Que peut apporter l'OLAP à l'analyse de réseaux d'informations bibliographiques ?

Sabine Loudcher, Cécile Favre, Wararat Jakawat

Que peut apporter l'OLAP à l'analyse de réseaux d'informations bibliographiques ?

Sabine LOUDCHER — Cécile FAVRE — Wararat JAKAWAT

Université de Lyon (Laboratoire ERIC), France
{sabine.loudcher, cecile.favre, wararat.jakawat}@univ-lyon2.fr

RÉSUMÉ. *Dans le contexte de la bibliométrie, les données peuvent être considérées comme des réseaux hétérogènes d'informations. La construction et l'analyse de tels réseaux posent de nombreux problèmes. A la question de comment synthétiser ou agréger un réseau, l'analyse OLAP peut apporter une réponse car l'agrégation et la visualisation sont deux points centraux de ce type d'analyse. Nous proposons alors dans cet article un nouveau cadre pour construire, stocker et analyser plusieurs réseaux d'informations et leur dynamique, tout en combinant l'OLAP et la fouille de données. Les deux idées principales sont d'une part d'être capable de construire et d'analyser plusieurs réseaux issus des mêmes données bibliographiques (réseau des auteurs, des citations, des conférences, ces réseaux traduisent plusieurs points de vue); d'autre part de combiner l'analyse OLAP à la fouille de données et à la fouille de textes pour enrichir les possibilités d'analyse.*

ABSTRACT. *In the context of bibliometrics, data can be considered as heterogeneous information networks. The construction and the analysis of such networks pose many problems. To the question of how to synthesize or aggregate a network, OLAP analysis can provide an answer since the aggregation and the visualization are two central points of this type of analysis. Then we propose in this paper a new framework to build, store and analyze several information networks and their dynamics, while combining OLAP and data mining. The main two ideas are firstly to be able to construct and analyze multiple networks from the same bibliographic data (network of authors, citations, conferences, these networks reflect different points of view); secondly to combine OLAP analysis with data mining and text mining to enhance analysis capabilities.*

MOTS-CLÉS : *OLAP, entrepôts de données, fouille de données, réseaux d'informations, données bibliographiques.*

KEYWORDS: *OLAP, data warehouse, data mining, information networks, bibliographic data*

1. Introduction

Les systèmes de communication ou de transports, les médias et réseaux sociaux comme Facebook, Twitter, LinkedIn, les wikis, les blogs, etc. sont devenus omniprésents sur le Web et ont fait évoluer le Web vers une nouvelle dimension : le Web social. Nous y trouvons également des plateformes dédiées plus particulièrement aux chercheurs, telles que Mendeley ou ResearchGate. Tous ces systèmes sont des réseaux d'informations composés d'un très grand nombre d'objets interconnectés. Ces réseaux sont souvent modélisés par des grands graphes où les sommets ou nœuds représentent des entités (comme des internautes, des ressources, des annotations, des documents, etc.) et les arêtes décrivent les liens ou les relations existant entre ces entités. Il peut y avoir plusieurs types d'entités et de liens, on parle alors de réseaux d'informations hétérogènes (Han, 2009). En plus de la structure topologique encodée par le graphe, des attributs peuvent être associés aux sommets ou aux arêtes, ce qui forme les réseaux multidimensionnels (Zhao *et al.*, 2011).

Parallèlement, depuis les années 50, s'est développée dans le monde académique, une discipline, appelée bibliométrie, visant à étudier l'activité scientifique au travers des publications au moyen des mathématiques et de méthodes quantitatives. Cette discipline trouve sa source dans l'intersection de différents domaines de recherche tels que la sociologie et l'histoire des sciences, les sciences de l'information, la linguistique computationnelle, mais aussi la statistique et l'analyse des réseaux, etc. Elle s'intéresse à l'ensemble du processus de l'activité scientifique, à la fois du point de vue de la production scientifique des auteurs, mais aussi du processus de sélection, des pratiques de citation, etc. L'analyse de réseaux contribue donc fortement aux recherches de cette discipline, étant donné les relations existantes entre auteurs (via la co-publication), entre publications (via le processus de citation), entre communautés thématiques, etc. L'analyse de tels réseaux permet de découvrir des connaissances comme la détection de communautés d'auteurs ou d'institutions, l'extraction des thèmes sur lesquels portent les publications, un ordonnancement des auteurs ou des conférences, etc. L'avènement du Web a facilité la mise à disposition de différentes bases de données comme DBLP, ACM, PubMed, NCBI, etc. Une masse importante de données est donc disponible.

