Représentation des documents textuels : Etude d'un domaine à travers des publications associées

J. Mothe[1,2], C. Chrisment[1], B. Dousset[1], S. Karouach[1]

1. CNRS – Institut de Recherche en Informatique de Toulouse – 118 Route de Narbonne 31062 Toulouse
mothe@irit.fr

2. Institut Universitaire de Formation des Maîtres de Midi-Pyrénées, 56 av. de l'URSS, 31078 Toulouse

Mots-clefs : information textuelle, exploration d'information, données géo-référencées, analyse multi-dimensionnelle, analyse de données

1 Introduction

Un système de recherche d'information (SRI) vise à permettre aux utilisateurs de retrouver tous les documents pertinents par rapport à leurs besoins en information, et seulement ceux-là. De nombreux systèmes sont ainsi utilisés pour permettre l'accès à de l'information électronique toujours disponible en plus grande quantité (moteurs de recherche sur le web, outils d'accès aux collections spécifiques). Quelque soit le système utilisé, un des problèmes majeurs rencontré est d'exprimer le besoin de façon à rendre la recherche efficace. Une solution pour résoudre ce problème est d'orienter l'utilisateur en structurant l'espace d'information. Le recours aux index est une première étape qui permet de représenter une collection via un ensemble de descripteurs. Cette représentation n'est pas toujours suffisante pour permettre l'expression optimale du besoin et pour accéder à de l'information élaborée. La représentation multi-facettes [1], que nous proposons, apporte une sémantique plus riche aux descripteurs. Cette représentation est mise à profit dans deux tâches complémentaires : l'accès à l'information 'brute', c'est à dire aux documents eux-mêmes et l'accès à de l'information élaborée extraite d'un ensemble de documents.

2 Représentation des documents textuels

Représentation multi-facettes des objets d'étude

Chaque document n'est pas seulement considéré comme un ensemble de termes, comme cela est le cas dans la majorité des moteurs d'indexation et de recherche [2,3], mais selon différents points de vues traduits au travers des facettes ou dimensions. Ainsi, un point clef de notre approche est que la recherche et l'exploration d'information s'effectuent dans un contexte sémantique, dépendant du domaine. Un contexte sémantique est représenté via un vocabulaire organisé selon des hiérarchies de concepts ou des ontologies, chacune représentant une facette des documents [4].

Représentation réduite d'un ensemble d'objets

La représentation de chaque objet considéré individuellement permet de fournir des index pour la recherche des documents pouvant répondre à un besoin d'information. Cette représentation n'est pas suffisamment synthétique pour répondre à des besoins en information élaborée qui doivent faire appel à des fonctions d'exploration de données. Les représentations réduites que nous proposons permettent de décrire globalement un ensemble de documents et sont compatibles avec les fonctions d'exploration définies dans la littérature [5]. Ces représentations se basent sur des tables de contingence, qui correspondent à une forme efficace de représentation d'information pour leur exploration.

3 Exploration / découverte de connaissances

Dans notre approche, *l'accès à l'information brute* est facilité par la possibilité qui est offerte aux utilisateurs de parcourir le contexte sémantique associé à la collection choisie et ainsi d'être guidés dans le choix des termes d'interrogation. De plus, des représentations globales d'un ensemble de documents leur permettent d'appréhender le contenu de la collection et sa répartition en fonction des thèmes ou concepts choisis. L'accès direct à des documents n'est pas toujours suffisant et des connaissances supplémentaires sur les corrélations entre les éléments représentatifs des documents peuvent être utiles en particulier lors de l'analyse d'un domaine. Nous proposons des cartes interactives permettant à l'utilisateur *de découvrir des connaissances* à partir d'un ensemble de documents ciblés. La construction de ces cartes s'appuie sur des méthodes d'analyse de données comme les méthodes de classification (supervisées ou non) et les méthodes d'analyse factorielle [6], mais également sur les mécanismes OLAP [7].

4 Cas des données géo-référencées

Dans le cas particulier de données géo-référencées, l'utilisation de cartes géographiques s'avère être un outil supplémentaire pour représenter de façon synthétique l'information découverte à partir d'un ensemble de documents. Ces cartes peuvent être construites à partir des fonctions d'analyse précédentes, pour, par exemple, mettre en évidence les pays ou régions qui possèdent les mêmes caractéristiques ou qui se comportent de façon identique par rapport aux éléments étudiés [9].

5 Exemple d'applications

Les méthodes que nous présentons ont été appliquées à l'analyse de différents domaines [8]. Dans le cadre de la veille scientifique et technologique, par exemple, notre approche permet de connaître les éléments importants d'un domaine (auteurs, techniques, organismes), leurs spécificités, leurs corrélations mais également leurs évolutions. Ce type d'analyse n'est possible qu'en se basant sur des sources de données qui sont forcément distribuées (web, bases de brevets, références bibliographiques). De plus, une analyse efficace nécessite la collaboration de différentes compétences : documentalistes, spécialistes de l'analyse de données et experts du domaine. Notre approche permet à ces différents acteurs de collaborer à distance. Cette collaboration se retrouve aussi au niveau des différents modules d'analyse qui coopèrent pour aboutir aux résultats finaux.

Références

[1] S.R. Ranganathan (1967). Prolegomena to library classification, New York Asia pub. house, 6-40.

[2] TREC (2002). Text Retrieval Conference, NIST, www://trec.nist.gov

[3] R. Baeza-Yates, B. Ribeiro-Neto (1999). Modern Information Retrieval, Addison-Wesley Ed., ISBN 0-201-39829-X.

[4] Mothe, J., Egret, D., Chrisment, C., Englmeier, K. H., Dkaki, T., & Lesteven, S. (2002, December). Knowledge discovery in bibliographic collections using concept hierarchies and visualization tools. Application to the astronomy domain. In Astronomical Telescopes and Instrumentation (pp. 246-253). International Society for Optics and Photonics.

[5] Fayyad, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, (1996) Advances in Knowledge Discovery and Data Mining, AAAI Press, ISBN 0-262-56097-6.

[6] J-P. Benzécri, (1992). Correspondence Analysis Handbook, Marcel Dekker Ed., New York.

[7] Mothe, J., Chrisment, C., Dousset, B., & Alaux, J. (2003). DocCube: Multi□dimensional visualisation and exploration of large document sets. Journal of the American Society for Information Science and Technology, 54(7), 650-659.

[8] B. Dousset, Tétralogie, http://atlas.irit.fr

[9] S. