Analyse en ligne d'objets complexes avec l'analyse factorielle

Loic Mabit*, Sabine Loudcher*, Omar Boussaid*

*Laboratoire ERIC, Université Lumière Lyon 2
Université de Lyon
5 avenue Pierre Mendès-France
69676 Bron Cedex
loic.mabit@gmail.com
{sabine.loudcher | omar.boussaid}@univ-lyon2.fr

Résumé. Les entrepôts de données et l'analyse en ligne OLAP (*On-line Analysis Processing*) présentent des solutions reconnues et efficaces pour le processus d'aide à la décision. Notamment l'analyse en ligne, grâce aux opérateurs OLAP, permet de naviguer et de visualiser des données représentées dans un cube multidimensionnel. Mais lorsque les données ou les objets à analyser sont complexes, il est nécessaire de redéfinir et d'enrichir ces opérateurs OLAP. Dans cet article, nous proposons de combiner l'analyse OLAP et la fouille de données (*data mining*) afin de créer un nouvel opérateur de visualisation d'objets complexes. Cet opérateur utilise l'analyse factorielle des correspondances.

1 Introduction

Les entrepôts de données et l'analyse en ligne OLAP (*On-line Analysis Processing*) présentent des solutions reconnues et efficaces pour le processus d'aide à la décision. Notamment l'analyse en ligne, grâce aux opérateurs OLAP, permet de naviguer et de visualiser des données représentées dans un cube multidimensionnel. Cette technologie est bien rodée quant il s'agit de données simples où les faits sont analysés à travers des indicateurs numériques, souvent additifs, selon des descripteurs en général qualitatifs. Cependant, l'avènement des données complexes a remis en cause ce processus d'entreposage et d'analyse en ligne.
Souvent les données complexes sont plutôt représentées par un ensemble de descripteurs de bas niveau et/ou sémantiques. Lorsqu'il s'agit d'analyser par exemple une image, une vidéo ou tout autre objet de l'univers, il est alors plus efficace de considérer chacun de ces éléments comme une entité, à part entière, à observer. Celle-ci est alors perçue comme un objet complexe. Ce dernier peut être considéré comme un agrégat hétérogène de données qui, une fois réunies, forment une unité sémantique.
La vocation de l'analyse en ligne (OLAP) est de permettre d'agréger des données pour résumer l'information qu'elles contiennent et de représenter celle-ci sous différents angles. L'utilisateur peut alors naviguer dans les données afin de les explorer. Les opérateurs OLAP sont définis pour des données classiques. Ils sont par conséquent inadaptés quand il s'agit d'objets complexes. Le recours à d'autres techniques, par exemple de fouille de données, peut s'avérer

intéressant (Ben Messaoud (2006), Imielinski et Mannila (1996), Han (1997)).

Nous proposons dans ce papier une approche d'analyse en ligne d'objets complexes en utilisant l'analyse factorielle. Cette technique permet de représenter les données en les projetant sur des axes factoriels. Elle sera ainsi utilisée comme une opération de visualisation dans le cadre de l'analyse en ligne. Pour illustrer nos propos, nous avons effectué une étude de cas sur les publications des chercheurs d'un laboratoire. En effet, une publication peut être considérée comme un objet complexe du fait que nous souhaitons l'observer comme une entité sémantique. Nous envisageons de l'analyser selon son premier auteur et ses co-auteurs, sa portée nationale ou internationale, son support tel un congrès national ou international ou un journal, etc. Nous voulons observer la diversité des thématiques dans lesquelles publient les chercheurs ainsi que la proximité de ces derniers lorsqu'ils travaillent dans les mêmes domaines. Comme nous pouvons le remarquer, contrairement à l'analyse en ligne (OLAP) classique qui propose des opérateurs arithmétiques pour agréger des données numériques, nous avons besoin, dans notre cas, de se baser sur le contenu sémantique des publications, tels que les mots-clés par exemple, pour étudier les objectifs d'analyse que nous venons de citer plus haut.

Pour présenter cette nouvelle approche d'analyse en ligne sur des objets complexes, nous avons organisé ce papier comme suit. Nous allons d'abord développer le principe de notre démarche dans la section 2. Nous présentons dans la section 3, l'étude de cas sur les publications qui illustre notre démarche. Nous avons validé cette dernière par l'implémentation d'une plate-forme logicielle que nous présentons succinctement dans la section 4. Nous terminons cet article par une section discussion et conclusion (section 5).