Pour permettre l'analyse de cette masse d'informations, de nombreux travaux se sont intéressés à l'analyse des graphes complexes et à l'apprentissage statistique relationnel (Jensen *et al.*, 2004; Getoor *et al.*, 2007; Nan Du *et al.*, 2010; Rossetti *et al.*, 2011). Une des questions porte sur comment synthétiser un réseau d'informations. Pour répondre à cette question, le problème de compression des graphes et plus récemment celui de compression des graphes issus du Web sont étudiés. Mais les travaux cherchent à réduire les graphes pour les stocker et les manipuler plus facilement alors que l'utilisateur a besoin d'une méthode plus intuitive pour contrôler et synthétiser les graphes (Tian *et al.*, 2008). Il peut vouloir contrôler la "résolution" des synthèses, exactement comme le fait l'OLAP (Online Analytical Processing - communément appelée l'analyse en ligne) dans les systèmes d'aide à la décision en faisant des forages vers le haut ou vers le bas. Synthétiser un graphe consiste à partir d'un

réseau d'entités à obtenir une vue plus agrégée de ces entités, vue qui doit prendre en compte les relations existantes entre les entités (Morfonios *et al.*, 2008). 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. Avec la construction de cubes OLAP, les utilisateurs peuvent visualiser et analyser les données à travers plusieurs dimensions (axes d'analyse) et à travers différents niveaux hiérarchiques pour chaque dimension. On parle d'analyse multidimensionnelle à plusieurs niveaux de granularité. Plusieurs auteurs proposent d'étendre l'OLAP à l'analyse des réseaux d'informations hétérogènes (Morfonios *et al.*, 2008; Tian *et al.*, 2008; Chen *et al.*, 2008). Ils montrent qu'il est nécessaire de faire évoluer l'OLAP pour l'adapter aux réseaux multidimensionnels et hétérogènes ; ce qui crée de nouveaux paradigmes de recherche.

Le 1er objectif de cet article est d'étudier les adaptations de l'OLAP traditionnel à l'OLAP sur les réseaux d'informations en considérant les données bibliographiques comme exemple de réseaux d'informations hétérogènes. Le 2ème objectif de l'article est de poser les bases d'un nouveau cadre d'analyse des données bibliographiques à partir de l'OLAP. Dans cet article de positionnement, il s'agit de décrire ce cadre et d'identifier les challenges qui seront à relever pour mettre en place un tel cadre. Les principes généraux de ce cadre sont les suivants. Pour une étude de publications scientifiques, nous voulons construire plusieurs réseaux, ces réseaux représentant différents points de vue du même problème. Notre objectif est de modéliser et de construire des réseaux multiples, puis de les stocker dans un entrepôt de données. Nous voulons ensuite développer un OLAP adapté pour visualiser et analyser les réseaux, tout en prenant en compte leur dynamique. La combinaison de l'OLAP avec des méthodes de fouille données et de fouille de textes sera un élément clé pour obtenir une pertinence des analyses.

Le reste de cet article est organisé de la façon suivante. La section 2 rappelle rapidement les concepts de base sur les réseaux d'information et l'OLAP. La section 3 présente le contexte de l'OLAP sur les réseaux d'informations, compare cette nouvelle analyse OLAP avec l'OLAP traditionnel et résume les principaux travaux dans le domaine. Le cadre pour l'analyse des données bibliographiques et les challenges qui en découlent sont exposés dans la section 4. Nous concluons dans la section 5.

2. Préliminaires

2.1. *Les réseaux d'informations*

Un réseau d'informations est formé d'un grand nombre d'objets (ou d'entités) interconnectés entre eux (Chen *et al.*, 2008; Qu *et al.*, 2011). Classiquement un réseau d'informations est modélisé à l'aide d'un graphe composé d'un ensemble de nœuds et d'un ensemble d'arêtes. Chaque noeud ou sommet représente un objet ou une entité et chaque arête ou lien exprime une relation entre deux entités. Un graphe est donc un ensemble de nœuds, dont certains sont directement reliés par un (ou plusieurs) lien(s).

