Graphes enrichis par des Cubes (GreC) : une approche innovante pour l'OLAP sur des réseaux d'information

Cécile Favre, Wararat Jakawat, Sabine Loudcher

Graphes enrichis par des Cubes (GreC) : une approche innovante pour l'OLAP sur des réseaux d'information

Cécile Favre[1], **Wararat Jakawat**[2], **Sabine Loudcher**[3]

1. *Université de Lyon, Université Lyon 2, ERIC EA 3083, France*
   cecile.favre@univ-lyon2.fr
2. *Computer Science Department, Prince of Songkhla University, Thailand*
   wararat.j@psu.ac.th
3. *Université de Lyon, Université Lyon 2, ERIC EA 3083, France*
   sabine.loudcher@univ-lyon2.fr

ABSTRACT. *In order to make the online analysis of information networks, several works combine OLAP and graphs, combination known as* Graph OLAP. *To complete these works which often offer to analyze cubes of graphs, in a different and complementary way, we propose a new approach called* GreC *(*Graphs enriched by Cubes*). Instead of building cubes of graphs, our proposal is to enrich the graphs with cubes, graphs where the nodes or edges of the network are described by cubes. This allows interesting analyses for the user who can navigate within a graph enriched by cubes according to different levels of analysis, with dedicated operators. In this article we recall the general framework of* GreC *and show how the classic concepts of OLAP must be revisited and expanded. Then we will focus on metadata in the model that ensures the genericity of the approach, on the implementation of a prototype and on elements of performances.*

RÉSUMÉ. *Afin de pouvoir faire de l'analyse en ligne de réseaux d'information, plusieurs travaux proposent de combiner l'OLAP et les graphes, combinaison connue sous le nom de* Graph OLAP. *Pour compléter ces travaux dont le principe fondamental est d'analyser des cubes de graphes, nous proposons une approche innovante appelée* GreC *(*Graphes enrichis par des Cubes*). Plutôt que de construire des cubes de graphes, notre proposition consiste à enrichir les graphes avec des cubes de données qui viennent décrire les nœuds et/ou les arêtes du réseau selon les besoins. Cela permet des analyses intéressantes pour l'utilisateur qui peut naviguer au sein d'un graphe enrichi de cubes selon différents niveaux d'analyse, avec des opérateurs dédiés. Dans cet article nous rappelons le cadre général de l'approche* GreC *et montrons comment les concepts classiques de l'OLAP doivent être revisités et étendus. Puis nous nous focalisons sur les métadonnées qui permettent d'assurer la généricité de l'approche, sur l'implémentation d'un prototype et donnons quelques éléments relatifs aux performances.*

KEYWORDS: *Réseau d'informations ; cube ; OLAP ; graph OLAP ; opération informationnelle ; opération topologique.*

MOTS-CLÉS : *Information network ; cube ; OLAP ; graph OLAP ; informational operation ; topological operation.*

1. Introduction

Historiquement, l'analyse OLAP (*Online Analytical Processing*) a été développée et utilisée dans un contexte de données assez classiques, souvent structurées dans des bases de données. L'émergence de nouveaux types de données à considérer, comme par exemple le texte, les images, les réseaux d'information, a soulevé de nouveaux défis à relever pour permettre une extension de cette technologie, entre autres en revisitant les concepts, en recherchant comment transposer ce qui existait à de nouveaux types de données, en développant de nouvelles approches prenant en compte ces nouveaux types de données, et ce pour tirer parti de la richesse de leurs spécificités. Dans le paysage des données complexes, les réseaux d'information constituent un type de données particulièrement riche compte-tenu non seulement de la multiplicité des données, mais aussi de leurs liens. La modélisation sous forme de graphes avec des nœuds et des arêtes peut prendre différentes formes selon les besoins de représentation : graphe valué ou non pour la pondération des arcs, graphe homogène (un seul type de nœud) ou hétérogène, etc.

Pour illustrer les réseaux d'information, considérons les données bibliographiques qui se prêtent particulièrement bien à la représentation sous forme de graphes. Ces données ont d'ailleurs fait l'objet des premières approches qui ont tenté de combiner les graphes et l'approche OLAP. Il apparaît qu'une des caractéristiques importantes des données bibliographiques réside dans le fait que, de par leur nature, elles sont liées entre elles et peuvent donner lieu à une représentation sous forme de graphe. Par exemple, le fait que deux auteurs aient publié ensemble induit le fait que sur un graphe d'auteurs, si nous nous intéressons à la co-publication, l'arête reliant ces deux auteurs pourra être valuée par le nombre de papiers que les personnes ont écrits ensemble. Par exemple dans la Figure 1, J. Han et Y.Sun ont collaboré au travers de 5 publications.

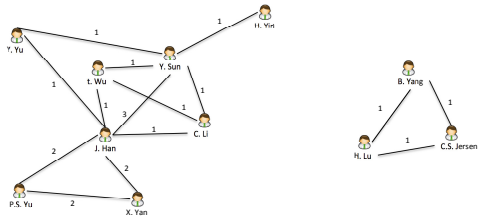

Figure 1. Graphe d'auteurs représentant les co-publications à un instant t.

Néanmoins, dans cet exemple, on peut constater que le pouvoir informatif de ce graphe reste assez faible. En effet, cette représentation ne prend pas en compte la dynamique des données (c'est une photo à un instant t de l'état des co-publications) ; par ailleurs, cette représentation ne permet pas de rendre compte de différentes informations caractérisant les publications dénombrées, telles que l'année, le lieu de publication, la thématique, etc.

