 L'approche Diamant

Les réseaux sociaux sont généralement représentés par un graphe afin de faciliter la visualisation des nœuds et des liens. Les problèmes engendrés par l'évolution exponentielle et dynamique, plus les limites des capacités des opérateurs OLAP traditionnels qui ne permettent pas l'analyse et la résolution des problèmes actuels d'une part et qu'ils ne prennent pas en considération le contenu textuel échangé entre les membres de ces réseaux d'une autre part, nous ont poussé à proposer une nouvelle approche pour résoudre un de ces problèmes qui est la prédiction des liens dans les réseaux sociaux. L'idée est de fusionner la notion d'agrégation textuelle issue de Text OLAP et le concept du graphe hétérogène issue de la théorie des graphes.

Cette approche sert à faire la prédiction des liens sémantiques dans un réseau social en se basant sur l'agrégation de contenu textuel produit par les membres de ce réseau en utilisant une fonction d'agrégation issue de Text OLAP. Notre approche baptisée $DIAMANT$ est une extension de l'approche TAG proposée dans la section 4.6 adaptée pour les réseaux sociaux. Elle se déroule en quatre étapes : (1) elle commence par l'extraction des mots-clés apparus dans le contenu produit par les membres du réseau ; (2) la création du graphe d'affinité entre les mots-clés afin de les agréger, puis l'extraction des différents circuits formés dans le graphe ; (3) la construction du graphe hétérogène qui correspond à chaque circuit, et enfin (4) la prédiction des liens sémantiques entre les différents membres.

Pour formaliser notre approche on va prendre $G_t<U,E>$ une capture d'un réseau social à l'instant t, où U représente l'ensemble des membres qui compose ce réseau $U=\{u_1, u_2, ..., u_n\}$ et E représente l'ensemble des liens entre les différents membres $E=\{e_1, e_2, ..., e_n\}$. Les membres du réseau partagent entre eux un ensemble de documents $D=\{d_1, d_2, ..., d_n\}$ où chaque membre a contribué dans la rédaction d'un ou plusieurs documents. L'ensemble $Kw=\{kw_1, kw_2, ..., kw_n\}$ représente l'ensemble des mots-clés extrait à partir des documents D fournis par l'ensemble des membres U.

Après l'extraction des mots-clés à partir des document en se basant sur l'outil d'extraction des mots-clés présenté dans la section 5.3, on obtient une matrice $MFDKW$

des fréquences où les lignes représentent les documents, les colonnes représentent les mots-clés et les cellules représentent la fréquence d'apparition de chaque mot-clé dans chaque document. Cette matrice est utilisée comme paramètre d'entrée afin de générer une autre matrice qui s'appelle la matrice d'affinité servant à trouver le degré d'affinité entre les différents mots-clés. Les lignes et les colonnes de cette matrice représentent les mots-clés et l'intersection entre eux représente leur affinité.

L'application de l'approche TAG nous a permis de créer un graphe d'affinité entre les mots-clés, puis elle donne les différents circuits générés dans le graphe. L'idée est de créer un réseau hétérogène contient plusieurs types de nœuds tels que les mots-clés, les documents et les membres de réseaux afin de faire la prédiction des liens sémantiques. Après l'obtention de tous les cycles, où les mots-clés représentent les nœuds, on établie des liens entre chaque mot-clé et les documents dans $D' = \{d'_1, d'_2, ..., d'_m\} \subset D$, qui le contiennent. On obtient l'ensemble des liens $LKwD$=$\{(kw_1, d'_1)$,(kw_2, d'_2),...,$(kw_n, d'_n)\}$, puis on crée des liens entre chaque document sélectionné et le membre qui ont contribué à sa rédaction. On obtient l'ensemble des liens LDU=$\{(d_1, u'_1)$,(d_2, u'_2),...,$(d_n, u'_n)\}$ où U'=$\{u'_1, u'_2, ..., u'_e\} \subset U$. La structure finale obtenue est se forme un diamant comme il est illustrer dans la figure 7.1.

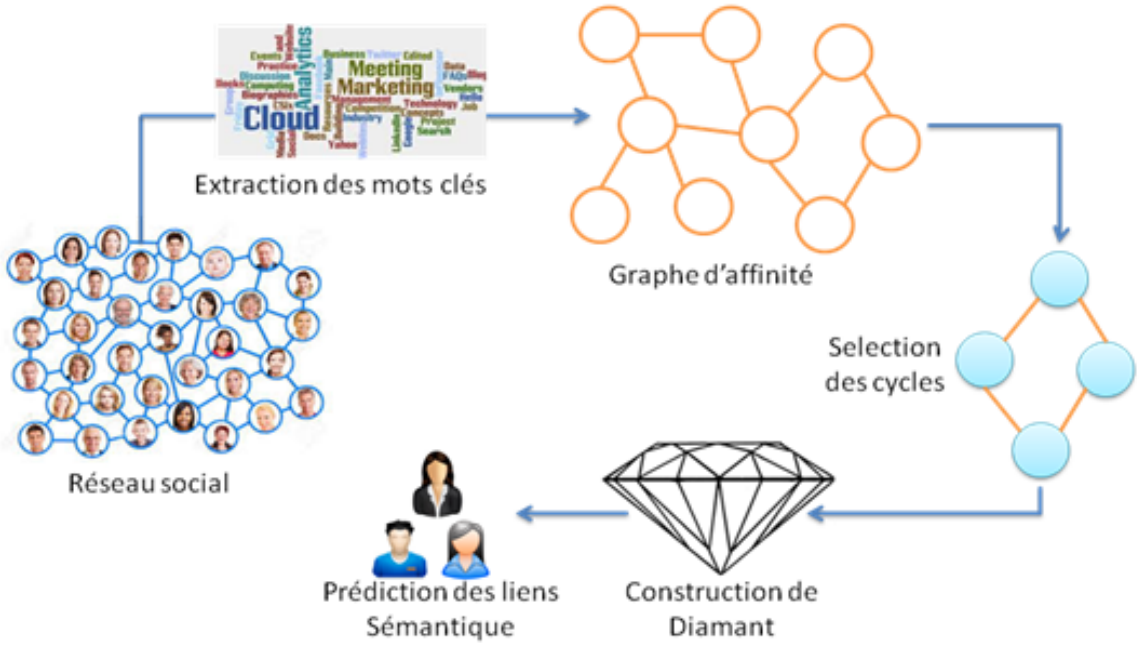

FIGURE 7.1 – Les étapes d'application de l'approche Diamant

L'étape suivante consiste à extraire les différents liens qui existent dans le nouveau graphe hétérogène qui sont classés en deux catégories, les liens de type co-auteur où

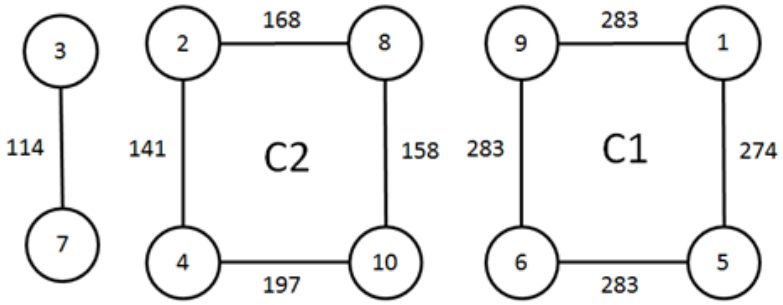

FIGURE 7.2 – Le graphe d'affinité

les membres contribuent à la production de même document et les liens sémantiques entre les membres lorsqu'ils partagent le même mot clés agrégé. Une fois l'opération d'extraction des liens est terminée, on calcule la crédibilité par la formule 7.1 pour chaque lien, afin de les trier selon leurs valeurs de crédibilité la plus grande, puis on donne à l'utilisateur la possibilité de définir le nombre des liens souhaités. La figure 7.1 illustre les différentes étapes d'application de notre approche *Diamant*.

$$CL(x,y) = 1 - \frac{|NF(x) - NF(y)| + |NsD(x) - NsD(y)|}{Max(NF(x), NF(y)) + Max(NsD(x), NsD(y))} \tag{7.1}$$

où NF représente le nombre des amis de X et NsD représente le nombre des documents partagés par l'auteur X.

7.3.1 L'algorithme Diamant

Dans cette section nous présentons l'algorithme proposé pour prédire les liens entre les membres du réseau en utilisant la fonction d'agrégation textuelle TAG décrite dans la section 4.6.

7.3.2 Exemple d'application

On reprend le même exemple utilisé dans la section précédente, où il y a 20 auteurs A=$\{a_1, a_2, ..., a_k\}$ qui partagent un ensemble de 13 articles scientifiques. Ces documents continent un ensemble de mots-clés qui est le suivant : Kw = { (T_1 = OLAP), (T_2 =XML), (T_3 = Datamining), (T_4 = Query), (T_5 =datawarehouse), (T_6 = Document), (T_7 =System), ($T8$ = F unction)(T_9 = Cube), ($T_1$0 =Network) }. Après l'application de la fonction TAG on a obtenu une matrice des fréquences, une matrice d'affinité et les circuits d'affinité entre les mots-clés.

Algorithme 3 DIAMANT

Entrées
    Un corpus de documents $D = \{D_1, D_2, ..., D_n\}$
    Un réseau social sous forme d'une matrice Auteur x Auteur NetSocial
Sorties
    Liste des liens prédits ordonnés selon leurs crédibilités
Début
// Construction des circuits d'agrégation
    TAG(D) ;
// Calcul de la somme des fréquences de chaque mot-clé
    Pour chaque Circuit faire
        Pour chaque Kw ∈ Circuit faire
            LierKwDocAuteur (Kw,D,A) ;
// Extraction des liens
        Pour chaque Document sélectionné dans le circuit faire
            Si LienCoauteurNexistePas(NetSocial, D, A') alors
            ListLiensPrédits.add(D,A') ;
        Pour chaque Kw sélectionné dans le circuit faire
            Si LienSementiqueNexistePas(NetSocial, Kw, D, A') alors
            ListLiensPrédits.add(Kw, D, A') ;
// Calcul de la crédibilité des liens
    Crédibilité(ListLiensPrédits) ;
    OrdonnerLiens(ListLiensPrédits) ;
Fin.