2 Principe et démarche

Dans les entrepôts de données, les données sont souvent représentées selon un modèle en étoile, Kimball et Merz (2000), Inmon (1996). Celui-ci est composé d'une table de faits centrale contenant une ou plusieurs mesures à observer, de tables de dimensions comprenant des descripteurs. Ces tables sont alors représentées dans une structure multidimensionnelle adaptée à l'analyse qu'on appelle les cubes de données. Un fait est donc représenté par un ensemble de modalités provenant des dimensions d'un cube et observé par une ou plusieurs mesures ayant des propriétés d'additivité plus ou moins fortes. La vocation de l'OLAP est de fournir à l'utilisateur des opérateurs pour résumer et naviguer dans les données afin d'y découvrir des informations pertinentes, Chaudhuri et Dayal (1997).

Cependant, dans le cas des cubes de données complexes, les faits représentent des objets complexes, les dimensions peuvent comporter des images, du texte, des descripteurs, ... Les mesures ne sont pas forcément additives. Compte tenu de ces particularités, les opérateurs usuels OLAP de navigation ne sont pas adaptés ou ne peuvent pas être utilisés. Il est donc nécessaire de définir de nouveaux opérateurs adaptés pour la navigation et la visualisation dans un cube de données complexes. Pour cela, nous proposons d'utiliser les principes des méthodes factorielles bien connues en fouille de données notamment grâce aux travaux de Benzecri (1982) et de L. Lebart et Piron (2004). Une méthode factorielle va permettre de visualiser les objets complexes tout en mettant en évidence les points de vue intéressants pour l'analyse. La méthode factorielle va ajuster au mieux le nuage ou le cube d'objets complexes. Quand les

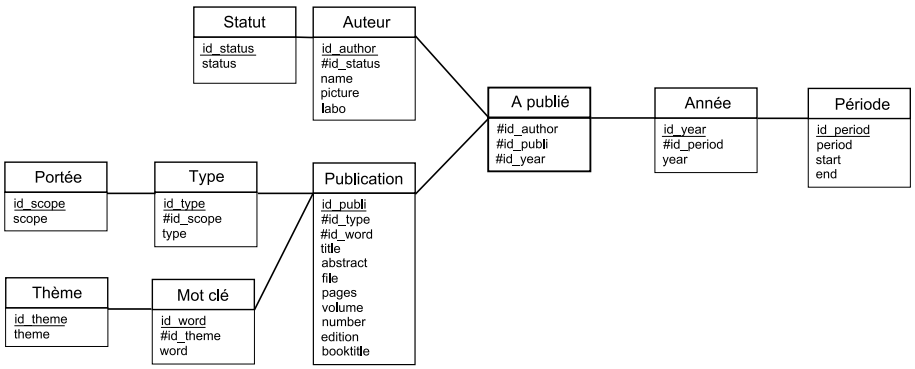

FIG. 1 – *Schéma de l'entrepôt de données*

faits représentent des objets complexes, souvent il n'y a pas de mesure au sens classique de la modélisation multidimensionnelle. En revanche, il est toujours possible de dénombrer les faits. La mesure COUNT est parfois la seule mesure utilisable quand on travaille dans des cubes de données complexes. Dans ce cas là, le cube de données complexes avec plusieurs dimensions et la mesure COUNT pour dénombrer les objets complexes peut être vu comme un tableau de contingence ou d'occurrences. Il est alors possible d'utiliser une analyse des correspondances pour visualiser les faits.

L'analyse des correspondances produit des axes factoriels qui peuvent être utilisés comme nouvelles dimensions, à savoir des "dimensions factorielles". Ces nouveaux axes ou dimensions constituent un nouvel espace de représentation dans lequel il est possible de projeter les faits ou les objets complexes. Le premier plan factoriel (c'est à dire l'espace de représentation créé à partir des deux premiers axes produits par l'analyse des correspondances) permet de visualiser les objets complexes contenus dans le cube. Ce plan factoriel est aussi appelé carte factorielle. L'utilisation de l'analyse des correspondances comme opérateur de visualisation se justifie pleinement car l'analyse des correspondances reprend l'objectif de navigation et d'exploration de l'OLAP. En plus d'un principe, nous proposons à l'utilsateur OLAP une démarche complète en plusieurs étapes : (1) choix du contexte d'analyse, (2) construction du cube d'objets complexes, (3) construction du tableau d'occurrences, (4) réalisation de l'analyse des correspondances, (5) visualisation des objets complexes sur la carte factorielle.

3 Etude de cas

Présentation. Pour illustrer notre démarche, nous nous sommes appuyés sur l'exemple de l'analyse des publications de notre laboratoire de recherche. Dans cet exemple, l'objectif est de synthétiser nos travaux à partir de mots-clés résumant nos publications afin d'identifier les grands champs de recherche sur lesquels nous travaillons. Pour répondre à cette problématique, nous proposons un outil d'analyse en ligne offrant la possibilité de visualiser les publications à travers leurs auteurs et les mots-clés sur une carte factorielle.