Une autre alternative de visualisation pour rendre compte de ces informations correspond à ce qui est proposé par l'analyse OLAP avec une représentation multidimensionnelle sous forme de cube. Par exemple, dans la Figure 2, il est possible d'analyser le *fait* (objet d'analyse) "production scientifique", au travers d'une *mesure* (indicateur) qui est le "nombre de publications", en fonction de différentes *dimensions* (axes d'analyse) qui sont ici au nombre de trois : "auteur", "temps" et "lieu" (venue). Ces dimensions peuvent être organisées sous forme de *hiérarchies*, organisées en différents niveaux de granularité. Par exemple, la dimension "lieu" a une hiérarchie en deux niveaux : un niveau avec le nom du lieu de publication et un niveau "domaine". La présence d'une dimension hiérarchisée est un élément important de la navigation dans les données. En effet, l'OLAP traditionnel propose différentes opérations de navigation dans les données. Parmi les plus utilisées, il y a les opérations qui permettent de naviguer à travers les niveaux de détail des données selon les dimensions hiérarchisées, avec un processus d'agrégation : le *Roll Up* (forage ascendant) qui permet d'obtenir les données à un niveau agrégé et le *Drill Down* (forage descendant) qui fait l'inverse. Il est à noter que dans cette représentation multidimensionnelle qui comporte davantage d'informations, le fait que les auteurs soient en lien au travers de ces publications (co-publications) n'apparaît pas du tout.

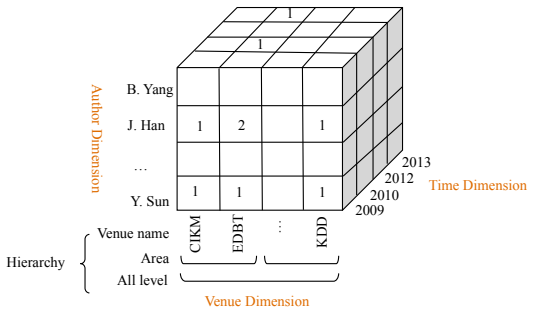

Figure 2. Structure d'un cube de données dans le contexte OLAP pour des données bibliographiques.

Ainsi, afin de tirer parti de ces deux visualisations (graphe et cube), un nouveau champ de recherche est apparu : *Graph OLAP* (Chen *et al.*, 2008). Le *Graph OLAP* a fait l'objet de plusieurs publications proposant des améliorations et des extensions (Jin *et al.*, 2010; Qu *et al.*, 2011; Zhao *et al.*, 2011). L'idée sur laquelle repose initialement le *Graph OLAP* consiste à construire un cube de graphes dans lequel il est

possible de naviguer, grâce à différents opérateurs OLAP qui ont été redéfinis pour prendre en compte ce nouveau cadre d'analyse. Dans ces approches de *Graph OLAP*, il s'agit de considérer des cubes définis selon des dimensions dites informationnelles, et les mesures contenues dans les cellules correspondent, non pas à des indicateurs numériques comme traditionnellement, mais à des graphes ou plus exactement à des sous-graphes. Par exemple, dans la Figure 3, par rapport aux données considérées ici, les dimensions informationnelles sont le temps, le lieu et les mots-clés. Ici le cube est présenté sous forme de "tranche", en considérant tous les mots-clés simultanément. Dans la cellule qui est définie par les valeurs "EDBT" pour le lieu et "2013" pour le temps, le réseau est composé des co-auteurs B. Yang et C.S. Jensen qui ont co-publié un papier (dans cette conférence, cette année-là), en considérant le nombre de papiers co-publiés qui value les arêtes du graphe.

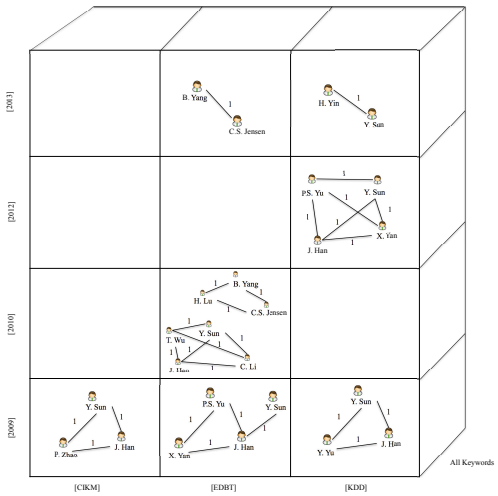

Figure 3. Cube de graphes sur des données bibliographiques pour analyser les liens de co-publication.

Dans les approches initiales de *Graph OLAP*, au niveau de la modélisation, deux types de dimensions ont été définis : les dimensions informationnelles et les dimensions topologiques. Les dimensions informationnelles vont donc conditionner les manipulations du cube. Ainsi, lors d'opérations informationnelles sur le cube, les graphes à l'intérieur des cellules vont être recalculés. Les dimensions topologiques, quant à elles, se rapportent à la modélisation des réseaux eux-mêmes dans les cellules. Les opérations topologiques sont caractérisées par un changement du type de nœuds dans les graphes. Par exemple, à partir d'un réseau d'auteurs présent dans une cellule, nous passons à un réseau d'institutions, si nous effectuons un *Roll Up* topologique selon la dimension auteur, dont la hiérarchie comprend un niveau institution.