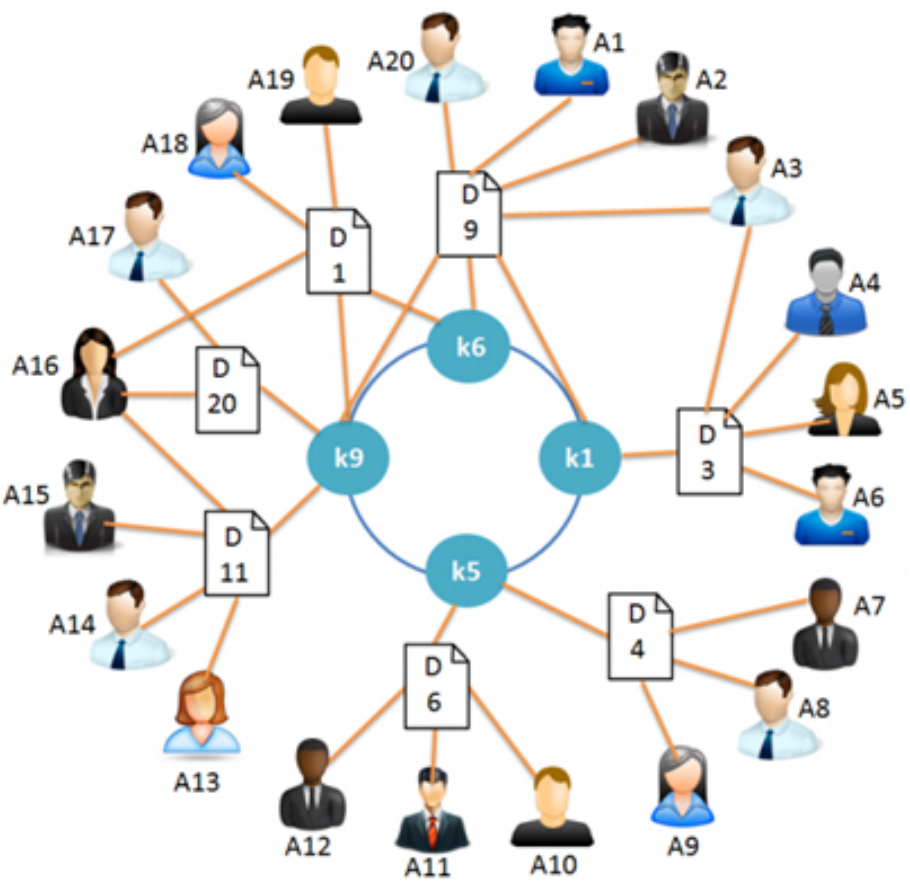

FIGURE 7.3 – La structure "Diamant" obtenue pour le circuit C_1

Puis on construit pour chaque circuit le Diamant qui convient. Une fois les graphes hétérogènes construits figure 7.3, on commence l'extraction des liens sémantiques. Une fois les liens extraits, on calcule leur crédibilité en utilisant la formule 7.1. Le nome de Diamant vient de la forme obtenue après la construction du graphe hétérogène entre les mots clés, les documents et les auteurs, comme le montre la figure 7.4.

Dans cet exemple, on peut distinguer deux types des liens : les liens co-auteurs et les liens sémantiques. Le premier type des liens représente les liens co-auteurs où les auteurs contribuent à la rédaction d'un ou plusieurs documents. Le deuxième type représente les liens sémantiques qui sont crées entre les auteurs qui partagent les mêmes mots clés agrégés et qui n'ont pas des co-auteurs. Dans notre cas on a trouvé 23 liens co-auteurs et 43 liens sémantiques. L'auteur A_1 est lié en tant que co-auteur avec les auteurs A_2, A_3 et A_{20} qui ont contribué à la rédaction du document D_9 d'une part et liés sémantiquement avec A_4, A_5, A_6 à travers le mot clé Kw_1 et avec A_13, A_{14}, A_{15}, A_{16}, A_{17} à travers le mot clé Kw_9 et aussi liés sémantiquement avec A_{18} et

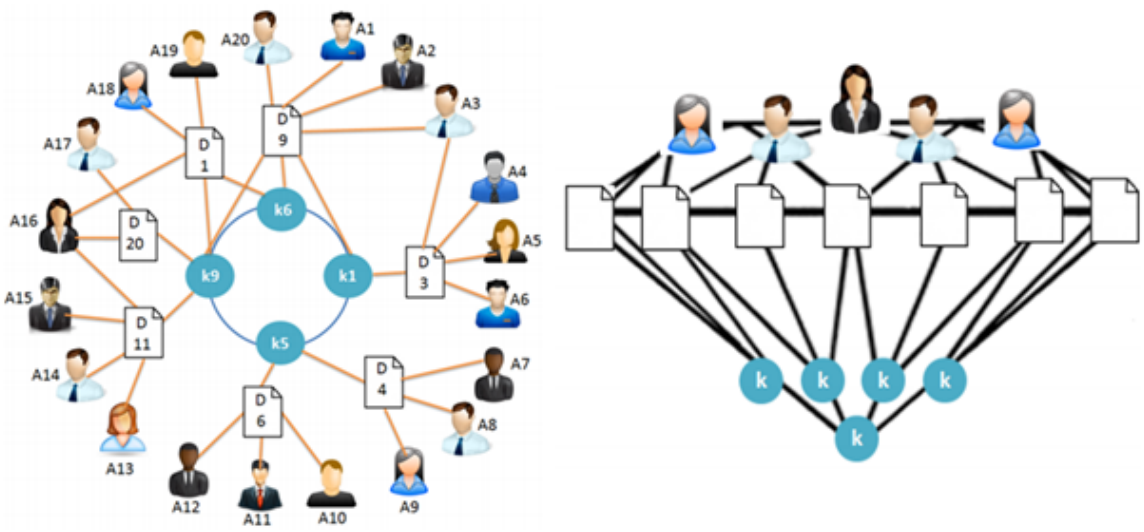

FIGURE 7.4 – La forme de diamant

A_{19} à travers le mot clé kw_6.

7.4 Étude expérimentale

Nous évaluons les performances de notre algorithme Diamant en le testant sur un réseau social réel et en le comparant à des techniques existantes. En premier lieu, nous préparons le réseau social de test qui représente les auteurs qui ont participé à la conférence IT Innovation durant la période 2008-2014. Nous implémentons notre approche avec cinq autres approches les plus connues dans la littérature, et nous comparons leurs performances en utilisant des mesures standard tel que le rappel, la précision et la F-mesure.

7.4.1 Description de réseau social utilisé

Le réseau social utilisé dans notre étude expérimentale baptisé «ITInnovation social network» est un réseau des chercheurs qui ont contribué à la rédaction des articles scientifiques communiqués dans la conférence ITInnovation et publié dans IEEE explorer durant la période entre 2008 jusqu'à 2014. Les caractéristiques de ce réseau sont résumé dans le tableau 7.2

Après les opérations de prétraitement appliquées sur les documents partagés entre les membres de ce réseau comme il est présenté dans la section 5.3, et pour simplifier le traitement de réseau, on a construit des matrices telles que la matrice Auteur x

TABLE 7.2 – Caractéristiques du réseau social ITInnovation

Nombre des auteurs	1285
Nombre des documents	600
Nombres des termes	800000
Nombres des mots clés	2000
Nombres des liens	10000

Auteur qui représente les liens de collaboration co-auteurs et une matrice auteur x document qui représente la contribution des auteurs dans la rédaction des documents.

Pour comparer les résultats obtenus par la prédiction des liens dans ce réseau social on a appliqué les approches implémentées sur un réseau initial, formé des liens entre les membres durant la période 2008 jusqu'à 2011, pour obtenir une prédiction du réseau en 2014. Ce dernier est comparé avec le réseau observé, formé des liens entre les membres durant la période 2008 jusqu'à 2014.

Plus formellement, pour suivre l'évolution d'un réseau social et de mesurer le degré de son changement dans plusieurs périodes, on considère $G_t =< V, E_t >$ et $G'_{t'} =< V, E'_{t'} >$ deux captures du réseau social dans deux périodes différentes t et t'. $G'_{t'}$ représente le nouveau réseau social obtenu après l'application d'une approche de prédiction de liens sur le réseau G_t. L'idée est de comparer les deux graphes $G_{t'}$ et $G'_{t'}$. Dans notre contexte t=2011 et t'=2014 où G_t représente l'état du graphe original en t=2011, $G'_{t'}$ représente l'état du nouveau obtenu en t'=2014 et $G_{t'}$ représente l'état du graphe original en t'=2014.

7.4.2 Mesures d'évaluation

Pour évaluer les résultats obtenus par les différentes approches de la prédiction des liens, nous avons utilisé comme mesures de performance : le rappel, la précision et la F-mesure qui est le rapport entre le rappel et la précision [Milen 2007] [Chawla 2012] [Yang 2015] [Daqing 2016]. Ces mesures sont les mêmes en termes d'appellation avec celui-ci utilisées dans l'agrégation textuelle. Mais dans le contexte de la prédiction des liens elles sont définies par d'autres formules afin de bien mesurer la qualité des liens obtenus et la performance des approches de prédiction de liens.