Il y a deux types de réseaux. D'une part les réseaux peuvent être homogènes. Ils contiennent alors un seul type d'objets et un seul type de liens comme dans les réseaux d'amis, d'auteurs ou de films. Les liens peuvent avoir des attributs ou des poids. D'autre part, les réseaux d'informations sont hétérogènes quand ils sont composés de plusieurs types d'objets et de plusieurs types de liens. La figure 1 montre deux exemples de réseaux de données bibliographiques. Dans la figure 1a, le réseau d'auteurs est homogène ; chaque noeud représente un auteur (*authorID*) et chaque arête exprime une relation de co-auteur dans un ou plusieurs papiers. Par exemple, les auteurs A et B ont écrit trois papiers ensemble et dans la même conférence. Dans la figure 1b, le réseau est celui des auteurs et des papiers, c'est un réseau hétérogène. Ils est composé de deux types de sommets : les auteurs et les papiers. Il y a trois types d'arêtes : "*est écrit par*" entre un papier et des auteurs, le second type représente la relation de co-auteur et le troisième exprime que plusieurs papiers peuvent être écrits par un même auteur.

Figure 1. *Exemples de réseaux d'informations*

2.2. *L'analyse OLAP*

Dans les systèmes à base d'entrepôts de données, l'analyse OLAP (Online Analytical Processing) permet de construire des cubes multidimensionnels de données, de visualiser les données sous plusieurs dimensions et de naviguer à l'intérieur des cubes (Chaudhuri *et al.*, 1997). Le modèle multidimensionnel est composé de faits représentés par des mesures et des dimensions. La notion de fait désigne l'objet que l'on veut analyser. Sur le fait, on observe une ou des mesures (indicateurs). Dans l'OLAP traditionnel, les mesures sont le plus souvent numériques et elles sont assorties d'une fonction d'agrégation (somme, moyenne, minimum, maximum, comptage, ...) pour calculer la valeur des mesures aux différents niveaux hiérarchiques des dimensions. Une dimension définit un axe d'analyse et est constituée d'un ou de plusieurs niveaux hiérarchiques, appelés aussi attributs. Dans l'exemple des publications, les articles peuvent être les faits, la mesure peut être le nombre de publications et les dimensions peuvent être les auteurs, le temps, le support (*venue*) de la publication, les mots-clés, etc. La dimension du temps *time* peut avoir quatre niveaux hiérarchiques *semester*,

year, *decade*, *all* ; la dimension *venue* a trois niveaux hiérarchiques : *support* (le nom de la conférence, du journal ou du livre, etc.), *research area* (comme *databases*, *data mining* ou *information retrieval*, etc.) et *all*. Une caractéristique intéressante du modèle multidimensionnel est que la mesure peut être agrégée selon une ou plusieurs dimensions, par exemple on peut calculer le nombre total de publications chaque année dans le domaine des bases de données dans les différents pays. Parmi les opérateurs OLAP qui permettent à l'utilisateur de manipuler et visualiser les données dans un cube, l'opérateur *roll-up* est le plus connu : il permet d'agréger les données selon une ou plusieurs dimensions.

3. OLAP et les réseaux d'informations

3.1. *OLAP traditionnel et OLAP sur des graphes*

Nous pensons que l'OLAP doit évoluer pour analyser des données issues de réseaux d'informations et modélisées sous forme de graphes. Les expressions *OLAP sur des graphes* et *OLAP sur des réseaux d'informations* généralisent celle d'*OLAP social* c'est à dire d'OLAP sur des données issues de réseaux sociaux. Dans le tableau 1, nous proposons une comparaison entre l'OLAP traditionnel et ce qu'est ou ce que doit être l'OLAP sur des graphes (*Graph OLAP*). Cette comparaison permet de mettre en évidence les apports de l'OLAP sur les graphes.

Traditionnellement les entrepôts de données permettent de stocker, interroger et visualiser des données structurées et très souvent relationnelles. Depuis près de 10 ans, de nombreux travaux de recherche s'intéressent aussi aux entrepôts de données semi-structurées, textuelles et/ou XML. En revanche, dans l'OLAP sur les graphes, les travaux se concentrent sur les réseaux hétérogènes où les informations sont interconnectées et de plusieurs types. Nous pensons que les réseaux hétérogènes d'informations peuvent être vus comme une généralisation des bases de données, des données semi-structurées et même de certains corpus de documents. Malgré une modélisation très souvent relationnelle, les enregistrements stockés dans ces bases de données sont isolés, on ne les considère pas comme des objets interconnectés et liés entre eux à travers des relations de types multiples. En adoptant un autre point de vue, on peut voir les bases de données comme des réseaux d'informations très riches, réseaux que l'on va avoir besoin d'analyser pour extraire des connaissances. Par exemple, à partir d'une base de données des publications comme DBLP ou PubMed où les articles sont liés via les auteurs, les citations, les institutions, les thèmes etc., on peut construire le réseau des co-auteurs pour visualiser les collaborations entre auteurs, le réseau des citations, le réseau des conférences, etc.