Initialement, la combinaison de l'OLAP et des graphes s'est donc faite au travers de plusieurs approches basées sur des cubes de graphes avec dans leurs cellules, un

graphe comme mesure (Tian *et al.*, 2008; Morfonios, Koutrika, 2008; Beheshti *et al.*, 2012; Yin *et al.*, 2012). Ces approches permettent de visualiser des "instantanés" de graphes en fonction des dimensions d'analyse choisies (c'est à dire l'état d'un graphe à un instant donné). Différents opérateurs ont été proposés pour naviguer dans le cube de graphes : des opérations informationnelles ou topologiques, selon si les opérations s'appliquent selon les dimensions du cube ou les dimensions des graphes. Dans (Loudcher *et al.*, 2015), nous proposons un état de l'art et une étude comparative de ces différentes approches. Cependant, dans cette combinaison de l'OLAP et des graphes basée sur des cubes de graphes, la visualisation plus globale du graphe est perdue, alors même que celle-ci est intéressante d'un point de vue analytique. Parallèlement, la dynamique des données est importante pour l'analyse du graphe, et ceci n'est pas toujours bien perceptible dans la visualisation des parties de graphe. En effet, malgré la présence d'une dimension temporelle, en croisant celle-ci avec une ou plusieurs autres dimensions, il est difficile de se rendre compte de la dynamique même d'un graphe (évolution des arêtes ou des nœuds).

Par conséquent, nous proposons une nouvelle façon de considérer la combinaison de l'OLAP et des graphes en construisant un graphe qui réponde aux besoins de l'utilisateur avec l'enrichissement par des cubes de données pour valuer les nœuds et/ou les arêtes selon les besoins d'analyse. De plus, la présence d'une dimension temporelle dans les cubes qui valuent les nœuds et/ou les arêtes va notamment permettre de rendre compte de la dynamique du graphe. Par ailleurs, pour enrichir l'analyse, notre attention s'est focalisée sur deux apports : d'une part les types de mesures possibles ; d'autre part les opérateurs de navigation proposés. Notre approche s'intitule *GreC* pour Graphes enrichis par des Cubes.

Le cadre général de l'approche *GreC* a déjà donné lieu à des publications (Jakawat *et al.*, 2016a; 2016b). Dans le présent papier nous nous attachons à l'opérationnalité de *GreC* sur l'ensemble du processus mais aussi à sa généricité au-delà du contexte des données bibliographiques qui est le contexte initial de développement de notre approche. En effet, cette approche pourrait par exemple être appliquée dans le cadre de l'analyse de messages Twitter en se focalisant sur le graphe des *followers* enrichi par des cubes informant de l'activité propre en nombre de tweets selon le temps, la thématique, etc. ; des cubes sur les arêtes se focalisant sur les mentions entre deux *Twittos* renseignant sur la production de tweets selon les mêmes axes se restreignant aux mentions des comptes dans les tweets produits par exemple. Dans ce papier, nous gardons comme fil conducteur les données bibliographiques, tout en discutant les points permettant la généricité de l'approche. Dans cette optique, nous présentons, dans ce papier, l'utilisation dans *GreC* de métadonnées pour lesquelles nous proposons une abstraction au travers d'un métamodèle de GreC.

La suite du papier est organisée de la manière suivante. Nous commencerons par rappeler le cadre général de l'approche *GreC* et monterons comment les concepts classiques de l'OLAP doivent être revisités et étendus (Section 2). Puis nous nous focaliserons sur l'opérationnalisation de l'approche avec l'introduction de métadonnées, de leur modélisation et avec des considérations calculatoires (Section 3). En-

fin nous présenterons l'implémentation d'un prototype pour montrer la faisabilité de l'approche et discuterons des performances (Section 4), pour finalement conclure dans une dernière section (Section 5).

2. Cadre général de l'approche *GreC*

Nous proposons l'approche *GreC* (Graphes enrichis par des Cubes) qui est une nouvelle façon de considérer la combinaison de l'OLAP et des graphes pour l'analyse de réseaux d'information. *GreC* est une approche originale et complémentaire des approches basées sur une construction d'un cube de graphes (Jakawat *et al.*, 2016b). Elle permet de construire un graphe qui réponde aux besoins d'analyse de l'utilisateur et de l'enrichir avec des cubes de données qui vont décrire et valuer les nœuds et/ou les arêtes selon les besoins d'analyse. L'utilisateur peut ainsi avoir une vue globale d'une partie du réseau avec des informations multidimensionnelles et faire des analyses intéressantes en naviguant au sein du graphe enrichi avec des opérateurs dédiés. *GreC* considère la structure du réseau pour permettre des opérations OLAP topologiques, et pas seulement des opérations OLAP classiques et informationnelles. La présence d'une dimension temporelle dans les cubes qui valuent les nœuds et/ou les arêtes permet de rendre compte d'une certaine façon de la dynamique du graphe.

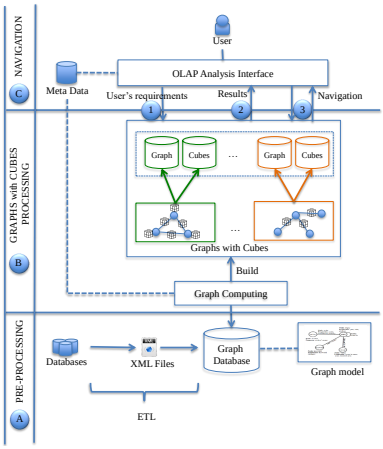

Figure 4. Processus de GreC.