Après l'application des approches étudiées, Nous distinguons 4 situations pour les liens du réseau social : la première situation s'appelle « True Positif (TP) » : où un nouveau lien apparu après l'application d'une approche de prédiction de liens et qui

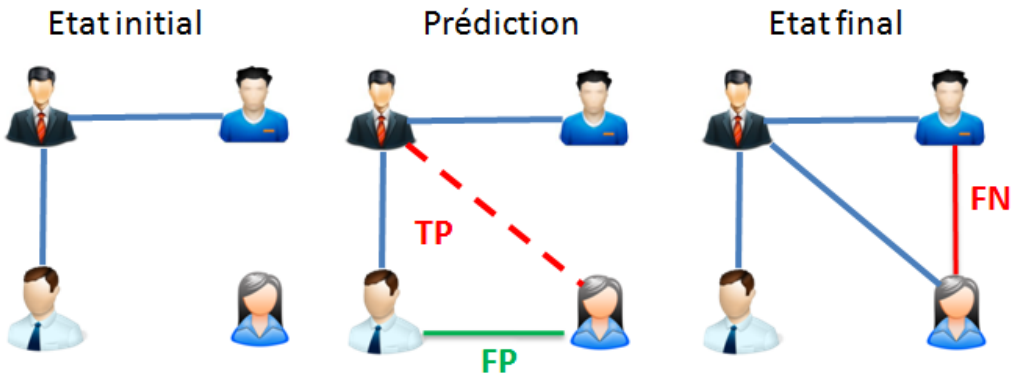

FIGURE 7.5 – Les situations des liens dans un réseau social

existe dans l'état final du réseau, c'est-à-dire : $e' \in E'_{t'}$ et $e' \in E_{t'}$. La deuxième situation s'appelle « False Positif (FP) » : où un nouveau lien apparu après l'application d'une approche de prédiction de liens et qui n'existe pas dans l'état final du réseau, c'est-à-dire : $e' \in E'_{t'}$ et $e' \in E_{t'}$. La troisième situation s'appelle « False Negatif (FN) » : où un lien qui existe dans l'état final du réseau mais qui n'est pas apparu après l'application d'une approche de prédiction de liens, c'est-à-dire : $e' \in E'_{t'}$ et $e' \in E_{t'}$. La dernière situation s'appelle « True Negatif (TN) » : où un lien qui n'existe pas dans l'état final du réseau et aussi qui n'est pas apparu après l'application d'une approche de prédiction de liens, c'est-à-dire : $e' \in E'_{t'}$ et $e' \in E_{t'}$. Cette combinaison de situation est illustrée dans la figure 7.5 [Yang 2015].

Dans notre contexte, le rappel correspond au nombre des liens apparus après la prédiction de liens et qui existent dans l'état final du réseau divisé par la somme du nombre de liens vrais positifs et faux négatif.

$$Rappel = \frac{TP}{TP + FN} \tag{7.2}$$

La précision correspond au nombre des liens apparus après la prédiction des liens et qui existent dans l'état final du réseau divisé par la somme du nombre de liens vrais positifs et faux positifs. Ce rapport est calculé à l'aide de la formule :

$$Prcision = \frac{TP}{TP + FP} \tag{7.3}$$

La F-mesure représente le rapport entre le rappel et la précision. Ce rapport est calculé à l'aide de la formule :

$$F - mesure = \frac{2 * Rappel * Prcision}{Rappel + Prcision} \tag{7.4}$$

7.5 Résultats et discussion

L'objective de cette étude expérimentale est de montrer la performance de notre approche Diamant qui se base sur les données textuelles par rapport aux autres approches. Nous avons implémenté les cinq approches les plus citées dans la littérature telle que l'approche Common Neighbours, l'approche de Jaccards, l'approche Adamic/Adar et l'approche preferential attachment et notre approche Diamant. Nous avons utilisé comme langage de programmation le langage JAVA sous l'environnement Netbeans. La première étape est d'exécuter les approches sur le même réseau social, puis on calcule la crédibilité de tous les liens obtenus par les différentes approches et on les trie par ordre décroisant afin qu'on puisse sélectionner les k premiers liens les plus crédibles. Une fois les k premiers liens sont sélectionnés on calcule le rappel, la précision et la F-mesure pour chaque approche. Les différents résultats obtenus sont présentés dans les tableaux 7.3, 7.4, 7.5, 7.6, 7.7 et 7.8.

TABLE 7.3 – Le nombre de liens observés et prédits par chaque approche dans $t' = 2014$

Nouveaux liens observés en 2014	**Nouveaux liens prédits en 2014**				
	C.Neighb.	Jaccard	Adamic	Pref.Att.	Diamant
10000	144671	144671	309861	301917	3248

TABLE 7.4 – Le nombre des liens prédits selon la crédibilité variée de 0,2 à 1

	1	**0,8**	**0,6**	**0,4**	**0,2**
C.Neighb.	3540	41578	85490	121498	141292
Jaccard	3540	41578	85490	121498	141292
Adamic A.	7488	75917	157738	231600	284189
Pref.Att.	7488	75918	157739	231601	284190
Diamant	161	1079	2940	3248	3248

Le tableau 7.3 présente le nombre des liens observés et prédit dans t'=2014, avec 10000 liens qui sont observés dans le réseau social réel dans t'=2014, et les différents autres valeurs représentent le nombre des liens prédits par les différentes approches dans t'=2014. On remarque que le nombre des liens prédits par notre approche est

TABLE 7.5 – Le pourcentage des liens prédits selon la crédibilité variée de 0,2 à 1

	1	**0,8**	**0,6**	**0,4**	**0,2**
C.Neighb.	2	29	59	84	98
Jaccard	2	29	59	84	98
Adamic A.	2	25	51	75	92
Pref.Att.	2	25	52	77	94
Diamant	5	33	91	100	100

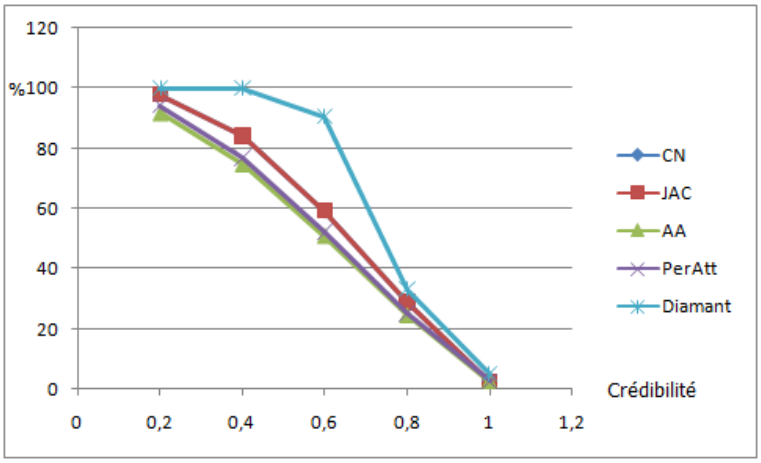

FIGURE 7.6 – Le pourcentage des liens prédits selon la crédibilité variée de 0,2 à 1 pour chaque approche

inférieur au nombre des liens observés dans le réseau réel, par contre pour les autres approches le nombre des liens prédits sont strictement supérieurs au nombre des liens observés (exemple Adamic/Adar génère 309861 liens dans t'=2014). D'après ces résultats on peut constater que les autres approches produisent beaucoup des liens inutiles qui sont des bruits et qui ne figureront plus dans le réseau réel, et qui augmentent la valeur du paramètre « False Positif (FP) » : "où un nouveau lien apparut après l'application d'une approche de la prédiction des liens et qui n'existe pas dans l'état final du réseau", ce qui génère une dégradation dans la précision des résultats de ces approches.

Le tableau 7.4 et le tableau 7.5 représentent les nombres et les pourcentages des liens prédits classés selon leurs crédibilités pour chaque approche. Les intervalles de crédibilité sont $[1, 0.9]$, $]0.9, 0.8]$, ..., $]0.1, 0]$. On constate que pour l'approche Common Neighbours le nombre des liens prédits qui ont une crédibilité égale à 1 est 3540 parmi les 14471 liens prédits ce qui représente 2% des liens. Idem pour les approches

TABLE 7.6 – Le rappel calculé (%) pour chaque approche selon les k premiers liens crédibles

	Les k premiers liens crédibles									
	1000	**2000**	**3000**	**4000**	**5000**	**6000**	**7000**	**8000**	**9000**	**10000**
C.Neighb.	10.8	6.4	6.9	7.4	6.3	5.4	5.1	5.8	5.9	5.9
Jaccard	1.9	4.4	3.9	3.2	3.9	3.8	4.2	4.7	4.5	4.3
Adamic	0	0.1	0.3	0.7	1.5	1.6	1.5	1.4	1.6	1.4
Pref.Att.	1.9	5.9	6.6	6.2	6.2	6.9	7.3	6.5	6.7	6.8
Diamant	49.46	46.52	48.75	43.51	38.13	33.93	30.56	27.81	25.55	23.55

TABLE 7.7 – La précision calculée (%) pour chaque approche selon les k premiers liens crédibles

	Les k premiers liens crédibles									
	1000	**2000**	**3000**	**4000**	**5000**	**6000**	**7000**	**8000**	**9000**	**10000**
C.Neighb.	11.0	6.5	7.0	7.5	6.4	5.5	5.1	5.8	6.0	6.0
Jaccard	2.0	4.5	4.0	3.2	4.0	3.8	4.2	4.7	4.5	4.4
Adamic	0	0.3	0.5	0.7	1.2	1.5	1.5	1.6	1.6	1.6
Pref.Att.	2.0	6.0	6.6	6.2	6.2	7.0	7.4	6.6	6.7	6.9
Diamant	97.90	87.00	95.13	94.88	94.88	94.88	94.88	94.88	94.88	94.88