Nous proposons d'analyser le fait d'avoir publié. Nous modélisons ce fait par un schéma en étoile avec trois dimensions : le temps, les auteurs, les publications (figure 1). Les données présentent la particularité d'être complexes dans la mesures où elles sont représentées par différents types de données et qu'elles sont soumises à une structure particulière. Par exemple, un auteur est représenté par son nom (chaîne de caractère structurée), par sa photographie

(image), son statut (professeur, maître de conférences, doctorant, ...), les dates d'arrivée et/ou de départ dans le laboratoire, etc. De plus, certaines dimensions ont une structure particulière comme par exemple la dimension "auteur" qui est évolutive ou la dimension "publication" qui est multiple.

Contexte d'analyse et cube d'objets complexes. L'étape suivante consiste à définir le contexte d'analyse. Celui-ci est défini par l'utilisateur, compte tenu de l'analyse qu'il souhaite réaliser. Par exemple, l'utilisateur peut choisir de travailler sur les publications de type "conférences" ayant été écrites au cours des trois dernières années, par des auteurs qui ont le statut de maître de conférences. La structure multidimensionnelle de l'entrepôt de données permet de construire le cube de donnée OLAP correspondant au choix de l'utilisateur. Dans le contexte des publications de type "conférences" ayant été écrites au cours des dernières années par des auteurs qui ont le statut de "maître de conférences", l'utilisateur peut construire, un cube des publications selon les mots-clés, l'année de parution et le nom du premier auteur (figure 2).

Tableau d'occurrences. L'analyse factorielle des correspondances admet en entrée un tableau d'occurrences. Dans notre démarche d'analyse en ligne d'objets complexes, l'idée est d'utiliser les opérateurs traditionnels de l'OLAP pour construire ce tableau d'occurrences. Dans notre exemple, celui-ci correspond au croisement l'ensemble des mots-clés avec l'ensemble des auteurs retenus dans le contexte d'analyse. Étant donné la structure de l'entrepôt de données, la démarche consiste à dénombrer les faits sur toutes les années en effectuant un roll-up sur la dimension "année". Nous obtenons ainsi un tableau bidimensionnel admettant les mots-clés pour ligne et les auteurs pour colonne (figure 2).

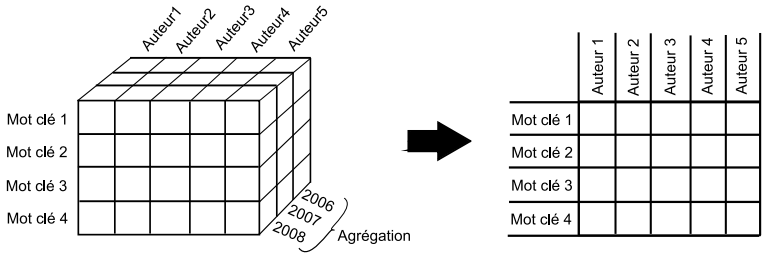

FIG. 2 – *Construction d'un tableau d'occurrences*

Analyse factorielle. La méthode démarre par le calcul des valeurs propres à partir desquelles sont déductibles les vecteurs propres qui définissent les axes factoriels. Comme ce sont les deux premiers axes qui contiennent le plus d'informations, la projection des données sur ceux-ci se révèle souvent pertinente. Sur notre exemple des publications, en se positionnant sur le premier plan factoriel, les publications (représentées par leurs mots-clés) et leurs auteurs s'associent de manière plus ou moins marquée. La figure 3 est un exemple très concis de ce qu'on peut observer sur les données réelles : les mots-clés 1 et 2 sont très proches ce qui laisse supposer que les publications concernées ont été écrites par des auteurs communs. En revanche, les mots-clés 3 et 4 sont assez distants et paraissent relever par conséquent de deux champs de recherche différents. Le placement des auteurs permet quant à lui d'avoir un aperçu sur le(s) champs de recherche auquel(s) ils se rattachent. Par exemple, les auteurs 3 et 5 paraissent

travailler sur des thèmes de recherche communs, tandis qu'ils s'opposent aux auteurs 1 et 2, qui eux-mêmes s'opposent à l'auteur 4. Nous obtenons ainsi une synthèse graphique sur ce que les chercheurs de notre laboratoire publient et avec qui.