La Figure 4 illustre le fonctionnement global de l'approche *GreC* appliquée, à titre d'exemple, sur les données bibliographiques. Le point de départ est la phase préparatoire (couche A), qui correspond au pré-traitement des données. Diverses bases de données bibliographiques sont fusionnées et intégrées dans des fichiers XML qui alimentent une base de données orientée graphe selon un modèle de données que nous avons défini (Jakawat *et al.*, 2016a). Il s'agit ici de considérer un graphe hétérogène comportant l'ensemble de toutes les données. Ensuite, dans la couche B, à partir du

graphe hétérogène des données de base, les graphes enrichis par des cubes sont construits (notons que l'ensemble des graphes est calculé en amont pour assurer de bonnes performances pour l'utilisateur, nous y reviendrons ultérieurement). Cette construction se décompose en deux étapes : construction du graphe lui-même, puis celle des cubes de données qui valuent les nœuds et/ou les arêtes. Ces deux étapes se répètent pour construire l'ensemble des graphes enrichis par les cubes. Dans la couche C, grâce à une interface de navigation, l'utilisateur exprime ses besoins d'analyse, ce qui permet de sélectionner le graphe adéquat. Une fois le graphe adéquat obtenu, l'utilisateur peut naviguer grâce à des opérateurs adaptés, à la fois par rapport au graphe, mais aussi par rapport aux cubes qui lui sont associés.

Pour permettre l'analyse en ligne de graphes enrichis par des cubes ainsi décrite, nous sommes amenées à redéfinir et à étendre les concepts de l'OLAP manipulés dans le contexte de *GreC*. Tout comme dans l'approche classique d'analyse en ligne, dans *GreC*, il s'agit d'analyser un fait. Par exemple, dans le cadre des données bibliographiques, il peut s'agir d'analyser la production scientifique ou la co-publication. En revanche, le fait n'est pas directement analysé au travers d'une mesure, mais au travers d'un graphe. En fonction du fait, et des besoins d'analyse, des métadonnées permettent de déterminer si des cubes de données valuent des nœuds et/ou des arêtes.

La notion de cube dans *GreC* correspond à un cube classique, qui contient dans chacune de ses cellules la valeur d'une ou plusieurs mesures numériques ; ces mesures peuvent être «simples» (additives) comme le nombre de publications ou elles peuvent être basées sur des graphes comme par exemple une mesure de degré de centralité.

Comme dans l'approche *Graph OLAP* initiale, nous retrouvons deux types de dimension : dimension informationnelle et dimension topologique. Les dimensions informationnelles correspondent aux dimensions définissant les cubes de données attenants aux nœuds ou aux arêtes. Les dimensions topologiques correspondent aux dimensions par rapports aux éléments représentés au niveau du graphe, avec dans les deux cas, la possibilité d'une hiérarchisation. Par exemple, la dimension topologique *auteur* est hiérarchisée avec un niveau *institution*. Ceci permettra de passer du graphe des auteurs au graphe des institutions par exemple. De plus, nous parlons, non pas d'opérateurs OLAP, mais d'opérateurs OLAP informationnels ou topologiques, déterminant ainsi si l'opération (que ce soit un *Roll Up*, *Drill Down*, etc.) est appliquée par rapport au graphe en question selon une dimension topologique, ou aux cubes de ce graphe selon une dimension informationnelle.

Pour rendre opérationnelle l'approche *Grec* sur l'ensemble du processus, nous avons besoin de définir des métadonnées (en plus du modèle de données spécifique à chaque réseau d'information) et de développer de nouveaux algorithmes pour construire les graphes et les cubes, calculer les mesures et adapter les concepts OLAP.

3. Opérationnalisation de GreC

Rendre opérationnelle l'approche GreC pour les données bibliographiques et aussi pour d'autres données nécessite de penser une mise en œuvre assurant la généricité de l'approche. Ceci passe en particulier par la définition de métadonnées.

3.1. Métadonnées pour la généricité

Pour permettre la mise en œuvre de l'approche *GreC*, il est nécessaire que les données de base à considérer soient modélisées dans un graphe hétérogène retraçant l'ensemble des données et leurs liens à prendre en compte. Ce modèle est spécifique à chaque réseau étudié et ne peut faire l'objet d'un modèle générique. Le modèle que nous proposons pour l'exemple de l'analyse des données bibliographiques est introduit dans la section 4.2 lors de la présentation des données pour les expérimentations.

En revanche, d'une façon générale, à partir du graphe hétérogène complet, pour extraire et construire les graphes enrichis par des cubes ainsi que pour assurer la généricité de l'approche *GreC* et notamment celle de l'interface de navigation, nous introduisons des métadonnées avec un modèle dédié (Figure 5), correspondant au métamodèle de *GreC*.

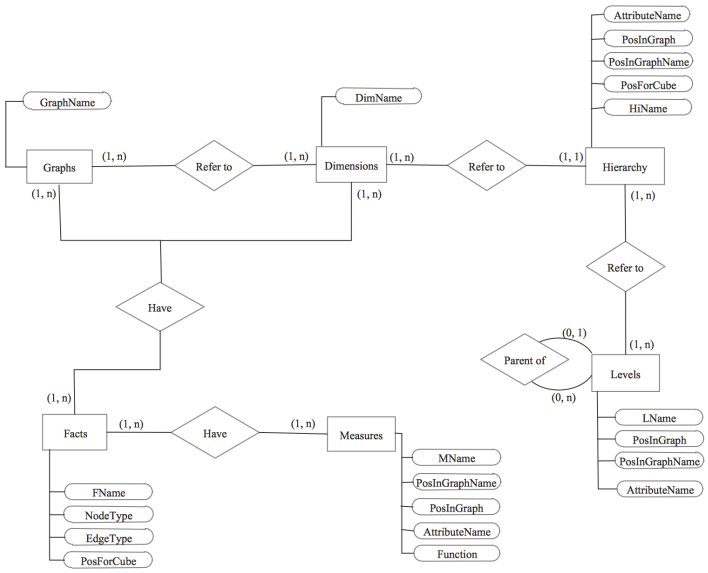

Figure 5. Métamodèle simplifié de GreC.