TABLE 7.8 – La F-mesure calculée (%) pour chaque approche selon les k premiers liens crédibles

	Les k premiers liens crédibles									
	1000	**2000**	**3000**	**4000**	**5000**	**6000**	**7000**	**8000**	**9000**	**10000**
C.Neighb.	10.9	6.4	6.9	7.4	6.3	5.4	5.1	5.8	5.9	5.9
Jaccard	1.9	4.4	3.9	3.2	3.9	3.8	4.2	4.7	4.5	4.3
Adamic	/	1.0	0.3	0.7	1.5	1.6	1.5	1.4	1.6	0.6
Pref.Att.	1.9	5.9	6.6	6.2	6.3	6.9	7.4	6.6	6.7	6.8
Diamant	65.72	60.62	64.44	59.67	54.40	50.00	46.24	43.01	40.20	37.74

Jaccards, Adamic/Adar et preferential attachment où les pourcentages des liens qui ont une crédibilité égale à 1 ne dépassent pas le 2% de nombre total des liens prédits par ces approches. Par contre, notre approche Diamant génère un nombre de 161 liens qui ont une crédibilité égale à 1 est qui représentent 5% des liens prédits. Pour une crédibilité supérieure ou égale à 0.6, le pourcentage des liens prédits qui se trouve dans cet intervalle est varie entre 51% jusqu'à 59% pour les approches Common Neighbours, Jaccards, Adamic/Adar et preferential attachment. En revanche, ce pourcentage de crédibilité s'augmente jusqu'à 91% des liens prédits par notre approche Diamant dans le même intervalle de [1, 0.6] et il atteindra 100% à partir de l'intervalle [1, 0.4]. Ces valeurs montrent que notre approche Diamant offre des résultats crédibles par rapport aux autres approches qui génèrent beaucoup de liens qui ont une crédibilité très faible.

Pour mesurer la performance de chaque approche en utilisant le rappel, la précision et la F-mesure on a varié un paramètre k qui représente les premiers liens les plus crédibles. Les résultats obtenus sont montrés dans les tableaux 7.6 , 7.7 et 7.8. Le tableau 7.6 représente le rappel obtenu par les 5 approches, on remarque que pour k=1000, l'approche Adamic/Adar ne génère aucun lien parmi les liens apparu dans le réseau final en t'=2014, ce qui signifie que la valeur du paramètre « False Negatif (FN) » est très élevé, et qu'est influée négativement sur le rappel obtenus. Pour les approches Jaccard et preferential attachment, elles génèrent un rappel de 1.9% lorsque k=1000. Pour Common Neighbours son rappel arrive jusqu'à 10.8%, par contre elle arrive à 49% avec notre approche Diamant ce qui signifier que le paramètre "True Positif (TP)" est très élevé. Ce paramètre représente le nombre des liens apparus après l'application de notre approche Diamant et qui existe dans l'état final du réseau en t'=2014. Pour k=4000, notre approche Diamant donne une valeur de rappel de 43% comparée à ceux de l'approche Common Neighbours, Jaccards, Adamic/Adar et preferential attachment avec les valeurs de rappel de 7.4%, 3.2%, 0.7% et 6.2 % respectivement. Pour k=10000, le rappel obtenu par notre approche Diamant est de 23.5% comparée à ceux de l'approche Common Neighbours, Jaccards, Adamic/Adar et preferential attachment avec les valeurs de rappel de 5.9%, 3.4%, 1.4% et 6.8% respectivement. La dégradation de rappel de notre approche pour les valeurs de k=4000 et k=10000 relative au nombre des liens obtenus par Diamant qui égale à 3248 qui signifie que le paramètre "False Negatif (FN)" est très élevé qui représente les nombre des liens qui existent dans l'état final du réseau mais qui n'est pas apparu après l'application de notre approche Diamant.

Pour la précision, lorsque k=1000, notre approche Diamant donne une précision

de 97.90% qui signifie que le paramètre « False Positif (FP) » est très faible, ce paramètre représente le nombre des nouveaux liens apparus après l'application de notre approche Diamant et qui n'existe pas dans l'état final du réseau t'=2014. Pour les autres approches la précision variée entre 0%, 2%, 2% et 11% pour les approches Adamic/Adar, Jaccards, preferential attachment et Common Neighbours respectivement. Pour k=4000, notre approche Diamant donne une valeur de précision de 94.88% comparée à ceux de l'approche Common Neighbours, Jaccards, Adamic/Adar et preferential attachment avec les valeurs de précision de 7.5%, 3.2%, 0.7% et 6.2% respectivement. Pour k=10000, la précision obtenue par notre approche Diamant est de 94.88% comparée à ceux de l'approche Common Neighbours, Jaccards, Adamic/Adar et preferential attachment avec les valeurs de précision de 6.0%, 4.4%, 1.6% et 6.9% respectivement.

Pour la F-mesure, qui représente le rapport entre le rappel et la précision, on constate que notre approche Diamant donne des meilleurs résultats par rapport aux autres approches. Le F-mesure de notre approche pour k=1000 arrive à 65.72% par rapport les autres approches qui ont des valeurs de 10.9%, 1.9% et 1.9% pour Common Neighbours, Jaccards et preferential attachment. Pour k=10000, la F-mesure de notre approche arrive jusqu'à 37.74% comparée à ceux de Common Neighbours, Jaccards, Adamic/Adar et preferential attachment avec une précision de 5.9%, 4.3%, 0.6% et 6.8% respectivement.

7.6 Conclusion

Dans ce chapitre nous avons présenté une application de l'agrégation textuelle dans le contexte OLAP au Social OLAP, en particulier au domaine des réseaux sociaux. Une nouvelle approche de prédiction des liens sémantiques dans les réseaux sociaux basée sur l'agrégation de contenu textuel est proposée. Nous avons présenté, dans un premier temps, les étapes d'application de notre approche, puis nous avons exposé les différents indicateurs utilisés pour mesurer la qualité de notre contribution. Ensuite, nous avons fait une description du réseau social sélectionné pour mener nos expérimentations. Les résultats de notre approche DIAMANT indiquent une nette amélioration dans les mesures de rappel, de précision et de F-mesure comparativement à ceux obtenus par les approches proposées dans la littérature. Cette amélioration est due à la prise en compte, dans DIAMANT, du contenu textuel introduit par les membres du réseau. Nous avons montré alors que les résultats de notre approche ont une meilleure

crédibilité et sont proche de la réalité. Enfin, notre approche peut être appliqué à d'autres jeux de données plus volumineux et avec un nombre de nœuds beaucoup plus important.

Chapitre 8

Conclusion et perspectives

8.1 Bilan et contributions

L'avènement des données complexes, a donné naissance à des nouveaux problèmes de recherche. Dans ce contexte, nous nous intéressons à l'adaptation du processus d'analyse en ligne des données complexes plus particulièrement aux données textuelles.

Le processus d'analyse en ligne OLAP offre aux décideurs un ensemble d'approches qui sert à faciliter la tâche de prise de décision. Son intérêt est de simplifier la manipulation des données structurées de façon multidimensionnelle à l'aide d'opérations de structuration, de granularité et de visualisation. Mais les constats faits sur l'augmentation de la masse des données complexes d'une part, et sur les limites des opérateurs OLAP classiques qui ne prennent pas en considération ce genre de données d'autre part, militent pour une évolution et une adaptation de l'OLAP. Cette évolution consiste en l'enrichissement de l'OLAP par des nouvelles approches adaptées aux données complexes par une analyse sémantique.

Pour répondre à cette problématique, nous avons choisi d'entamer le problème d'agrégation des données issues de documents textuels. Nous avons aussi essayé de prendre en considération la sémantique véhiculée par les données pour améliorer le processus décisionnel. Nos solutions sont basées sur la combinaison des points forts de l'OLAP, de la fouille de données, de la recherche d'information et de la théorie des graphes.

L'une des contributions résultantes de cette démarche de combinaison est la proposition d'un nouveau processus d'agrégation des données textuelles dans l'analyse OLAP, baptisé GOTA. Il tient compte de l'agrégation de l'information porteuse de sens qui représente le corpus de documents. GOTA se base sur les résultats obtenus par l'application des opérations de prétraitement et d'épuration des données textuelles. Notre apport, dans cette contribution, concerne plus particulièrement l'agrégation des mots-clés en se basant sur la technique K-means qui classifie les mots-clés selon la distance entre eux que nous avons mesuré en utilisant la distance de similarité de *Google* « Google Similarity distance ».

Une autre contribution établit une fonction permettant d'obtenir une vision synthétique des données textuelles analysées. Cette fonction, baptisée TAG permet d'agréger automatiquement un ensemble de mots-clefs en un ensemble réduit et plus synthétique. L'approche TAG exploite de la théorie des graphes pour la représentation et l'extraction des circuits les plus représentatifs du corpus.

Pour assurer des données textuelles de qualité, nous avons proposé un outil d'extraction des mots-clés issus du corpus de documents. Cet outil se base sur les données disponibles dans la plateforme Microsoft Research Academia. Son objectif est de déterminer les mots-clés qui sont utilisés par les fonctions d'agrégation proposées. Les résultats obtenus par l'outil d'extraction des mots-clés, sont utilisés comme données d'entrés dans notre plateforme OLAP-TAS dans laquelle nous avons implémenté six approches d'agrégation des données textuelles à des fins de comparaisons. Nous avons évalué les approches implémentées sur deux corpus réels d'articles scientifiques, le premier dans le domaine informatique, et le second dans le domaine médical. Nous avons testé la performance des approches, en utilisant comme mesures le rappel, la précision, la F-mesure et le temps d'exécution. Nos deux approches on donné des performances meilleures que les quatre autres approches.