Dans l'OLAP traditionnel, les cubes contiennent (en entrée) des faits définis par des dimensions (attributs) et mesures et permettent de générer (en sortie) des agrégats c'est à dire des faits dont la mesure a été agrégée selon des dimensions. Dans l'OLAP sur les graphes, les cubes peuvent contenir en entrée des graphes définis par une structure (entités et arêtes) et par des attributs. Nous pensons que l'agrégation d'un graphe

doit prendre en compte le contenu et la structure du graphe et fournit en sortie un graphe plus général.

Dans l'OLAP sur les graphes, en reprenant la terminologie de (Chen *et al.*, 2008), il peut y avoir plusieurs types de dimensions : les dimensions informationnelles (comme dans l'OLAP traditionnel) et les dimensions topologiques. Par exemple, en reprenant le réseau des auteurs et des articles de la figure 1b, *venue* et *time* sont deux attributs informationnels qui peuvent être utilisés comme dimensions informationnelles. Avec leur hiérarchie respective {*semester*, *year*, *decade*, *all*} et {*support*, *research area*, *all*}, ils permettent d'obtenir par exemple un réseau d'auteurs pour la conférence *ICDM* toutes années confondues (*all years*) et un autre pour le domaine du *data mining* en *2010*. *authorID* est un attribut qui définit un nœud, et le nombre de collaborations entre deux auteurs est un attribut de l'arête. La hiérarchie {*institute*, *country*, *continent*, *all*} associée à l'attribut *authorID* du nœud peut être utilisée comme dimension topologique et permet de regrouper par exemple tous les auteurs d'une même institution en un nœud plus général. Ainsi en résulte un nouveau graphe montrant les interactions entre les institutions. Nous pensons que les dimensions topologiques constituent une vraie valeur ajoutée dans la modélisation car elles permettent de modéliser les relations qui existent entre les objets.

Concernant les mesures, elles sont traditionnellement numériques, basées sur des indicateurs et elles sont assorties de fonctions d'agrégation comme la somme, la moyenne ou le comptage pour résumer, synthétiser ou agréger les faits. Dans les travaux sur les entrepôts de données XML ou textuelles, on peut toutefois trouver d'autres types de mesures et fonctions d'agrégation plus adaptées aux données semi-structurées. Outre les mesures numériques issues de la théorie des graphes (degré de voisinage, densité, centralité, diamètre etc.), dans l'OLAP sur les graphes on doit pouvoir avoir comme mesure un graphe. Un graphe peut être selon les cas un fait ou une mesure. Si la mesure est un graphe, il est nécessaire de développer de nouvelles fonctions d'agrégation adaptées aux graphes. Nous pensons que cette agrégation doit se faire en prenant en compte à la fois les attributs décrivant les entités mais aussi les relations qui existent entre les entités. L'OLAP traditionnel est finalement un cas particulier de cette nouvelle forme d'OLAP, cas dans lequel on ne tient pas compte des relations entre les faits pendant l'agrégation.

Dans l'OLAP sur les graphes, compte-tenu du fait que les dimensions peuvent être informationnelles ou topologiques, l'agrégation peut avoir deux sémantiques différentes. En reprenant la terminologie introduite par Chen *et al.* on parle d'OLAP informationnel (*I-OLAP*) et d'OLAP topologique (*T-OLAP*). Quand on fait un forage vers le haut selon une dimension informationnelle, la structure du réseau ne change pas, c'est une navigation de type *I-OLAP*. Par exemple, en faisant un *roll-up* sur les dimensions *venue* et *time*, on peut regrouper les collaborations entre auteurs dans des conférences d'un même domaine pendant une période donnée. En revanche, un *roll-up* selon une dimension topologique réorganise le réseau pour avoir une vue plus généralisée d'un ensemble d'objets ; c'est une navigation de type *T-OLAP*. La structure topologique du graphe originel est modifiée. Par exemple, on peut généraliser un ré-

seau d'auteurs par un réseau de collaborations entre institutions. En revanche dans l'OLAP traditionnel, il n'y a qu'une seule sémantique car le *roll-up* revient à faire un forage vers le haut dans l'OLAP informationnel.