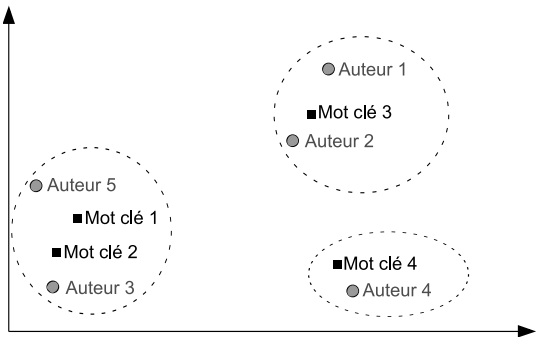

FIG. 3 – *Représentation des données sur les deux premiers axes factoriels*

Visualisation. Notre outil d'analyse en ligne retient les deux premiers axes factoriels comme nouvelles dimensions, dans le sens où les coordonnées des objets projetés agissent comme des descripteurs de leur position sur ces axes. Il est ainsi possible d'effectuer un drill-down de ces objets sur une carte factorielle.

4 Implémentation

Pour valider notre démarche, nous avons conçu une plate-forme logicielle implémentée sous la forme d'une application Web 2.0 Open Source. Elle s'installe sur un serveur Web Linux interprétant PHP5 et un serveur de base de données MySQL. Le logiciel R et son package FactoMiner doivent également être installés sur le serveur Web. D'un point de vue technique, l'interface graphique est gérée par le framework ExtJS qui a l'avantage de contenir un excellent support Ajax. D'un point de vue utilisateur, l'application se présente sous la forme d'un accordéon à déployer pas à pas. Ainsi, l'utilisateur procède aisément par étape dans le processus d'analyse : définition du contexte d'analyse, exécution de l'analyse factorielle et visualisation.

5 Discussion et conclusion

Nous avons développé dans ce papier une approche d'analyse en ligne d'objets complexes c'est à dire d'entités sémantiques. Les analyses que l'on peut effectuer doivent être basées sur le contenu des objets. Notre démarche a montré la faisabilité et l'intérêt d'utiliser l'analyse factorielle pour permettre la navigation et la visualisation en ligne des données complexes tout en se basant sur le contenu sémantique des données.

L'approche préconisée dans ce papier constitue seulement une première proposition. Il convient de poursuivre cette approche en la formalisant, en la dotant d'indicateurs sur la qualité de représentation, en donnant plus de sens à la nouvelle représentation, en l'appliquant

à d'autres types de données complexes (par exemple géographiques, dans le domaine médical, etc.). De plus, nous travaillons dès à présent sur plusieurs problèmes comme par exemple celui des cubes de données complexes ayant des dimensions qualitatives et quantitatives. Cet aspect revêt une importance particulière dans le cadre des techniques de fouille de données. Il convient de prendre en compte cette hétérogénéité des dimensions pour la visualisation des objets complexes. D'autre part, lorsqu'il existe une mesure autre que la mesure COUNT, nous réfléchissons pour pouvoir prendre en compte cette mesure dans la visualisation des objets complexes. Par exemple, si cette mesure est de type numérique ou quantitatif, on pourrait pondérer les faits par cette mesure dans l'analyse factorielle.

Dans le problème particulier de l'analyse des publications du laboratoire, à l'heure actuelle, ce sont les auteurs qui affectent, à partir d'un thésaurus, les mots-clés à leurs publications. Ce mode d'affectation peut être fastidieux et peut biaiser l'analyse. Nous envisageons d'utiliser des techniques de recherche d'informations pour détecter automatiquement les mots-clés de chaque publication à partir du titre ou du résumé. Ainsi un couplage de l'OLAP et de la recherche d'informations permettrait d'enrichir une nouvelle fois l'analyse en ligne d'objets complexes.

Références

Benzecri, J. (1982). *L'analyse des données (tome 2 : l'analyse des correspondances)*. Dunod, Paris, 4ème édition.

Chaudhuri, S. et U. Dayal (1997). An overview of data warehousing and OLAP technology. *In ACM-SIGMOD, Record 26(1)*.

Han, J. (1997). Olap mining : An integration of olap with data mining. In *Proceedings of the 7th IFIP Conference on Data Semantics*, Volume 39(11), Leysin, Switzerland.

Imielinski, T. et H. Mannila (1996). A database perspective on knowledge discovery. In *Communications of the ACM*, Volume 39(11), pp. 58–64.

Inmon, W. H. (1996). *Building the Data Warehouse*. John Wiley & Sons.

Kimball, R. et R. Merz (2000). *The data webhousing*. Eyrolles.

L. Lebart, A. M. et M. Piron (2004). *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 3ème édition.

Messaoud, R. B. (2006). *Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes*. Ph. D. thesis, Université Lumière Lyon 2, Lyon, France.

Summary

Datawarehouses and On-line Analysis Processing (OLAP) have aknowledge and efficient solutions in the decision help process. In this paper, we suggest to combine OLAP analysis and data mining in order to create a new operator of visualization of complex objects. This operator uses the correspondence analysis.