Enfin, face au problème de la prédiction des liens dans les réseaux sociaux, qui est considéré comme un domaine d'application, nous avons proposé une approche baptisée Diamant, qui est une extension de l'approche TAG, et qui sert à faire de la prédiction des liens sémantiques entre les membres du réseau en exploitant les documents textuels qu'ils partagent. Nous avons comparé les performances de notre approche avec celles de cinq autres approches en utilisant une capture d'un réseau social réel. L'étude expérimentale montre la crédibilité de nos résultats et nous avons constaté que les performances de notre proposition sont nettement meilleures que celles des autres approches.

8.2 Perspectives

Les travaux réalisés dans cette thèse ouvrent diverses perspectives de recherche. Tout d'abord, nous continuons à croire que la combinaison de l'OLAP, la fouille de données et la RI est une solution adéquate pour l'analyse des données complexes. Nous proposons dans ce qui suit quelques perspectives à ces travaux pour améliorer davantage nos approches d'agrégation de données textuelles.

Les études expérimentales menées dans cette thèse se basent sur les documents

scientifiques. Ces documents, une fois écrits sont figés dans le temps. Cependant, deux contraintes que ne sont pas prises en considération : les versions des documents et leurs types. Par exemple, les pages web ont tendance à évoluer constamment. Il est naturel d'envisager la gestion des types et les versions de documents. Cela permet de proposer de nouvelles fonctions d'agrégation qui prennent en considération les données et les structures en même temps.

Nous avons testé la mise en œuvre de nos approches GOTA et TAG sur des corpus de 600 et 13000 documents et l'approche Diamant sur un réseau social de 1200 nœuds. La taille des corpus influe sur la performance des approches en termes de temps d'exécution et de qualité des résultats. Dans ce contexte, nous envisageons le passage à l'échelle par l'utilisation des corpus de taille plus importante pour évaluer l'ensemble des approches implémentées.

Au-delà de ces perspectives directes, nous envisageons plusieurs autres directions pour nos futurs travaux. Par exemple, il serait intéressant d'expérimenter les approches que nous avons proposées avec des données collectées autres que les publications scientifiques, chose que nous n'avons malheureusement pu faire jusqu'à présent du fait de la difficulté à obtenir des données à partir de réseaux sociaux connus comme Facebook par exemple. Cela nous permettrait d'étudier la généricité de nos propositions. Il serait par ailleurs intéressant de considérer la problématique de la prédiction des liens sémantiques comme une introduction pour résoudre des problèmes liés aux réseaux sociaux tels que la détection des communautés, la sélection des leaders et la détection des évènements.

Annexe A

Liste des publications

Nous avons régulièrement communiqué sur l'avancée de nos travaux au sein de différentes conférences et revues.

A.1 Revue internationale

BOUAKKAZ Mustapha, LOUDCHER Sabine, et OUITEN Youcef. OLAP textual aggregation approach using the Google similarity distance. International Journal of Business Intelligence and Data Mining, vol. 11, no 1, p. 31-48 (2016). DOI : http ://dx.doi.org/10.1504/IJBIDM.2016.076425

A.2 Conférences internationales

BOUAKKAZ Mustapha, LOUDCHER Sabine, et OUITEN Youcef. A New Tool for Textual Aggregation In Information Retrieval.International Conference on Enterprise Information Systems (ICEIS), Rome p. 7-12 (2016).

BOUAKKAZ Mustapha, LOUDCHER Sabine, et OUITEN Youcef. GOTA : Using the Google Similarity Distance for OLAP Textual Aggregation. International Conference on Enterprise Information Systems (ICEIS), Barcelona p. 24-30 (2015).

BOUAKKAZ Mustapha, LOUDCHER Sabine, et OUITEN Youcef. Automatic textual aggregation approach of scientific articles in OLAP context. Innovations in Information Technology (INNOVATIONS), 2014 10th International Conference on. IEEE p. 30-35 UAE (2014).

Bibliographie

[Abiteboul 2003] S. Abiteboul. *Managing an XML Warehouse in a P2P Context.* In Advanced Information Systems Engineering, pages 4–13, 2003. (Cité en page 16.)

[Abiteboul 2006] S. Abiteboul. *Entrepôts de contenu autour de XML et des services Web.* In EDA, pages 1–2, 2006. (Cité en page 16.)

[Adamic 2003] L. Adamic et E. Adar. *Friends and neighbors on the web.* In Social networks, volume 25, pages 211–230. Elsevier, 2003. (Cité en page 82.)

[Agrawal 1997] R. Agrawal, A. Gupta et S. Sarawagi. *Modeling multidimensional databases.* In ICDE, pages 232–243, 1997. (Cité en page 13.)

[Agrawal 2000] R. Agrawal, R. Bayardo et R. Srikant. *Athena : Mining based interactive management of text databases.* In Advances in Database Technology-EDBT 2000, pages 365–379. Springer, 2000. (Cité en page 24.)

[Aimé 2015] X. Aimé. *Eléments de réflexion sur l'utilisation de corpus pour la construction d'ontologies.* In IC2015, 2015. (Cité en page 32.)

[Aknouche 2013] R. Aknouche et A. Ounas. *Decisional architecture for text warehousing : ETL-text process and multidimensional model TWM.* In Proceedings of the 19th International Conference on Management of Data, pages 101–104. Computer Society of India, 2013. (Cité en page 40.)

[Aknouche 2014] R. Aknouche, O.Asfari et F. Bentayeb. *Integration Process for Multidimensional Textual Data Modeling.* In International Workshop in Software Evolution and Modernization SEM / ENASE, pages 119–126, 2014. (Cité en pages 3 et 16.)

[Alexa 2015] Alexa. *Top Ranking International Websites.* In Online accessed 1-Dec-2015, http ://www.alexa.com/, 2015. (Cité en page 74.)

[Allali 2011] O. Allali, C. Magnien et M. Latapy. *Link prediction in bipartite graphs using internal links and weighted projection.* In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on, pages 936–941. IEEE, 2011. (Cité en page 88.)

[Asfari 2013] O. Asfari, F. Bentayeb et N. Benblidia. *CXT-cube : contextual text cube model and aggregation operator for text OLAP.* In Proceedings of the sixteenth

international workshop on Data warehousing and OLAP, pages 27–32. ACM, 2013. (Cité en pages 29 et 36.)

[Azabou 2015] M. Azabou, K. Khrouf, C. Soule-Dupuy et N. Valless. *Diamond multi-dimensional model and aggregation operators for document OLAP*. In Research Challenges in Information Science (RCIS), 2015 IEEE 9th International Conference on, pages 363–373. IEEE, 2015. (Cité en page 40.)

[Baeza-Yates 1999] R. Baeza-Yates et B. Ribeiro-Neto. Modern information retrieval, volume 463. ACM press New York, 1999. (Cité en page 17.)

[Baten 2014] K. Baten et P. Hadermann. *Le syntagme verbal en FLE : complexité, variation, systématicité*. In CAHIER AFLS, volume 19, pages 23–56, 2014. (Cité en page 30.)

[Béchet 2009] N. Béchet. *Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc, 2009. (Cité en pages 30, 32 et 36.)

[Bechikh 2013] A. Bechikh et H. Haddad. *Extraction et filtrage de syntagmes nominaux pour la Recherche d'Information*. In EGC, pages 247–252, 2013. (Cité en page 30.)

[Benamara 2007] F. Benamara, C. Cesarano, A. Picariello, D. Recupero et V. Subrahmanian. *Sentiment Analysis : Adjectives and Adverbs are better than Adjectives Alone*. In ICWSM. Citeseer, 2007. (Cité en pages 30 et 36.)

[Benatallah 2012] B. Benatallah et H. Nezhad. *A framework and a language for online analytical processing on graphs*. In International Conference on Web Information Systems Engineering, pages 213–227. Springer, 2012. (Cité en page 93.)

[BenMessaoud 2004] R. BenMessaoud et S. Loudcher. *OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données*. In 4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04), pages 35–46, 2004. (Cité en pages 3, 4, 27 et 36.)

[BenMessaoud 2006] R. BenMessaoud et S. Loudcher. *A data mining-based OLAP aggregation of complex data : Application on XML documents*. In International booktitle of Data Warehousing and Mining (IJDWM), volume 2, pages 1–26. IGI Global, 2006. (Cité en page 42.)

[Bhide 2016] M. Bhide, S. Mittapalli et S. Padmanabhan. *Star and snowflake schemas in extract, transform, load processes.* Google Patents, 2016. US Patent 9,298,787. (Cité en page 11.)

[Bilhaut 2007] F. Bilhaut, D. Franck et E. Patrice. *Indexation sémantique et recherche d'information interactive.* In CORIA, volume 7, pages 65–76, 2007. (Cité en page 17.)

[Bimonte 2006] S. Bimonte, A. Tchounikine et M Miquel. *Geocube, a multidimensional model and navigation operators handling complex measures : Application in spatial olap.* In International Conference on Advances in Information Systems, pages 100–109. Springer, 2006. (Cité en page 13.)

[Bollobás 2013] B. Bollobás. Modern graph theory, volume 184. Springer Science and Business Media, 2013. (Cité en pages 48, 49 et 51.)

[Bothorel 2015] C. Bothorel, J. Cruz, M. Magnani et B. Micenkova. *Clustering attributed graphs : models, measures and methods.* In Network Science, volume 3, pages 408–444. Cambridge Univ Press, 2015. (Cité en page 91.)

[Bouakkaz 2014] M. Bouakkaz, S. Loudcher et Y. Ouinten. *Automatic textual aggregation approach of scientific articles in OLAP context.* In The 10th International Conference on Innovations in Information Technology, pages 30–35. IEEE, 2014. (Cité en pages 5, 6 et 40.)