Tableau 1. *Comparaison entre l'OLAP traditionnel et l'OLAP sur des graphes*

	OLAP traditionnel	OLAP sur des graphes
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	Mesures agrégées	Un réseau agrégé plus général
Dimensions	Uniquement informationnelles	Informationnelles et topologiques
Hiérarchies	Oui (uniquement informationnelles)	Oui (informationnelles et topologiques)
Mesures	Généralement numériques avec fonctions d'agrégation numériques (COUNT, SUM, MOY)	Graphe et mesures numériques de la théorie des graphes. Fonctions d'agrégation spécifiques
Opérations	Opérations classiques (de type I-OLAP)	Opérations I-OLAP et T-OLAP

3.2. *Les différentes approches*

Les premiers travaux de l'OLAP sur les réseaux sociaux puis sur les graphes hétérogènes d'informations remontent, à notre connaissance, à 2008 dans trois équipes différentes. Puis très récemment d'autres équipes se sont intéressées à ce sujet émergent.

Pour faire de la recherche dans le Web social, Morfonios et Koutrika proposent d'explorer l'espace des entités en partant d'un type d'entités et en retournant une vue agrégée des entités, vue basée sur les relations existant entre les entités (Morfonios *et al.*, 2008). Ils préconisent d'utiliser l'OLAP pour faire cette recherche et ils modélisent le réseau, la requête de l'utilisateur et la réponse à la requête sous forme de graphes. Ils utilisent un modèle multidimensionnel où chaque type d'entités correspond à une dimension et les faits expriment les relations entre les entités. La recherche de l'utilisateur, exprimée sous forme de requête, est considérée comme une agrégation de la table de faits. Les auteurs mènent une réflexion pour matérialiser et accélérer les agrégations à la volée et ils proposent plusieurs types de requêtes.

L'équipe de J. Han est à l'origine de plusieurs contributions (Chen *et al.*, 2008; Qu *et al.*, 2011; Zhao *et al.*, 2011). Chen *et al.* présentent les fondements et proposent un cadre général appelé *Graph OLAP* (Chen *et al.*, 2008). *Graph OLAP* est utilisé comme cadre formel par d'autres équipes de recherche. Qu *et al.* reprennent l'OLAP topologique pour proposer un 1er opérateur *T-OLAP*. Ils s'attaquent au problème de calculer efficacement la mesure lorsque l'utilisateur fait une requête de type T-OLAP qui modifie la structure du réseau (Qu *et al.*, 2011). Ils proposent deux propriétés *T-Distributiveness* et *T-Monocity* qui sont en fait deux propriétés classiques de la mesure redéfinies dans le cadre de *T-OLAP*. Zhao *et al.* introduisent le terme de réseaux multidimensionnels (*multidimensional networks*) et proposent un nouveau modèle de cube, appelé *Graph Cube* (Zhao *et al.*, 2011). *Graph Cube* supporte efficacement les

requêtes OLAP sur de larges réseaux multidimensionnels et inclut un nouveau type de requêtes appelées *crossboid queries* par opposition aux requêtes OLAP classiques appelées *cuboid queries*. L'agrégation d'un réseau se fait à la fois en fusionnant les entités et en agrégeant les liens. Les auteurs disent qu'ils obtiennent un réseau agrégé et enrichi par la structure.

Les travaux de Tian *et al.* sont les plus proches de ceux de Chen *et al*. Tian *et al.* proposent les opérateurs *SNAP* (Summarization by Grouping Nodes on Attributes and Pairwise Relationships) et *k-SNAP* pour fusionner des nœuds avec des attributs identiques, pour combiner les arêtes correspondantes et pour agréger le graphe en sous graphe tout en prenant en compte les liens entre les nœuds (Tian *et al.*, 2008). Au sens de la terminologie introduite dans *Graph OLAP*, nous qualifions les opérateurs *SNAP* et *k-SNAP* d'opérateurs OLAP topologiques (*T-OLAP*).

Plus récemment, d'autres équipes se sont intéressées à ce sujet. Kampgem *et al.* veulent récolter des données statistiques interconnectées et provenant de sources différentes pour les transformer, les intégrer et les stocker dans un entrepôt de données afin de les analyser avec la technologie OLAP (Kampgen *et al.*, 2011). Pour cela ils proposent une correspondance entre les données statistiques interconnectées et le modèle multidimensionnel tout en prenant en compte la sémantique véhiculée par ces données. Pour cela ils utilisent le vocabulaire de *RDF Data Cube*. On peut toutefois regretter que dans la modélisation proposée les auteurs ne tiennent pas compte des aspects topologiques des réseaux et qu'ils utilisent seulement les concepts traditionnels de l'OLAP.

