Thèse de Wararat JAKAWAT – Résumé détaillé

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

## 1. Contexte

Depuis plusieurs années, l'informatique décisionnelle a connu un essor important, à la fois en terme de projets pour les entreprises afin de se doter d'outils pour permettre un support aux décisions, notamment grâce à des systèmes proposant une navigation adéquate dans les données, mais également en terme de sujet de recherche, en particulier dans le milieu académique.

Dans ce cadre, l'OLAP (OnLine Analytical Processing) est alors devenu une brique importante des architectures décisionnelles, mais également un objet de recherche scientifique, étudié sous différents angles, parmi lesquels la prise en compte de la complexité des données.

L'OLAP, en se basant sur une modélisation multidimensionnelle des données et différents opérateurs, a pour objectif d'aider l'utilisateur à naviguer dans les données, à les résumer, à les détailler… Cette modélisation multidimensionnelle introduit la notion de fait à analyser, au travers d'indicateurs dénommés mesures, en fonction d'axes d'analyse nommés dimensions. Ces axes d'analyse peuvent être organisés selon différents niveaux de détails, d'où le terme de hiérarchie de dimension.

Historiquement, l'OLAP a été appliqué dans un contexte de données assez classiques, notamment dans le cas de données commerciales. L'émergence de nouveaux types de données à considérer, comme par exemple 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, 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.

Tout au long de la thèse, bien que l'approche développée soit générique, le travail a été appliqué au domaine des données bibliographiques, tentant de contribuer à l'analyse de ces données, et de fait au domaine de la scientométrie. Ces données se prêtent particulièrement bien à la représentation sous forme de graphes, et ces données ont d'ailleurs fait l'objet des premières approches qui ont tenté de combiner les graphes et l'approche OLAP, ce qui a donné lieu au « Graph-OLAP », qui constitue le contexte de cette thèse.

## 2. Préambule

Afin d'illustrer nos propos, un exemple de données bibliographiques est utilisé.

Il apparaît qu'une des caractéristiques importantes de ces 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 dans un même papier 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.

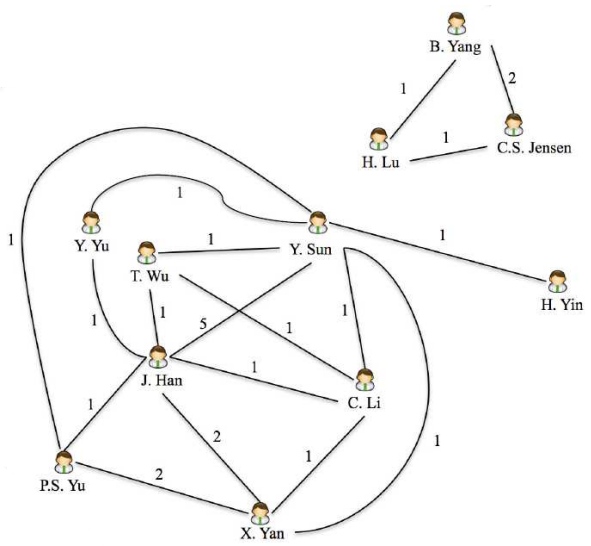


*Figure 1 : Graphe d'auteurs représentant les co-publications*

Néanmoins, dans cet exemple, il ait aisé de 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.

Ainsi, 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** production scientifique, au travers d'une **mesure** qui est le nombre de publications, en fonction de différentes **dimensions** qui sont ici : auteur, temps, et conférence. Ces dimensions peuvent être organisées sous forme de **hiérarchies**, comme c'est le cas pour les conférences, qui sont organisées en domaines. La métaphore du cube de données est utilisée lorsque trois dimensions sont mobilisées. A l'intersection des valeurs prises par ces dimensions, il est possible de visualiser la mesure au cœur de la cellule. Par exemple J. Han a contribué à deux publications publiées à la conférence EDBT en 2009. Notons par ailleurs que l'analyse en ligne fournit des opérateurs pour ensuite naviguer dans ces données, selon un modèle multidimensionnel qui a été préétabli. Ceci permet par exemple d'obtenir un cube, en ne considérant plus les conférences, mais les domaines de celles-ci, au travers d'une opération « Roll Up » qui fournirait donc le nombre de publications agrégé par auteur, par année et par domaine.