[Bouakkaz 2015] M. Bouakkaz, S. Loudcher et Y. Ouiten. *GOTA : Using the Google Similarity Distance for OLAP Textual Aggregation.* In International Conference on Enterprise Information Systems (ICEIS), pages 24–30, 2015. (Cité en pages 5, 6 et 39.)

[Bouakkaz 2016a] M. Bouakkaz, S. Loudcher et Y. Ouinten. *OLAP textual aggregation approach using the Google similarity distance.* In International booktitle of Business Intelligence and Data Mining, volume 11, pages 31–48, 2016. (Cité en pages 4, 5 et 6.)

[Bouakkaz 2016b] M. Bouakkaz, S. Loudcher et Y. Ouiten. *A New Tool for Textual Aggregation In Information Retrieval.* In International Conference on Enterprise Information Systems (ICEIS), pages 7–12, 2016. (Cité en pages 5 et 6.)

[Boubekeur 2008] F. Boubekeur. *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets.* PhD thesis, Université Paul Sabatier-Toulouse III, 2008. (Cité en page 65.)

[Bougouin 2013] A. Bougouin. *État de l'art des méthodes d'extraction automatique de termes-clés*. In Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), 2013. (Cité en page 59.)

[Boukraâ 2010] D. Boukraâ, O. Boussaïd et F. Bentayeb. *Olap operators for complex object data cubes*. In East European Conference on Advances in Databases and Information Systems, pages 103–116. Springer, 2010. (Cité en pages 13 et 25.)

[Boyd 2008] D. Boyd. Taken out of context : American teen sociality in networked publics. ProQuest, 2008. (Cité en page 74.)

[Bringay 2010] S. Bringay, A. Laurent, P. Poncelet, M. Roche et M. Teisseire. *Bien cube, les données textuelles peuvent s' agréger!* In 10èmes journées d'Extraction et Gestion des Connaissances, numéro 19, pages 585–596, 2010. (Cité en pages 28, 36, 65 et 71.)

[Campos 1] R. Campos, G. Dias, A. Jorge et A. Jatowt. *Survey of temporal information retrieval and related applications*. In ACM Computing Surveys (CSUR), volume 47, page 15. ACM, 1. (Cité en pages 15 et 22.)

[Capocci 2006] A. Capocci, V. Servedio, F. Colaiori et L. Buriol. *Preferential attachment in the growth of social networks*. In Physical Review E, volume 74, page 036116. APS, 2006. (Cité en page 82.)

[Chang 2012] Y. Chang et K. Hung-Yu. *Link prediction in a bipartite network using Wikipedia revision information*. In 2012 Conference on Technologies and Applications of Artificial Intelligence, pages 50–55. IEEE, 2012. (Cité en page 88.)

[Chaudiron 2004] Stéphane Chaudiron et Madjid Ihadjadene. *Évaluer les systèmes de recherche d'information : Nouveaux modèles de l'utilisateur*. In Hermès , 2004, 39, fascicule thématique" Critique de la raison numérique". CNRS Editions, Paris (FRA), 2004. (Cité en page 64.)

[Chawla 2012] N. Chawla et R. Lichtnwalter. *Link prediction : fair and effective evaluation*. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pages 376–383. IEEE Computer Society, 2012. (Cité en page 100.)

[Chelmis 2013] C. Chelmis et V. Prasanna. *Social link prediction in online social tagging systems*. In ACM Transactions on Information Systems (TOIS), volume 31, page 20. ACM, 2013. (Cité en page 85.)

[Chen 2008] C. Chen, X. Yan et F. Zhu. *Graph OLAP : Towards online analytical processing on graphs.* In 2008 Eighth IEEE International Conference on Data Mining, pages 103–112. IEEE, 2008. (Cité en pages 91 et 92.)

[Chowdhury 1984] G. Chowdhury. *Introduction to modern information retrieval.* Facet publishing, 1984. (Cité en page 17.)

[Chrisment 2005] C. Chrisment, F. Ravat et O. Teste. *Les entrepots de donnÃ©es.* In Techniques de l'ingenieur, 2005. (Cité en page 11.)

[Chun 2012] H. Chun et S. Moon. *Fragile Online Relationship : A First Look at Unfollow Dynamics in Twitter.* In HCI 2012, pages 1694–1694, 2012. (Cité en page 89.)

[Cilibrasi 2007] R. Cilibrasi et P. Vitanyi. *The google similarity distance.* In IEEE Transactions on Knowledge and Data Engineering, volume 19, pages 370–383, 2007. (Cité en page 43.)

[Codd 1972] E. Codd. Relational completeness of data base sublanguages. IBM Corporation, 1972. (Cité en page 24.)

[Codd 1993] E. Codd, S. Codd et C. Salley. *Providing OLAP to user analyst : an IT mandate.* In Rapport technique, 1993. (Cité en pages 1 et 9.)

[Daille 2000] B. Daille, S. Barreaux et F. Boudin. *Indexation d'articles scientifiques Presentation et resultats du defi fouille de textes DEFT 2016.* In PARIS Inalco du 4 au 8 juillet 2000, page 1, 2000. (Cité en page 19.)

[Daqing 2016] H. Daqing et W. Jeng. *Scholarly Collaboration on the Academic Social Web.* In Synthesis lectures on information Concepts, Retrieval, and Services, volume 8, pages 22–37. Morgan, 2016. (Cité en page 100.)

[Dehkordi 2013] M. Dehkordi. *A novel association rule hiding approach in OLAP data cubes.* In Indian booktitle of Science and Technology, volume 6, pages 4063–4075, 2013. (Cité en page 42.)

[Deo 2016] Narsingh Deo. Graph theory with applications to engineering and computer science. Courier Dover Publications, 2016. (Cité en page 51.)

[Dietzel 2011] S. Dietzel, F. Kargl, G. Heijenk et F. Schaub. *Modeling in-network aggregation in VANETs.* In IEEE Communications Magazine, volume 49, pages 142–148, 2011. (Cité en page 25.)

[Dietzel 2014] S. Dietzel, J. Petit, F. Kargl et B. Scheuermann. *In-network aggregation for vehicular ad hoc networks.* In Communications Surveys and Tutorials, IEEE, volume 16, pages 1909–1932. IEEE, 2014. (Cité en page 25.)

[Donna 2000] H. Donna. *What we have learned, and not learned, from TREC.* In Proc. of the BCS IRSG, pages 2–20, 2000. (Cité en page 20.)

[Du 2010] N. Du, H. Wang et C. Faloutsos. *Analysis of large multi-modal social networks : patterns and a generator.* In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 393–408. Springer, 2010. (Cité en page 91.)

[Dumais 2004] S. Dumais. *Latent semantic analysis.* In Annual review of information science and technology, volume 38, pages 188–230. Wiley Online Library, 2004. (Cité en page 34.)

[Dunlavy 2011] D. Dunlavy, T. Kolda et E. Acar. *Temporal link prediction using matrix and tensor factorizations.* In ACM Transactions on Knowledge Discovery from Data (TKDD), volume 5, page 10. ACM, 2011. (Cité en page 87.)

[Easley 2010] D. Easley et J. Kleinberg. Networks, crowds, and markets : Reasoning about a highly connected world. Cambridge University Press, 2010. (Cité en page 75.)

[Ellison 2013] N. Ellison et D. Boyd. *Sociality through Social Network Sites.* Citeseer, 2013. (Cité en page 74.)

[Etcheverry 2012] L. Etcheverry et A. Vaisman. *Enhancing OLAP analysis with web cubes.* In The Semantic Web : Research and Applications, pages 469–483. Springer, 2012. (Cité en page 25.)

[Fangbo 2013] T. Fangbo, K. LeiHou, J. Han et C. Zhai. *EventCube : multi-dimensional search and mining of structured and text data.* In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1494–1497. ACM, 2013. (Cité en page 42.)

[Fuglede 2004] B. Fuglede et F. Topsoe. *Jensen-Shannon divergence and Hilbert space embedding.* In IEEE International Symposium on Information Theory, pages 31–31, 2004. (Cité en page 33.)

[Fuhr 2001] N. Fuhr. *XIRQL : A query language for information retrieval in XML documents.* In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 172–180, 2001. (Cité en page 16.)

[Getoor 2007] L. Getoor. Introduction to statistical relational learning. MIT press, 2007. (Cité en page 91.)

[Golfarelli 2002] M. Golfarelli, S. Rizzi et E. Saltarelli. *WAND : A CASE Tool for Workload-Based Design of a Data Mart.* In Decimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati, pages 422–426, 2002. (Cité en page 39.)

[Grossman 2012] D. Grossman et O. Frieder. Information retrieval : Algorithms and heuristics, volume 15. Springer Science and Business Media, 2012. (Cité en page 17.)

[Gupta 2015] A. Gupta et N. Sardana. *Significance of Clustering Coefficient over Jaccard Index.* In Contemporary Computing (IC3), 2015 Eighth International Conference on, pages 463–466. IEEE, 2015. (Cité en page 81.)

[Gyssens 1997] M. Gyssens et L. Lakshmanan. *A foundation for multi-dimensional databases.* In VLDB, volume 97, pages 106–115, 1997. (Cité en page 24.)

[Han 1997] Jiawei Han. *OLAP Mining : An Integration of OLAP with Data Mining.* In In Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7, pages 1–9, 1997. (Cité en page 42.)

[Hassan 2013] Ali Hassan, Franck Ravat, Olivier Teste, Ronan Tournier et Gilles Zurfluh. *Opérateurs OLAP dans les bases de données multidimensionnelles multifonctions.* In EDA, pages 69–78, 2013. (Cité en pages 2 et 13.)

[Hofmann 1999] T. Hofmann. *Probabilistic latent semantic indexing.* In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57. ACM, 1999. (Cité en page 28.)