Yin *et al.* reprochent au cadre *Graph OLAP* d'être plus adapté à la modélisation des réseaux d'informations homogènes qu'à la modélisation des réseaux hétérogènes (Yin *et al.*, 2012). Ils proposent de rajouter aux concepts des dimensions informationnelles et topologiques celui de dimensions-entités (*entity dimensions*) pour ainsi modéliser l'hétérogénéité des entités et des relations et baptisent leur nouveau cadre *HMGraph OLAP*. Ils assortissent ce nouveau concept de dimension de deux nouveaux opérateurs (*Rotate* et *Stretch*) pour inverser le rôle des entités et des relations et pour découvrir des relations implicites entre les entités. Les dimensions topologiques permettent de créer de nouveaux types de nœuds alors que les dimensions-entités changent la signification des arêtes et le nombre de nœuds dans le réseau. Dans *HMGraph OLAP*, Yin *et al.* proposent également un modèle multidimensionnel en constellation pour modéliser les réseaux hétérogènes dans un entrepôt de données et un modèle de cube de graphes appelé *HMGraph Cube*.

Beheshti *et al.* proposent *GOLAP*, un modèle dans lequel ils considèrent à la fois les objets ou entités ainsi que les liens qui existent entre ces objets (Beheshti *et al.*, 2012). Ils redéfinissent les concepts de l'OLAP pour les adapter aux graphes et ils utilisent les notions de dossiers et de nœuds-chemins pour partitionner les graphes. Ils étendent également le langager *SPARQL* pour pouvoir réaliser efficacement des opérations multidimensionnelles sur les cubes et proposent deux nouveaux opérateurs *UPDATE* et *UPSERT*.

Une des limites majeures de ces travaux présentés réside dans le fait de construire un entrepôt de données pour un seul réseau et pour un réseau homogène. Notre proposition est de construire un entrepôt s'intéressant à plusieurs réseaux, et ainsi capter des points de vue différents d'un ensemble de données. En effet selon le point de vue adopté, à partir d'un même ensemble de données bibliographiques, différents réseaux peuvent être extraits comme par exemple le réseau de citations et le réseau de co-auteurs. De plus, nous voulons également être capables de prendre en compte des réseaux hétérogènes.

4. Cadre

Notre proposition de cadre de travail est représentée dans la figure 2. Le point de départ est constitué des bases de données bibliographiques (trois bases de données recensant des publications dans le domaine de l'informatique : DBLP, ACM et le réseau PASCAL). Ces bases nous permettent de collecter les informations sur les publications sous la forme de données XML. Notre idée est ensuite de construire différents réseaux à partir de ces bases : réseau des co-auteurs, réseau de citations, réseau de thématiques, réseau des conférences et d'autres encore. Les réseaux sont hétérogènes et représentés sous la forme de graphes dans lesquels on trouve des entités et des liens de différents types : (i) le réseau des co-auteurs avec des auteurs comme nœuds et des arêtes représentant les liens de co-rédaction d'un papier ; (ii) le réseau des citations où les nœuds représentent les papiers où les liens entre un papier et les papiers qu'il cite sont représentés par des arcs orientés ; (iii) le réseau des thématiques peut relier des papiers ou des auteurs abordant les mêmes thématiques ; (iv) le réseau des conférences liant deux conférences quand les même auteurs ont publié dans ces deux conférences.

La 1ère étape consiste à construire les réseaux et à les enrichir par des connaissances extraites automatiquement. Différentes techniques peuvent être utilisées pour extraire des connaissances à partir de ces réseaux, telles que la fouille de données et l'analyse de réseaux sociaux. Par exemple, les méthodes de classification non supervisée sont particulièrement adaptées pour détecter des communautés, en regroupant des thématiques ou des auteurs. Ces groupes détectés comprennent alors des entités ayant des propriétés communes. Ces connaissances peuvent permettre de construire des hiérarchies de dimension pour enrichir l'analyse en ligne. Il est également intéressant d'envisager la dimension temporelle pour suivre l'évolution de ces groupes. Un autre exemple est celui des méthodes d'ordonnancement qui permettent d'évaluer ou de trier des objets d'un réseau avec une fonction mathématique ou statistique et ainsi trouver le top-100 des meilleurs chercheurs ou les thématiques les plus populaires. Les méthodes d'analyse de réseaux sociaux, quant à elles, peuvent être utilisées pour étudier les liens, les citations et calculer des indicateurs comme un score pertinent pour découvrir des collaborateurs sur des sujets similaires. Toutes ces techniques vont permettre d'enrichir les différents réseaux par exemple en fournissant des nouvelles dimensions ou des niveaux de dimensions.