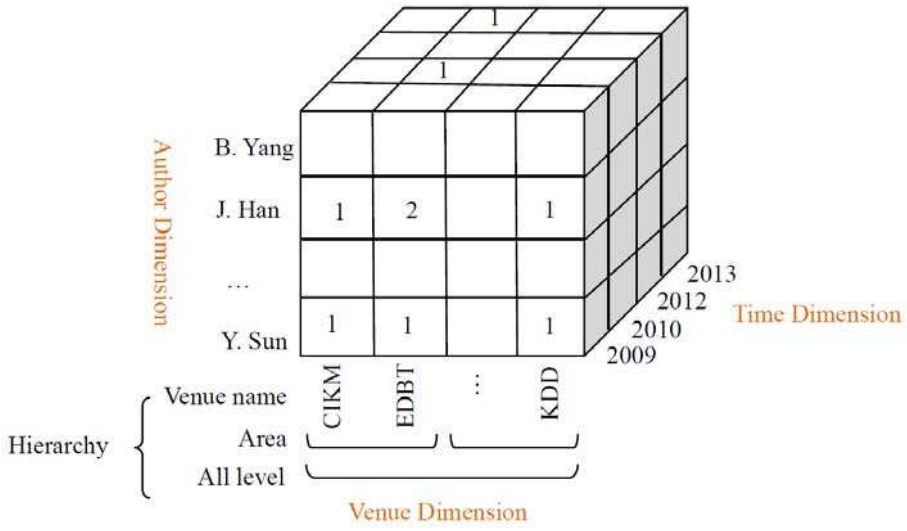


*Figure 2 : structure d'un cube de données dans le contexte OLAP*

Dans cette représentation 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. Ainsi, afin de tirer parti de ces deux visualisations (graphe et cube), un nouveau champ de recherche est apparu portant comme dénomination « Graph OLAP » [CYZ+08]. Cette approche a fait l'objet de plusieurs publications proposant des améliorations et des extensions. Nous présentons ici l'idée sur laquelle elle repose qui consiste en la construction

d'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 contexte d'analyse.

Dans cette approche « Graph OLAP », il s'agit ainsi de considérer des cubes définis selon des dimensions (dites informationnelles), et la mesure contenue dans les cellules correspond à aux sous-graphes adéquats. Par exemple, dans la Figure 3, par rapport aux données considérées ici, pour la conférence EDBT en 2013, le réseau est composé des co-auteurs B. Yang et C.S. Jensen qui ont co-publié un papier, en considérant le nombre de papiers qui value les arêtes du graphe.