En adoptant le formalisme du modèle conceptuel entité-association (EA), les principales entités sont les "faits", "mesures", "dimensions", "hiérarchies", "niveaux" et "graphes". Ceci a pour but de représenter les contextes d'analyse possibles pour

l'utilisateur, en partant des faits à analyser et de comment le faire : avec quel graphe, quels cubes, grâce à l'instanciation de ces métadonnées.

Les principales associations permettent alors de déterminer quels graphes sont possibles par rapport à un fait à analyser (association entre *FACTS* et *GRAPHS*) ; quels indicateurs sont possibles (association entre *FACTS* et *MEASURES*). L'entité *DIMENSIONS* est associée à la fois à l'entité *GRAPHS* et *FACTS*, ce qui permet de préciser les dimensions topologiques et informationnelles respectivement, avec ensuite une association avec l'entité *HIERARCHY*, associée elle-même à l'entité *LEVELS*, ce qui permet de déterminer les niveaux de navigation. La définition des mesures et des dimensions informationnelles déterminent la structure du cube.

Chaque entité est bien sûr décrite par différents attributs. Ici, nous mettons en avant les attributs ayant un rôle particulier. Pour l'entité *FACTS*, notons que le fait est précisé au travers du type de nœuds constituant le graphe qui permet d'analyser ce fait (attribut *NodeType*). Chaque fait sera analysé au travers d'un graphe homogène. Nous y trouvons également l'attribut *PosForCube* qui permet de caractériser la position des cubes qui vont enrichir le graphe : au niveau des arêtes, des nœuds, ou à la fois des arêtes et des nœuds. L'attribut *EdgeType* permet de préciser le chemin type dans le graphe initial hétérogène qui permettra la construction du graphe homogène en fonction du fait défini. Au niveau de l'entité *MEASURES*, il est précisé la façon d'agréger les données avec l'attribut *Function*, qui prend par exemple la valeur "numeric" si la mesure peut s'agréger simplement par additivité, ou "degree" s'il s'agit d'une mesure de type calcul de degré basée sur les graphes et nécessitant donc un recalcul de la mesure en fonction des choix d'analyse ou des opérations appliquées, et ce à partir des données sources du graphe initial hétérogène.

Cette modélisation peut s'illustrer sur l'exemple des données bibliographiques en instanciant ces métadonnées :

– Si l'utilisateur veut analyser la co-publication, l'entité *FACTS* permet de tracer le fait que le réseau des co-publications est un graphe où les noeuds sont les auteurs et les arêtes entre deux d'entre eux indiquent qu'ils ont co-écrit ensemble. Cela permet de préciser aussi que ce graphe a des cubes seulement sur les arêtes, puisque l'analyse est centrée ici sur la co-publication et non sur la publication, avec dans ce dernier cas, des cubes qui seraient à la fois sur les arêtes et sur le nœuds, spécifiant que des publications peuvent être écrites par un unique auteur sans collaboration.

– Si le fait est la co-publication, les mesures peuvent être le nombre de papiers, cela peut aussi être une mesure basée sur le graphe comme le degré de centralité. Dans le premier cas, nous avons donc des cubes au niveau des arêtes, mais pour le degré de centralité qui permet d'estimer l'activité d'un nœud, le cube contenant ce type de mesure se trouverait au niveau des nœuds. Ainsi, dans ce dernier cas, il s'agit pour chaque auteur du graphe d'avoir un cube qui caractérise le nombre total de liens (avec des co-auteurs différents) en fonction des dimensions du cube choisies, permettant d'analyser les auteurs actifs dans des collaborations variées.

– Si le fait est la co-publication, plusieurs dimensions comme *time* et *venue* peuvent être utilisées au niveau des cubes qui seront définis. Ainsi, si la mesure est le nombre de papiers pour ce fait, cela induit qu'au niveau des arêtes, il y a des cubes qui déterminent le nombre de papiers co-publiés par année et par conférence.

– Une dimension peut être structurée selon une hiérarchie. Par exemple la dimension institution a une hiérarchie du style : *author name* / *institution name* / *country*, *country* étant un niveau plus élevé de *institution name*. Dans le cadre de l'analyse de co-publication, cette hiérarchie de dimension topologique permettra de faire des opérations pour analyser le phénomène de co-publication à l'échelle des auteurs, mais également des institutions, ou même des pays pour analyser l'internationalisation des collaborations scientifiques.

Le modèle des métadonnées est ainsi conçu pour assurer la généricité de l'approche. Il permet de faire le lien entre les faits à analyser, les graphes à construire et l'emplacement des cubes (au niveau des nœuds et/ou des arêtes). Il permet également de décrire les concepts OLAP (faits, mesures, dimensions, etc.) et de stocker leur instanciation. Les métadonnées sont utilisées par les différents algorithmes qui construisent les graphes et les cubes, qui calculent les mesures et qui réalisent les opérations OLAP redéfinies. Elles conditionnent également l'interfaçage de l'application.

3.2. Considérations calculatoires

Rappelons que le point de départ est un graphe hétérogène avec l'ensemble des données. En fonction des besoins d'analyse exprimés par l'utilisateur, un graphe est proposé à ce dernier, il pourra naviguer dans les données de celui-ci grâce à différents opérateurs OLAP adaptés à l'approche GreC : navigation dans le graphe ou dans les cubes qui enrichissent le graphe.