Les réseaux et les connaissances extraites doivent être chargés dans l'entrepôt de données. Cette 2ème étape se fait à travers le processus d'extraction, transformation et chargement (ETL pour *Extract, Transform and Load*). La structure de l'entrepôt est basée sur un modèle multidimensionnel et elle doit être capable de stocker les données, les réseaux et les connaissances associées. Ce modèle reste à définir et constitue un réel challenge de modélisation, du fait que les modèles qualifiés de classiques ne répondent pas au besoin de représentation de tous ces éléments.

Nous pensons que le fait peut être un nœud simple (un auteur, une institution) ou un réseau (réseau de co-publications). La mesure peut être classique et correspondre à un indicateur numérique (le nombre de papiers, le nombre de citations, un nombre de téléchargements, etc) ou provenir de la théorie des graphes (la centralité, le diamètre, la similarité, etc.). Mais la mesure peut aussi être textuelle comme par exemple des mots-clés. Et enfin la mesure peut également être un réseau. Dans notre modèle nous souhaitons pouvoir prendre en compte tous les types de mesure et les fonctions d'agrégation adaptées à chaque type. Par exemple, les fonctions d'agrégation numériques peuvent être utilisées sur les mesures classiques pour calculer un indicateur par groupe d'auteurs ou par institution. L'agrégation d'un graphe correspondrait au fait de regrouper des auteurs en fonction de leurs relations.

Une fois que nous disposons d'un entrepôt consolidant toutes les données, notre idée est de créer un ensemble d'outils OLAP pour la visualisation et la navigation. Ils pourront être utilisés pour analyser par exemple la dynamique des réseaux (auteurs, publications, ...) à travers le temps et ainsi voir les thématiques les plus populaires chaque année ou les auteurs les plus actifs. Un autre outil pourra permettre de naviguer au sein des réseaux pour répondre à des requêtes utilisateurs telles que la navigation parmi les auteurs qui collaborent et qui travaillent sur des thématiques similaires. Pour créer ces outils ou possibilités d'analyse, nous pensons qu'il faut combiner les approches de fouille de données, de fouille de texte aux opérateurs OLAP.

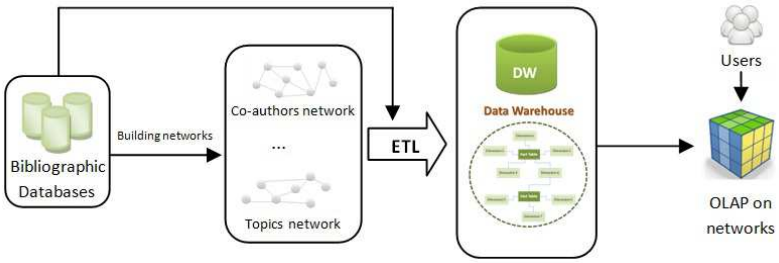

Figure 2. *Proposition d'un cadre de travail*

Ainsi, en considérant ce cadre de travail, nous avons identifié plusieurs challenges.

Tout d'abord, nous devons construire différents réseaux à partir des bases de données bibliographiques et en extraire des connaissances. Pour cette tâche, les algorithmes existants seront considérés.

Un 1er challenge très important consiste en la définition du modèle multidimensionnel capable de considérer plusieurs réseaux, des réseaux hétérogènes et les connaissances extraites. Il s'avère que les modèles classiques ne pourront pas répondre à nos besoins et nous serons donc amenés à inventer un nouveau type de modèle. Ce nouveau modèle devra être capable de prendre en compte les différents types de noeuds, de liens et de mesures. Des fonctions d'agrégation spécifiques aux différents types de mesure doivent être développées et elles devront prendre en compte à la fois la structure et le contenu des réseaux. De plus, malgré l'intérêt qui leur est porté, il n'existe pas de modèle multidimensionnel standard pour les données bibliographiques. Cela pourrait être l'occasion d'aller vers une proposition unifiée des différents modèles.

Ensuite, nous aurons à considérer la phase d'ETL. Les données proviennent de sources différentes et hétérogènes et doivent être intégrées. Plusieurs questions se posent alors. Comment prendre en compte l'usage de techniques de fouille de données pour enrichir les réseaux ? Comment alimenter l'entrepôt à la fois avec les réseaux et les connaissances extraites ?