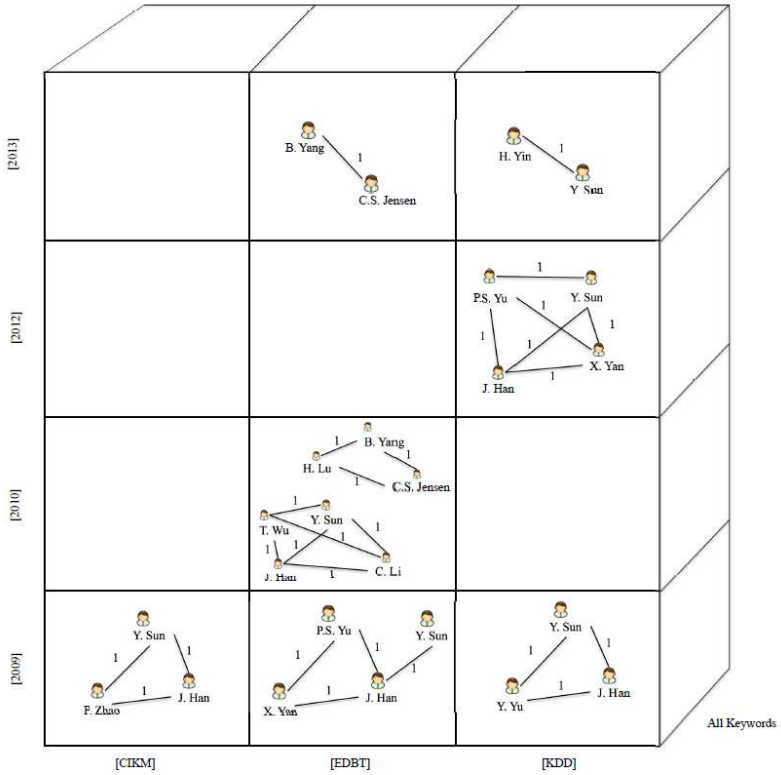


*Figure 3 : Cube de graphes sur des données bibliographiques*

Notons que dans cette approche de « Graph OLAP », au niveau de la modélisation, ont été définis deux types de dimensions : les dimensions informationnelles et les dimensions topologiques. Les dimensions informationnelles vont conditionner les manipulations du cube. Lors d'opérations informationnelles, les graphes à l'intérieur des cellules vont être recalculés. Les dimensions topologiques, quant à elles, se rapportent à la modélisation des réseaux 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, 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.

Cette approche permet alors de manipuler à la fois les éléments définissant le cube et les éléments des graphes.

## 3. Motivation et contributions

Initialement, si la combinaison de l'OLAP et des graphes s'est donc faite au travers d'une proposition de cubes de graphes [CYZ+08], nous proposons dans le cadre de cette thèse une nouvelle façon de considérer la combinaison de l'OLAP et des graphes. Précédemment, il s'agissait de considérer des cubes de données, avec dans leurs cellules, un graphe comme mesure. Cette approche permet de visualiser des « instantanés » de graphes en fonction des dimensions d'analyse choisies. Différents opérateurs ont été proposés pour naviguer dans le cube de graphes : en distinguant des opérations informationnelles ou topologiques, selon si les opérations s'opèrent selon les dimensions du cube ou les dimensions des graphes.En revanche, dans cette approche, 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 son évolution, 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 l'évolution même d'un graphe (évolution des arêtes ou des nœuds).

Ainsi, nous avons voulu développer une nouvelle approche de « Graph-OLAP » combinant l'analyse OLAP et les graphes d'une façon originale, et sans doute complémentaire des premières approches s'inscrivant dans une construction d'un cube de graphes. En effet, plutôt que de visualiser des cubes de graphes, l'approche que nous proposons est de construire un graphe qui répond aux besoins d'analyse de l'utilisateur et de l'enrichir avec des cubes de données qui vont valuer les nœuds et/ou les arêtes selon les besoins d'analyse. Ainsi, la présence d'une dimension temporelle au niveau des cubes qui valuent les nœuds et/ou les arêtes va notamment permettre de rendre compte de l'évolution du graphe. Par ailleurs, pour permettre une richesse d'analyse, notre attention s'est focalisée sur deux apports : d'une part les types de mesure possibles ; d'autre part les opérateurs de navigation proposés. Ce travail a donné lieu à l'implémentation d'un prototype pour étudier la faisabilité de l'approche. Et, compte-tenu de l'importance pour l'utilisateur d'obtenir des temps de réponse satisfaisants, une étude des performances a été réalisée.

L'état de l'art réalisé dans le cadre de cette thèse, ainsi que les différentes contributions ont donné lieu à des publications de différentes formes : journaux internationaux [LJMF15, JFL16], conférence internationale [JFL15] et nationale [LFJ13], atelier de portée internationale [JFL13].