[Inmon 1996] W. Inmon. *Building the Data Warehouse.* In John Wiley and Sons, volume ISBN : 0764599445, 1996. (Cité en pages 1 et 8.)

[Jadav 2012] J. Jadav et M. Panchal. *Association rule mining method on OLAP cube.* In International booktitle of Engineering Research and Applications (IJERA), volume 2, pages 1147–1151. Citeseer, 2012. (Cité en page 42.)

[Jalam 2002] R. Jalam et J. Chauchat. *Pourquoi les n-grammes permettent de classer des textes Recherche de mots-clefs pertinents a l'aide des n-grammes caracteristiques.* In 6th International Conference on Textual Data Statistical Analysis, France, pages 381–390, 2002. (Cité en page 19.)

[Jeh 2002] G. Jeh et J. Widom. *SimRank : a measure of structural-context similarity.* In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 538–543. ACM, 2002. (Cité en page 85.)

[Jensen 2004] D. Jensen, J. Neville et B. Gallagher. *Why collective inference improves relational classification.* In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 593–598. ACM, 2004. (Cité en page 91.)

[Kimball 1996] R. Kimball. *The data warehouse toolkit : Practical techniques for building dimentionnel data warehouses.* In John Wiley, 1996. (Cité en pages 10 et 11.)

[Kister 2012] L. Kister et E. Jacquey. *Relations syntaxiques entre lexiques terminologique et transdisciplinaire : analyse en texte intégral.* In SHS Web of Conferences, volume 1, pages 909–919. EDP Sciences, 2012. (Cité en page 31.)

[Klug 1982] A. Klug. *Equivalence of relational algebra and relational calculus query languages having aggregate functions.* In booktitle of the ACM (JACM), volume 29, pages 699–717. ACM, 1982. (Cité en page 24.)

[Kohomban 2007] U. Kohomban et W. Lee. *Optimizing classifier performance in word sense disambiguation by redefining word sense classes.* In Proceedings of the International Joint Conference on Artificial Intelligence, pages 1635–1640, 2007. (Cité en page 30.)

[Kompaoré 2008] N. Kompaoré. *Fusion de systemes et analyse des caractéristiques linguistiques des requêtes : vers un processus de RI adaptatif.* PhD thesis, Université Paul Sabatier-Toulouse III, 2008. (Cité en page 64.)

[Kou 2015] G. Kou et Y. Peng. *An application of latent semantic analysis for text categorization.* In International booktitle of Computers Communications and Control, volume 10, pages 357–369, 2015. (Cité en page 33.)

[Kumar 2008] N. Kumar, L. Zhang et S. Nayar. *What is a good nearest neighbors algorithm ?* In European conference on computer vision, pages 364–378. Springer, 2008. (Cité en page 81.)

[Kunegis 2010] J. Kunegis, E. DeLuca et S. Albayrak. *The link prediction problem in bipartite networks.* In International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, pages 380–389. Springer, 2010. (Cité en page 88.)

[Kwak 2012] H. Kwak, S. Moon et L. Wonjae. *More of a Receiver Than a Giver : Why Do People Unfollow in Twitter ?* In ICWSM, 2012. (Cité en page 89.)

[Laurent 2002] A. Laurent. *Bases de donnees multidimensionnelles floues et leur utilisation pour la fouil le de donnees.* PhD thesis, Universite de Paris6, 2002. (Cité en page 13.)

[Lauw 2007] H. Lauw, E. Lim et H. Pang. *TUBE (Text-cUBE) for discovering documentary evidence of associations among entities.* In Proceedings of the 2007 ACM symposium on Applied computing, pages 824–828. ACM, 2007. (Cité en pages 27, 36, 65 et 71.)

[Lee-Hoon 2006] S. Lee-Hoon, K. Pan-Jun et J. Hawoong. *Statistical properties of sampled networks.* In Physical Review E, volume 73, page 016102. APS, 2006. (Cité en page 79.)

[Lefrançois 2014] M. Lefrançois, F. Gandon et A. Giboin. *Méthodologie d'ingénierie des connaissances pour la représentation des définitions lexicographiques dans le cadre de la théorie Sens-Texte.* In TOTh-8th International Conference on Terminology and Ontology : Theories and applications. Institut Porphyre, Savoir et Connaissance, 2014. (Cité en page 30.)

[Leicht 2006] E. Leicht, P. Holme et M. Newman. *Vertex similarity in networks.* In Physical Review E, volume 73, page 026120. APS, 2006. (Cité en page 81.)

[Liben-Nowell 2007] D. Liben-Nowell et J. Kleinberg. *The link-prediction problem for social networks.* In booktitle of the American society for information science and technology, volume 58, pages 1019–1031. Wiley Online Library, 2007. (Cité en pages 81 et 82.)

[Lim 2010] A. Lim et C. Lee. *Processing online analytics with classification and association rule mining.* In Knowledge-Based Systems, volume 23, pages 248–255. Elsevier, 2010. (Cité en page 42.)

[Lin 2008] X. Lin, B. Ding, J. Han, F. Zhu et B. Zhao. *Text cube : Computing ir measures for multidimensional text database analysis.* In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, pages 905–910. IEEE, 2008. (Cité en pages 28 et 36.)

[Linyuan 2009] L. Linyuan, J. Ci-Hang et T. Zhou. *Similarity index based on local paths for link prediction of complex networks.* In Physical Review E, volume 80, page 046122. APS, 2009. (Cité en pages 84 et 85.)

[Longrée 2016] D. Longrée. *De la lemmatisation et l'annotation morphosyntaxique des textes latins á l'exploitation statistique des données : les méthodes informatiques du LASLA.* 2016. (Cité en page 19.)

[Loudcher 2015] S. Loudcher, W. Jakawat et E. Morales. *Combining OLAP and information networks for bibliographic data analysis : a survey.* In Scientometrics, volume 103, pages 471–487. Springer, 2015. (Cité en pages 91 et 92.)

[Lü 2011] L. Lü et T. Zhou. *Link prediction in complex networks : A survey.* In Physica A : Statistical Mechanics and its Applications, volume 390, pages 1150–1170. Elsevier, 2011. (Cité en page 80.)

[Madden 2013] M. Madden, A. Lenhart et S.Cortesi. *Teens, social media, and privacy.* In Pew Research Center, volume 21, 2013. (Cité en page 74.)

[Malhotra 2015] A. Malhotra, M. Gündel, A. Rajput, H. Mevissen, A. Saiz et X. Pastor. *Knowledge retrieval from pubmed abstracts and electronic medical records with the multiple sclerosis ontology.* In PloS one, volume 10, page e0116718. Public Library of Science, 2015. (Cité en page 32.)

[Malinowski 2004] E. Malinowski et E. Zimanyi. *OLAP hierarchies : A conceptual perspective.* In Advanced Information Systems Engineering, pages 477–491. Springer, 2004. (Cité en page 25.)

[Mansmann 2006] S. Mansmann et M. Scholl. Extending visual olap for handling irregular dimensional hierarchies. Springer, 2006. (Cité en page 25.)

[Milen 2007] P. Milen et I. Ryutaro. *Finding experts by link prediction in co-authorship networks.* In International conference on finding Experts on the web with semantics, volume 290, pages 42–55. CEUR-WS, 2007. (Cité en page 100.)

[Mohanty 2012] S. Mohanty et J. Debasish. *Secure data aggregation in vehicular-adhoc networks : A survey.* In Procedia Technology, volume 6, pages 922–929. Elsevier, 2012. (Cité en page 25.)

[Morfonios 2008] K. Morfonios et G. Koutrika. *OLAP Cubes for Social Searches : Standing on the Shoulders of Giants ?* In WebDB. Citeseer, 2008. (Cité en page 91.)

[Mothe 2003] J. Mothe, C. Chrisment, B. Dousset et J. Alaux. *DocCube : Multi-dimensional visualisation and exploration of large document sets.* In booktitle of the American Society for Information Science and Technology, volume 54, pages 650–659. Wiley Online Library, 2003. (Cité en pages 26 et 36.)

[Munasinghe 2013] L. Munasinghe et R. Ichise. *Link prediction in social networks using information flow via active links.* In IEICE TRANSACTIONS on Infor-

mation and Systems, volume 96, pages 1495–1502. The Institute of Electronics, Information and Communication Engineers, 2013. (Cité en page 87.)

[Nashi 2010] X. Nashi, L. Xumin et G. Yong. *Research on k-means clustering algorithm : An improved k-means clustering algorithm.* In Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on, pages 63–67. IEEE, 2010. (Cité en page 43.)

[O'Madadhain 2005] J. O'Madadhain, J. Hutchins et P. Smyth. *Prediction and ranking algorithms for event-based network data.* In ACM SIGKDD Explorations Newsletter, volume 7, pages 23–30. ACM, 2005. (Cité en page 87.)

[Ozsoyoglu 1987] G. Ozsoyoglu et V. Matos. *Extending relational algebra and relational calculus with set-valued attributes and aggregate functions.* In ACM Transactions on Database Systems (TODS), volume 12, pages 566–592. ACM, 1987. (Cité en page 24.)

[Palpanas 2000] T. Palpanas. *Knowledge discovery in data warehouses.* In ACM Sigmod Record, volume 29, pages 88–100. ACM, 2000. (Cité en page 42.)

[Papadimitriou 2011] A. Papadimitriou, P. Symeonidis et Y. Manolopoulos. *Friendlink : link prediction in social networks via bounded local path traversal.* In Computational Aspects of Social Networks (CASoN), 2011 International Conference on, pages 66–71. IEEE, 2011. (Cité en page 82.)

[Papadimitriou 2012] A. Papadimitriou, S Panagiotis et Y Manolopoulos. *Fast and accurate link prediction in social networking systems.* In booktitle of Systems and Software, volume 85, pages 2119–2132. Elsevier, 2012. (Cité en page 85.)