Enfin, un autre défi crucial réside au niveau de l'analyse. Il s'agit de fournir un ensemble d'outils pour analyser en ligne les données. Des méthodes d'analyse innovantes doivent être proposées aux utilisateurs en s'inspirant à la fois de la philosophie OLAP, mais aussi de la visualisation, dont celle des graphes. Nous pensons qu'il sera possible de développer ces nouveaux outils en combinant les méthodes de fouille de données, de fouille de textes aux opérateurs OLAP et aux principes de la visualisation. L'introduction de techniques de fouille de texte et d'analyse en ligne dans le processus des analyses bibliométriques constitue une direction de recherche prometteuse.

5. Conclusion

Dans ce papier, nous avons discuté de l'intérêt de combiner l'analyse en ligne OLAP et les réseaux d'information en présentant un état de l'art sur l'usage de ces deux domaines afin de souligner comment il est possible de les combiner. Nous avons proposé un cadre d'analyse des données bibliographiques bénéficiant de ces deux domaines en mentionnant quels sont les challenges à traiter. Les principales idées sont donc : (i) construire les différents réseaux à partir de sources de données bibliographiques telles que DBLP, ACM... (réseau des co-auteurs, réseau des citations, réseau des thématiques, réseau des conférences) ; (ii) construire un entrepôt de données avec le modèle approprié pour explorer les informations ; (iii) appliquer des techniques de fouille de données pour enrichir les informations (comme par exemple détecter les communautés pour enrichir les hiérarchies de dimension de l'entrepôt de données) ; et (iv) développer un outil approprié (inspiré du processus de navigation OLAP) pour visualiser ces données.

A terme, ce travail devrait contribuer à enrichir les méthodes disponibles pour la bibliométrie, proposant un outil de navigation utile pour observer la production et

tester des hypothèses. Un tel outil de navigation, caractérisé par une intuitivité d'usage, pourra être utile aux chercheurs s'intéressant à la bibliométrie, qui peuvent être issus de domaines très variés, allant de l'histoire de la science aux sciences de l'information, en passant par le domaine de la sociologie.

6. Bibliographie

Beheshti S., Benatallah B., Motahari-Nezhad H., « A Framework and a Language for On-Line Analytical Processing on Graphs », *13th International Conference on Web Information Systems Engineering (WISE'12)*, p. 213-227, 2012.

Chaudhuri S., Dayal U., « An Overview of Data Warehousing and OLAP Technology », *ACM SIGMOD*, vol. 26, n° 1, p. 65-74, 1997.

Chen C., Yan X., Zhu F., Han J., Yu P., « Graph olap : Towards online analytical processing on graphs », *IEEE International Conference on Data Mining (TCDM'08)*, p. 103-112, 2008.

Getoor L., B. T., « Introduction to Statistical Relational Learning », *Adaptive computation and machine learning*, 2007.

Han J., « Mining Heterogeneous Information Networks by Exploring the Power of Links », *12th International Conference on Discovery Science (DS'09)*, p. 13-30, 2009.

Jensen D., Neville J., Gallagher B., « Why collective inference improves relational classification ? », *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, p. 593-598, 2004.

Kampgen B., Harth A., « Transforming statistical linked data for use in OLAP systems », *7th International Conference on Semantic Systems (I-SEMANTICS'11)*, p. 33-40, 2011.

Morfonios K., Koutrika G., « OLAP Cubes for Social Searches : Standing on the Shoulders of Giants ? », *International Workshop on the Web and Databases (WebDB)*, 2008.

Nan Du W. H., Faloutsos C., « Analysis of Large Multi-modal Social Networks : Patterns and a Generator », *Lecture Notes in Computer Science*, 2010.

Qu Q., Zhu F., Yan X., Han J., Yu P., Li H., « Efficient Topological OLAP on Information Networks », *16th International Conference on Database systems for advanced applications (DASFAA'11)*, vol. 1, p. 389-403, 2011.

Rossetti G., Berlingerio M., Giannotti F., « Scalable Link Prediction on Multidimensional Networks », *11th IEEE International Conference on Data Mining Workshops (ICDMW)*, p. 979-986, 2011.

Tian Y., Hankins R., Patel L., « Efficient Aggregation for Graph Summarization », *ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*, p. 567-580, 2008.

Yin M., Wu B., Zeng Z., « HMGraph OLAP : a Novel Framework for Multi-dimensional Heterogeneous Network Analysis », *15th International Workshop on Data warehousing and OLAP (DOLAP'12)*, p. 137-144, 2012.

Zhao P., Li X., Xin D., Han J., « Graph Cube : On Warehousing and OLAP Multidimensional Networks », *ACM SIGMOD International Conference on Management of Data (SIGMOD'11)*, p. 853-864, 2011.