## 4. L'approche GreC

Pour parvenir à développer l'approche GreC (Graphes enrichis par des Cubes), cette thèse présente différentes contributions. Premièrement, il s'agit du cadre général de l'approche. Ensuite, nous focalisons notre attention sur l'aspect des données avec d'une part les données de base et leur représentation, et d'autre part, les méta-données qui permettent d'assurer la généricité de l'approche, au-delà du contexte des données bibliographiques. Puis il s'agit de présenter la brique qui permet de construire les graphes GreC, ce qui nécessite tout d'abord de redéfinir/d'étendre les concepts manipulés dans le contexte de GreC. Il s'agit alors de pouvoir proposer les éléments qui permettent la navigation de l'utilisateur (opérateurs). Seront enfin présentés les éléments relatifs à l'implémentation et à l'étude de performances. Chacun de ces éléments est détaillé par la suite.

### a. Cadre général de GreC

Dans la Figure 4, le cadre général de GreC permet de découvrir le fonctionnement global de l'approche. Le point de départ est la phase préparatoire (couche A), qui correspond au pré-traitement des données. A partir de diverses bases de données bibliographiques, celles-ci sont fusionnées, intégrées dans des fichiers XML qui permettent d'alimenter une base de données en graphes selon le modèle de données défini (cf point b.). Il s'agit ici de considérer un graphe hétérogène comportant l'ensemble de toutes les données. Ensuite, il y a la strate de GreC (couche B) avec les graphes correspondants. En effet, à partir du graphe hétérogène des données de base, la construction des graphes enrichis par les cubes est réalisée (notons que l'ensemble des graphes est calculé en amont pour assurer de bonnes performances pour l'utilisateur comme ce sera précisé dans le point g.). Cette construction se décompose en deux étapes : construction du graphe lui-même, puis celle des cubes de données ensuite pour valuer les nœuds et/ou les arêtes. Ces deux étapes se répètent pour construire l'ensemble des graphes enrichis par les cubes.

Côté utilisateur, avec la navigation (couche C), l'expression des besoins d'analyse 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.