Différents algorithmes ont été implémentés pour mettre en œuvre *GreC*, en se basant sur l'usage des métadonnées, le graphe initial hétérogène et les besoins d'analyse, cela recouvre notamment les étapes suivantes :

1. la construction du graphe pour l'utilisateur ;
2. la construction des cubes pour valuer les nœuds et/ou les arêtes ;
3. le calcul des mesures numériques (simples ou basées sur les graphes) pour le remplissage des cubes.

En raison des limitations de place, nous ne pouvons donner tous les détails. Pour les mesures numériques classiques qui sont basées sur des comptages, le calcul du cube se fait de façon classique par interrogation des données. Pour le calcul de mesures basées sur le graphe, cela nécessite des parcours de graphes, en fonction du type de la mesure. Par exemple, dans la Figure 6 est représenté le cube contenant la mesure numérique du degré de centralité de l'auteur J. Han en fonction des années et des conférences, et ici en particulier pour EDBT 2009. L'illustration de son calcul pour une cellule est donc basé sur le graphe associé à J. Han, où E1, E4 et E5 correspondent à des arêtes pour lesquelles des papiers à EDBT 2009 sont concernés (d'après les

informations stockées). Il s'agit alors, pour cette mesure de comptabiliser les arcs en question.

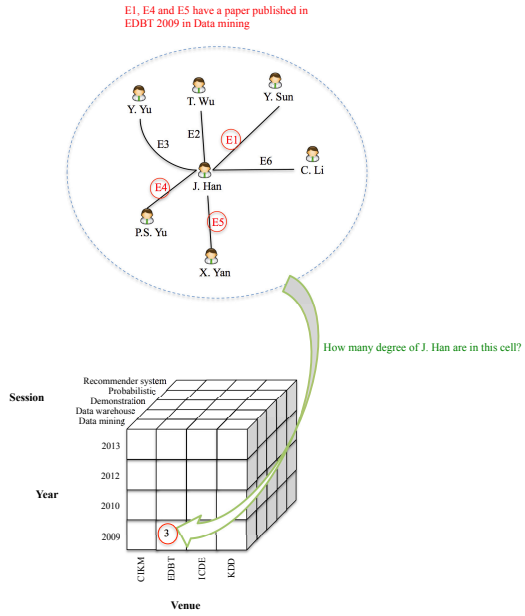

Figure 6. Illustration de cube contenant une mesure numérique calculée à partir du graphe.

Concernant les opérateurs OLAP, ceux-ci sont adaptés aux manipulations du graphe et des cubes. Cela concerne des opérateurs travaillant sur les informations hiérarchiques tels que *Roll Up* et *Drill Down*, mais aussi d'autres opérateurs qui permettent par exemple la sélection de données. Selon les cas, le graphe est amené à changer de structure (type de nœuds) ou les cubes sont recalculés.

Notons que pour éviter certains problèmes d'additivité, le retour aux données sources est souvent nécessaire pour différents calculs lors de la phase de manipulation des cubes ou du graphe, au travers d'une représentation à base de chemins.

Ce retour aux données sources est important d'un point de vue calculatoire pour deux raisons principales. La première est que cela permet de prendre en compte le fait que lorsqu'un *Roll Up* topologique est fait, les résultats demeurent cohérents. Par exemple, prenons le cas de l'analyse des co-publications avec le nombre de papiers comme mesure. Supposons qu'un papier a été écrit par deux auteurs du même établissement, ce papier sera comptabilisé pour les co-publications entre auteurs ; si nous passons au niveau des établissements, ce papier ne sera pas comptabilisé car il ne s'agit pas d'une collaboration inter-établissements. La deuxième raison réside dans le fait que nous prenons en compte l'évolution des données, notamment le fait qu'un auteur peut changer d'établissement dans le temps. Ainsi, nous nous ramenons tou-

jours à la donnée de base qui est la publication. Il s'agit de fait de pouvoir récupérer l'affiliation indiquée pour l'auteur dans le papier en question. Ainsi, l'affiliation d'un auteur qui évolue dans le temps est bien prise en compte dans les différents calculs de graphes et dans les opérations OLAP appliquées.

4. Implémentation et performances

4.1. Caractéristiques du prototype

L'implémentation de GreC a été réalisée en combinant les données bibliographiques de DBLP, ACM et Microsoft Research Area. Ceci est notamment justifié par la complémentarité des données, en termes de récupération des informations sur les affiliations des auteurs des papiers entre autres.

L'architecture de l'implémentation est présentée dans la Figure 7. Les données bibliographiques de base ont été centralisées dans le système NoSQL Neo4j. Les différents graphes correspondant aux différents faits et leurs cubes associés sont générés à partir des données de base du réseau hétérogène et des métadonnées. Les cubes sont ensuite stockés également dans Neo4j.

Les métadonnées sont stockées dans le système relationnel Oracle. Les interfaces pour l'utilisateur ont été développées en Java. Pour assurer la généricité de l'approche, leur contenu est généré en fonction des métadonnées, et des besoins d'analyse exprimés par l'utilisateur.

Concernant la navigation, comme une phase de pré-traitement permet de précalculer les éléments, cette brique consiste à la sélection et la visualisation des données adéquates.

4.2. Présentation des données de base

Pour permettre la mise en œuvre de l'approche *GreC*, les données de base doivent être modélisées. La Figure 8 montre le modèle proposé pour les données bibliographiques. Ce modèle couvre l'ensemble des données considérées et intègrent les liens entre les données à prendre en compte. Ces données de base correspondent à un graphe hétérogène.

4.3. Performances

L'étude de performances développée a permis d'analyser différents points. Elle a été conduite avec Java 1.7.0_75 sur un ordinateur portable avec un processeur Intel core i5 2.4 GHz avec 8 GB de RAM sur Mac OS X version 10.9.2. Quatre jeux de données de tailles différentes ont été constitués avec les caractéristiques décrites dans la Table 1, afin de mesurer l'approche en fonction du volume de données considérées.

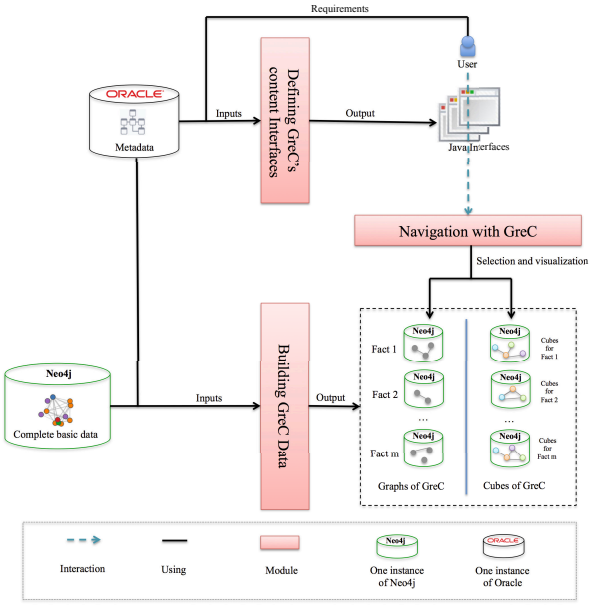

Figure 7. Architecture de l'implémentation de GreC.

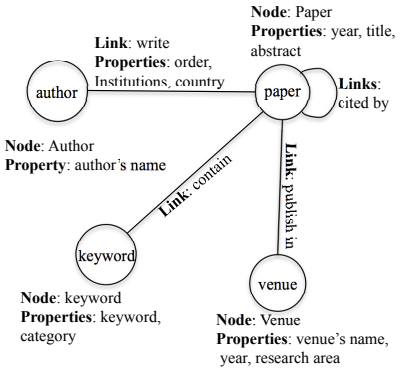

Figure 8. Modèle du graphe des données bibliographiques de base.

Table 1. Jeux de données pour l'expérimentation de GreC.

Jeux de Données	Nb de Publications	Réseau de co-auteurs		Réseau d'institutions	
		Nb de nœuds	Nb d'arêtes	Nb de nœuds	Nb d'arêtes
D1	1000	2216	4322	696	959
D2	2000	3790	8094	1157	1820
D3	3000	5335	12150	1573	2711
D4	4000	7038	16107	2051	3575

Le premier point porte sur la construction du graphe (des différents graphes) pour l'utilisateur. Cet algorithme est une optimisation d'un algorithme existant (Beheshti *et al.*, 2012). La Figure 9 reprend la construction du graphe de co-auteurs (Q1) et celle du graphe représentant les liens de co-publication inter-institutions (Q2). L'étude de performances démontre que sur des jeux de données de plus en plus importants, l'adaptation proposée est plus performante, et cela est possible car le nombre de lectures du graphe initial hétérogène a été optimisé.

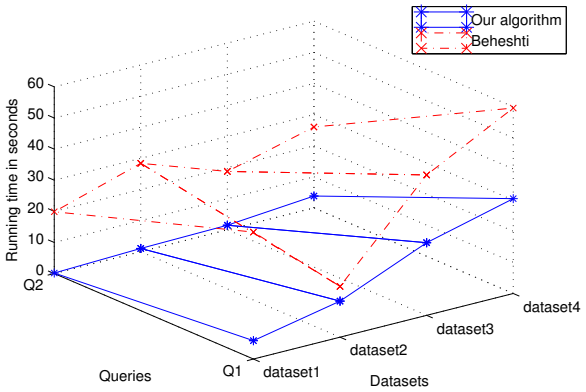

Figure 9. Temps d'exécution de la construction de graphe.

Le deuxième point correspond aux calculs des cubes, et donc a fortiori des mesures. Il apparaît que le temps est raisonnable lorsqu'il s'agit de mesures numériques simples. Le temps augmente fortement lorsqu'il s'agit de mesures basées sur le graphe (qui nécessitent un parcours de graphe pour le calcul), et ce de façon proportionnelle à la taille du graphe. Ceci dépend néanmoins du type de mesure, par exemple, si la mesure se base sur un calcul de voisinage comme le degré de centralité, le temps reste raisonnable comme le montre la Figure 10.

Le troisième point concerne la navigation par l'utilisateur dans les données. Cette étude montre que compte-tenu de l'intérêt d'une mesure à base de graphe en terme analytique et de l'enjeu du temps de traitement, le pré-calcul de ces informations est nécessaire, mais pourrait être partiel selon le contexte . Cette étude montre que compte-tenu des choix de pré-calcul, des temps acceptables de navigation sont obtenus selon les types de requêtes testés et le volume de données. En effet, si les requêtes de type sélective sur des membres de dimension peuvent aller jusqu'à une dizaine de secondes, globalement, le temps d'exécution des requêtes est largement inférieur à 1 seconde pour des requêtes de type *Roll Up* des auteurs aux institutions par exemple. Ainsi, pré-calculer les graphes et les cubes constitue une stratégie intéressante pour optimiser les temps d'exécution pour l'utilisateur.