[Pedersen 2006] T. Pedersen, C. Jensen et C. Dyreson. *Method and systems for making OLAP hierarchies summarisable*, 2006. US Patent 7,133,865. (Cité en page 25.)

[Pérez 2007] J. Pérez, R. Berlanga, M. Aramburu et T. Pedersen. *R-cubes : OLAP cubes contextualized with documents.* In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, pages 1477–1478. IEEE, 2007. (Cité en pages 27 et 36.)

[Porhiel 2013] S. Porhiel. *Le détachement en position initiale : rôle phrastique ou discursif/textuel ? Exemple du syntagme à propos de X.* In Linguistik online, volume 26, 2013. (Cité en page 31.)

[Poudat 2006] C. Poudat, G. Cleuziou et V. Clavier. *Catégorisation de textes en domaines et genres*. In Document numérique, volume 9, pages 61–76. Lavoisier, 2006. (Cité en pages 30 et 36.)

[Qiang 2011] Q. Qiang, F. Zhu, X. Yan et J. Han. *Efficient topological OLAP on information networks*. In International Conference on Database Systems for Advanced Applications, pages 389–403. Springer, 2011. (Cité en page 92.)

[Ramos 2003] J. Ramos. *Using tf-idf to determine word relevance in document queries*. In Proceedings of the first instructional conference on machine learning, 2003. (Cité en page 19.)

[Ravat 2007] F. Ravat, O. Teste et R. Tournier. *Querying Multidimensional Databases*. In Advances in Databases and Information Systems, pages 298–313, 2007. (Cité en pages 9, 11, 32 et 36.)

[Ravat 2008] F. Ravat, O. Teste et R. Tournier. *Algebraic and graphic languages for OLAP manipulations*. In Int booktitle of Data Warehousing and Mining, pages 17–46, 2008. (Cité en pages 10, 33, 36, 65 et 71.)

[Redkar 2015] H. Redkar, S. Bhingardive, D. Kanojia et P. Bhattacharyya. *World WordNet Database Structure : An Efficient Schema for Storing Information of WordNets of the World*. In AAAI, pages 4290–4291, 2015. (Cité en page 32.)

[Rijsbergen 2000] V. Rijsbergen et C. Keith. *Getting into information retrieval*. In Lectures on information retrieval, pages 1–20. 2000. (Cité en page 22.)

[Rizzi 2006] S. Rizzi, A. Alberto et J. Lechtenborger. *Research in data warehouse modeling and design : dead or alive ?* In Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, pages 3–10. ACM, 2006. (Cité en page 39.)

[Rodriguez 2015] R. Rodriguez, L. Martinez et F. Herrera. *A linguistic 2-tuple multicriteria decision making model dealing with hesitant linguistic information*. In Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, pages 1–7. IEEE, 2015. (Cité en pages 31 et 36.)

[Rossetti 2011] G. Rossetti, M. Berlingerio et F. Giannotti. *Scalable link prediction on multidimensional networks*. In 2011 IEEE 11th International Conference on Data Mining Workshops, pages 979–986. IEEE, 2011. (Cité en page 91.)

[Rubin 2008] D. Rubin, N. Shah et N. Noy. *Biomedical ontologies : a functional perspective*. In Briefings in bioinformatics, volume 9, pages 75–90. Oxford Univ Press, 2008. (Cité en page 33.)

[Sauvagnat 2005] K. Sauvagnat. *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*. PhD thesis, Toulouse 3, 2005. (Cité en page 20.)

[Shen 2015] Y. Shen, W. Rong, Z. Sun, Y. Ouyang et X. Zhang. *Question/Answer Matching for CQA System via Combining Lexical and Sequential Information*. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. (Cité en pages 31, 32 et 36.)

[Shoshani 1997] A. Shoshani. *OLAP and statistical databases : Similarities and differences*. In Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, pages 185–196. ACM, 1997. (Cité en page 24.)

[Shoshani 2003] A. Shoshani. *Multidimensionality in statistical, OLAP, and scientific databases*. In Chapitre II, Multidimensional Databases : Problems and Solutions, Maurizio Rafanelli (Ed.), Idea Publishing Group (IGP), pages 46–68, 2003. (Cité en page 25.)

[Silva 2012] R. Silva, J. Moura-Pires et M. Santos. *Spatial clustering in solap systems to enhance map visualization*. In International booktitle of Data Warehousing and Mining (IJDWM), volume 8, pages 23–43. IGI Global, 2012. (Cité en page 25.)

[Sullivan 2001] D. Sullivan. Document warehousing and text mining : techniques for improving business operations, marketing, and sales. 2001. (Cité en pages 15 et 16.)

[Sun 2012] A. Sun. *Short text classification using very few words*. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pages 1145–1146. ACM, 2012. (Cité en page 31.)

[Szymanski 2015] Terrence Szymanski et Gerard Lynch. *UCD : Diachronic Text Classification with Character, Word, and Syntactic N-grams*. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). United States, 2015. (Cité en page 31.)

[Tamine 2000] L. Tamine. *Optimisation de requetes dans un systeme de recherche d'information approche basee sur l'exploitation de techniques avancees de l'algorithmique generique*. 2000. (Cité en page 21.)

[Tian 2008] Y. Tian, R. Hankins et J. Patel. *Efficient aggregation for graph summarization.* In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 567–580. ACM, 2008. (Cité en page 91.)

[Torlone 2003] R. Torlone. *Multidimensionality in statistical, OLAP, and scientific databases.* In Chapitre II, Multidimensional Databases : Problems and Solutions, Maurizio Rafanelli (Ed.), Idea Publishing Group (IGP), pages 69–90, 2003. (Cité en page 25.)

[Torres 2014] M. Torres, J. Samos et E. Garví. *Closing Ontologies to Define OLAP Systems.* In International booktitle of Information Retrieval Research (IJIRR), volume 4, pages 1–16. IGI Global, 2014. (Cité en page 32.)

[Tournier 2008] R. Tournier. *Analyse en ligne (OLAP) de documents.* PhD thesis, Universite Paul Sabatier Toulouse 3, 2008. (Cité en pages 3, 16 et 26.)

[Tseng 2006] F. Tseng et C. Annie. *The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence.* In Decision Support Systems, volume 42, pages 727–744, 2006. (Cité en page 15.)

[Vassiliadis 1999] P. Vassiliadis et T. Sellis. *A survey of logical models for olap databases.* In SIGMOD, pages 64–69, 1999. (Cité en page 13.)

[Verma 2007] R. Verma, P. Chen et W. Lu. *A semantic free-text summarization system using ontology knowledge.* In Proc. of Document Understanding Conference, 2007. (Cité en pages 32 et 36.)

[Viswanathan 2011] G. Viswanathan et M. Schneider. *On the requirements for user-centric spatial data warehousing and SOLAP.* In Database Systems for Adanced Applications, pages 144–155. Springer, 2011. (Cité en page 25.)

[Wang 2015] P. Wang, B. Xu, W. Wu et X. Zhou. *Link prediction in social networks : the state-of-the-art.* In Science China Information Sciences, volume 58, pages 1–38. Springer, 2015. (Cité en page 82.)

[Wartena 2008] C. Wartena et R. Brussee. *Topic detection by clustering keywords.* In Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on, pages 54–58. IEEE, 2008. (Cité en pages 33, 36, 65 et 71.)

[Xia 2012] S. Xia, D. BingTian, Ee-Peng L et Y. Zhang. *Link prediction for bipartite social networks : the Role of Structural Holes.* In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, pages 153–157. IEEE, 2012. (Cité en page 88.)

[Xu 2012] Y. Xu, J. Merigo et H. Wang. *Linguistic power aggregation operators and their application to multiple attribute group decision making.* In Applied Mathematical Modelling, volume 36, pages 5427–5444. Elsevier, 2012. (Cité en page 31.)

[Yang 2015] Y. Yang, R. Lichtenwalter et N. Chawla. *Evaluating link prediction methods.* In Knowledge and Information Systems, volume 45, pages 751–782. Springer, 2015. (Cité en page 100.)

[Yin 2012] M. Yin, B. Wu et Z. Zeng. *HMGraph OLAP : a novel framework for multi-dimensional heterogeneous network analysis.* In Proceedings of the fifteenth international workshop on Data warehousing and OLAP, pages 137–144. ACM, 2012. (Cité en page 92.)

[Yu 2009] Y. Yu, C. Lin, Y. Sun, C. Chen, J. Han, B. Liao, T. Wu, C. Zhai, D. Zhang et B. Zhao. *iNextCube : Information network-enhanced text cube.* In Proceedings of the VLDB Endowment, volume 2, pages 1622–1625. VLDB Endowment, 2009. (Cité en pages 28 et 36.)

[Zargayouna 2004] H. Zargayouna. *Contexte et sémantique pour une indexation de documents semi-structurés.* In CORIA, volume 4, pages 161–177, 2004. (Cité en page 18.)

[Zhang 2009] D. Zhang, C. Zhai, J. Han, A. Srivastava et N. Oza. *Topic modeling for OLAP on multidimensional text databases : topic cube and its applications.* In Statistical Analysis and Data Mining, volume 2, pages 378–395. Wiley Online Library, 2009. (Cité en pages 28 et 36.)

[Zhao 2011] X. Zhao, X. Li et D. Xin. *Graph cube : on warehousing and OLAP multidimensional networks.* In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pages 853–864. ACM, 2011. (Cité en page 92.)

[Zhou 2009] T. Zhou, L. Linyuan et Y. Zhang. *Predicting missing links via local information.* In The European Physical booktitle B, volume 71, pages 623–630. Springer, 2009. (Cité en page 82.)