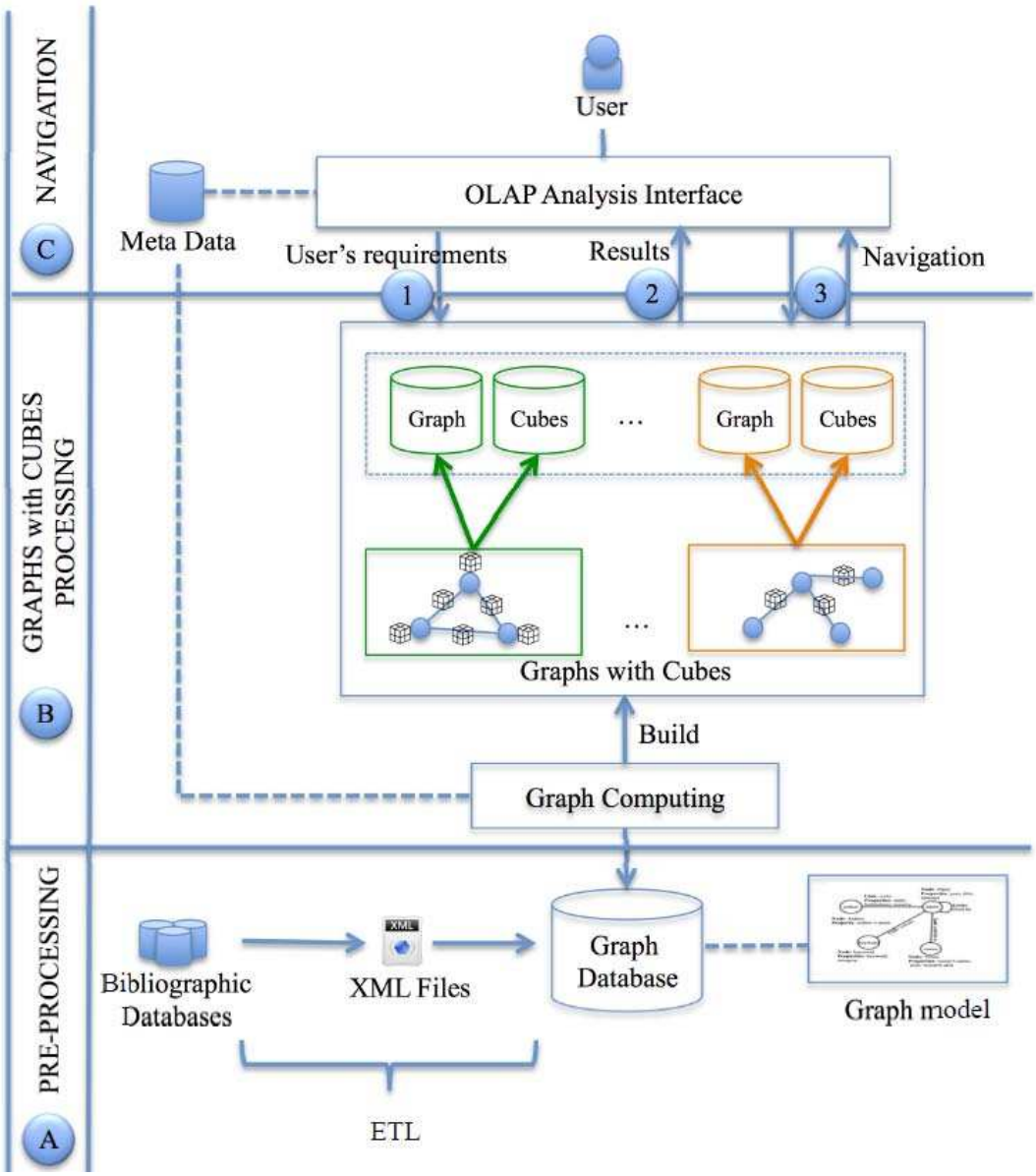


*Figure 4 : Processus de GreC*

## b. Modèle de graphe

Pour permettre la mise en œuvre de notre approche, il est nécessaire que les données de base à considérer soient modélisées (Figure 5). Concernant les données bibliographiques considérées dans notre travail, il s'agissait donc de proposer un modèle retraçant l'ensemble des données et leurs liens à prendre en compte. Ces données de base correspondent à un graphe hétérogène. A partir des données brutes représentées dans ce graphe hétérogène, pour en extraire des graphes enrichis par les cubes et assurer la généricité de l'interface qui va permettre de les manipuler, nous avons développé un modèle de métadonnées.

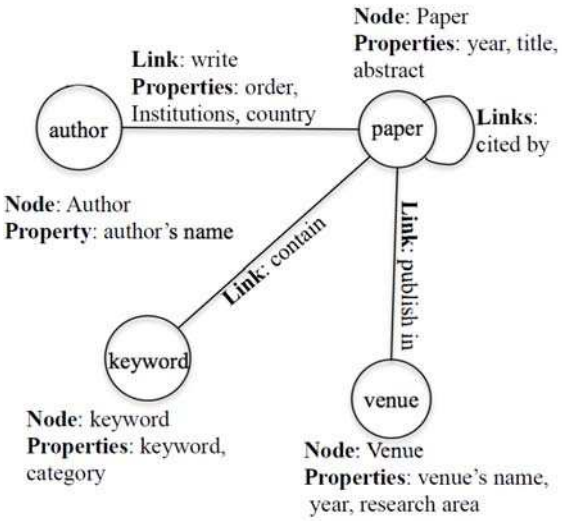


*Figure 5 : Modèle du graphe des données bibliographiques de base*

## c. Métadonnées

La Figure 6 permet de visualiser le modèle des métadonnées qui a été conçu pour assurer la généricité de l'approche sur différents aspects : à la fois pour permettre de faire le lien entre l'objet d'analyse, les graphes à construire et l'emplacement des cubes (au niveau des nœuds et/ou des arêtes) et les concepts OLAP nécessaires (leur instanciation) pour l'approche : faits, mesures, dimensions, etc. Nous revisitons ces concepts dans la section suivante par rapport à l'approche GreC.