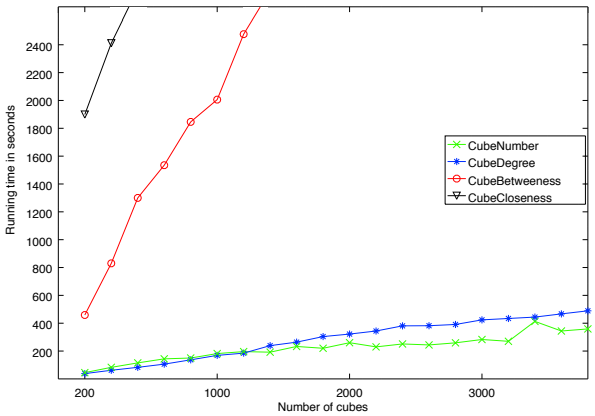

Figure 10. Temps d'exécution de la construction des cubes.

5. Conclusion et perspectives

L'approche GreC proposée constitue une vision innovante et complémentaire dans le domaine du *Graph OLAP*, en permettant à la fois une vue globale du graphe, tout en l'enrichissant d'informations pertinentes et riches au travers de cubes qui valuent les nœuds et/ou les arêtes selon les besoins d'analyse. La présence d'une dimension temporelle dans ces cubes permet d'avoir des éléments sur la dynamique du graphe et de prendre en compte les modifications de données au cours du temps. Nous nous sommes particulièrement focalisés ici sur l'explicitation des éléments qui permettent de donner un caractère générique à cette approche. Nous avons également précisé l'implémentation de *GreC* et donné des éléments de performance qui tendent à montrer la pertinence de l'approche à la fois du point de vue des possibilités d'analyse, mais également de la rapidité de leur obtention selon des choix qui optimisent le temps pour la navigation dans les données.

Ce travail ouvre de nombreuses perspectives. Une première perspective consiste à doublement étendre les possibilités d'analyse offertes par *GreC*. D'une part, il s'agit d'explorer la possibilité d'utiliser des mesures de centralité pour les arêtes. D'autre part, nous voulons introduire dans *GreC* des mesures textuelles. En effet, une grande partie de l'information contenue dans les réseaux d'information est textuelle. L'idée est de combiner les approches *Graph OLAP* aux approches *Text OLAP* afin de proposer une approche complète pour l'analyse des réseaux d'information.

La deuxième perspective concerne l'analyse de l'évolution du graphe. Au-delà de l'aspect temporel des cubes, une piste à explorer serait d'envisager des opérations binaires (différence, intersection, etc.) entre deux graphes issus de *GreC*. Ceci induit de redéfinir ces opérateurs au regard de l'approche *GreC*.

Concernant les données bibliographiques notamment, les publications sont souvent écrites par plus de deux auteurs. Cela pose alors la question de la possibilité

d'avoir recours aux hypergraphes, avec toutes les adaptations qui découleraient de ce choix.

Enfin, il s'agit de se focaliser davantage sur l'utilisateur, avec d'une part développer la possibilité de mieux cerner le graphe ou sous-graphe à analyser (système de recommandation par exemple) et de procéder à une évaluation utilisateur à grande échelle, en terme non seulement d'usage et de performances.

References

Beheshti S.-M.-R., Benatallah B., Motahari-Nezhad H. R., Allahbakhsh M. (2012). A framework and a language for on-line analytical processing on graphs. In *13th International Conference on Web Information Systems Engineering (WISE'12)*, p. 213-227.

Chen C., Yan X., Zhu F., Han J., Yu P. S. (2008). Graph OLAP: Towards online analytical processing on graphs. In *8th IEEE International Conference on Data Mining (ICDM'08)*, p. 103-112.

Jakawat W., Favre C., Loudcher S. (2016a). Graphs enriched by cubes for OLAP on bibliographic networks. *International Journal of Business Intelligence and Data Mining*, Vol. 11, No. 1, pp. 85–107.

Jakawat W., Favre C., Loudcher S. (2016b). OLAP cube-based graph approach for bibliographic data. In *42nd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM'16), Student Research Forum.*

Jin X., Han J., Cao L., Luo J., Ding B., Lin C. X. (2010). Visual cube and on-line analytical processing of images. In *19th ACM International Conference on Information and Knowledge Management (CIKM'10).*

Loudcher S., Jakawat W., Morales E. P. S., Favre C. (2015). Combining OLAP and information networks for bibliographic data analysis: a survey. *Scientometrics*, Vol. 103, No. 2, pp. 471–487.

Morfonios K., Koutrika G. (2008). Olap cubes for social searches: Standing on the shoulders of giants? In *International Workshop on the Web and Databases (WebDB).*

Qu Q., Zhu F., Yan X., Han J., Yu P., Li H. (2011). Efficient topological olap on information networks. In *Proceedings of the 16th International Conference on DAtabase Systems For Advanced Applications (DASFAA'11)*, Vol. 1, p. 389-403.

Tian Y., Hankins R., Patel L. (2008). Efficient aggregation for graph summarization. In *ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*, p. 567-580.

Yin M., Wu B., Zeng Z. (2012). Hmgraph olap: a novel framework for multi-dimensional heterogeneous network analysis. In *15th International Workshop on Data warehousing and OLAP (DOLAP'12)*, p. 137-144.

Zhao P., Li X., Xin D., Han J. (2011). Graph cube: On warehousing and olap multidimensional networks. In *ACM SIGMOD International Conference on Management of Data (SIGMOD'11)*, p. 853-864.