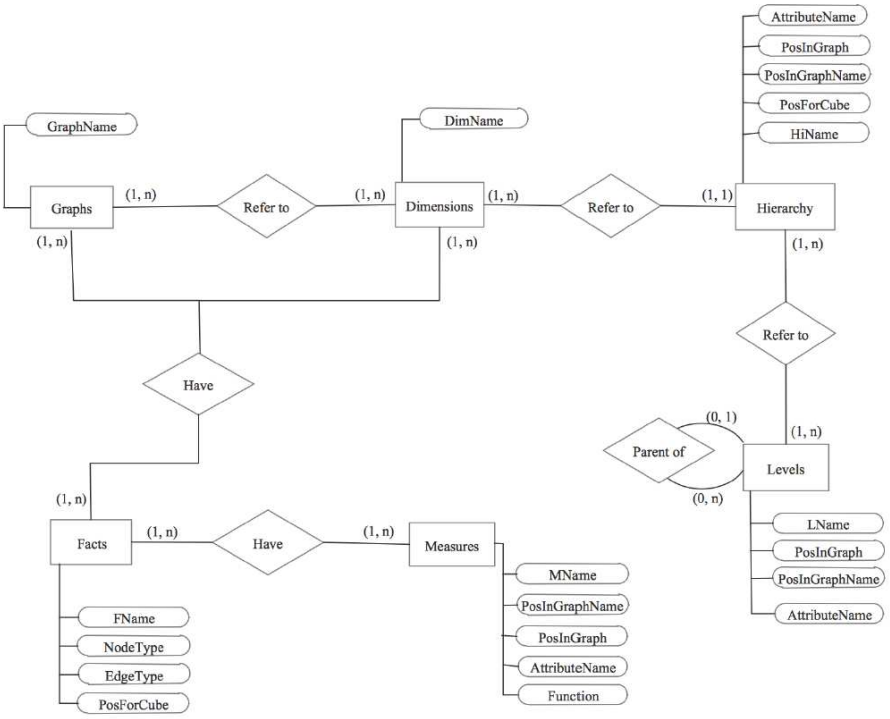


*Figure 6 : Modèle des métadonnées de GreC*

### d. Revisiter les concepts classiques et les étendre

Tout comme dans l'approche classique d'analyse en ligne, dans l'approche 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 général. En fonction du fait, et des besoins d'analyse, les métadonnées permettent de déterminer si des cubes de données valuent des nœuds et/ou des arêtes.

La notion de cube correspond alors à un cube OLAP classique, qui contient dans chacune de ses cellules une mesure numérique ; cette mesure peut-être « simple » (additive) comme le nombre de publications ou une mesure basée sur des graphes comme par exemple une mesure de centralité, là où pour l'autre approche de « Graph OLAP », la mesure était un graphe.

Comme dans l'approche « Graph OLAP » initiale, nous retrouvons deux types de dimension : dimension informationnelle et dimension topologique. Les dimensions informationnelles correspondent donc aux dimensions définissant les cubes de données attenants aux nœuds et/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. Nous parlons alors, non pas d'opérateur OLAP, mais d'opérateur OLAP informationnel ou topologique, déterminant ainsi si l'opération (que ce soit un « Roll Up », « Drill Down », etc.) est appliquée par rapport au graphe en question, ou aux cubes de ce graphe.

### e. Algorithmes

Rappelons que le point de départ correspond à 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.

Les différents algorithmes proposés pour implémenter GreC recouvrent différentes étapes du processus :

1/ la construction du graphe pour l'utilisateur, en fonction du graphe de départ, des besoins d'analyse de l'utilisateur et des méta-données

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)

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.

### f. Implémentation

L'implémentation de GreC a été réalisée en combinant les données bibliographiques de DBLP, ACM et Microsoft Research Area. 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. Ils 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 pré-calculer les éléments, cette brique consiste à la sélection et la visualisation des données adéquates.

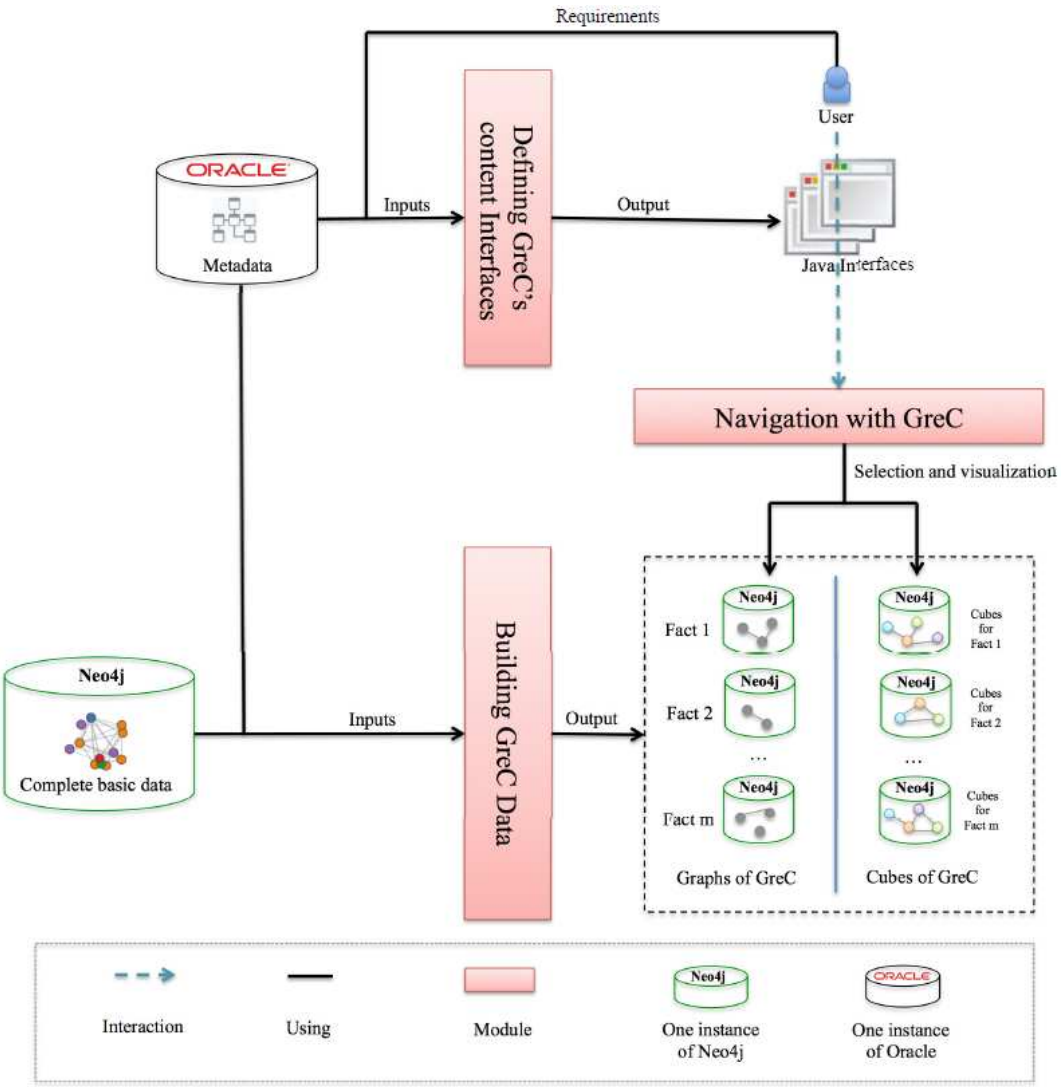


*Figure 7 : Architecture de l'implémentation de GreC*

## g. Performances

L'étude de performances développée a permis d'analyser différents points. Le premier porte sur la construction du graphe pour l'utilisateur. Cet algorithme est une adaptation de celui de [BBMA12]. L'étude de performances démontre que sur des jeux de données de plus en plus importants, l'adaptation proposée est plus performante car elle permet de limiter le nombre de lectures. Le deuxième point correspond aux calculs des cubes, et donc a fortiori des mesures. Il apparaît que le temps est raisonnable, sauf pour les mesures à base de graphes. Cette étude montre que compte-tenu de l'intérêt d'une mesure à base de graphe et de l'enjeu du temps de traitement, le pré-calcul de ces informations est nécessaire. Le troisième point concerne la navigation par l'utilisateur dans les données. Cette étude montre que compte-tenu des choix de pré-calcul, des temps acceptables de navigation sont obtenus.

## 5. Discussion et perspectives

L'approche GreC proposée constitue une nouvelle vision complémentaire dans le domaine du « Graph OLAP », en permettant à la fois une vue globale du graphe, tout en l'enrichissant d'informations au travers de cubes qui valuent les nœuds et/ou les arêtes. 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.

Ce travail a ouvert de nombreuses perspectives, les principales sont exposées ici.

Premièrement, il s'agit d'étendre les possibilités d'analyse offertes par GreC. Pour ce faire, il s'agit d'explorer la possibilité d'utiliser des mesures de centralité pour les arêtes, d'avoir recours à des mesures textuelles…

Deuxièmement, concernant 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 à partir de deux graphes issues de l'approche GreC : la différence, l'intersection, etc. Ceci induit de redéfinir ces opérateurs au regard de l'approche GreC.

Parallèlement, le recours à la visualisation de graphe comme nous l'avons présenté permet de porter attention au lien entre deux nœuds, entre deux auteurs par exemple. Or, 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.

## 6. Références principales

[BBMA12] Seyed-Mehdi-Reza Beheshti, Boualem Benatalla, Hamid Reza Motahari-Nezhad, and Mohammad Allahbakhsh. A framework and a language for on-line analytical processing on graphs. In 13th International Conference on Web Information Systems Engineering (WISE'12), pages 213-227, 2012.

[CYZ+08] Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S Yu. Graph OLAP: Towards online analytical processing on graphs. In 8th IEEE International Conference on Data Mining (ICDM'08), pages 103-112, 2008.

[JFL13] Wararat Jakawat, Cécile Favre, and Sabine Loudcher. OLAP on information networks: A new framework for dealing with bibliographic data. In 1st International Workshop on Social Business Intelligence Bibliography (SoBI'13), collocated with the East-European Conference on Advances in Databases and Information Systems (ADBIS'13), pages 361-370, 2013.

[JFL15] Wararat Jakawat, Cécile Favre, and Sabine Loudcher. 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, 2015.

[JFL16] Wararat Jakawat, Cécile Favre, and Sabine Loudcher. Graphs enriched by cubes for OLAP on bibliographic networks. International Journal of Business Intelligence and Data Mining (IJBIDM'16), 11(1):85-107, 2016.

LFJ13] Sabine Loudcher, Cécile Favre, and Wararat Jakawat. Que peut apporter l'OLAP à l'analyse de réseaux d'informations bibliographiques ? In 4ème conférence sur les modèles et l'analyse des réseaux : approches mathématiques et informatiques (MARAMI'13), 2013.

[LJMF15] Sabine Loudcher, Wararat Jakawat, Edmundo Pavel Soriano Morales, and Cécile Favre. Combining OLAP and information networks for bibliographic data analysis: a survey. Scientometrics, 103(2):471-487, 2015.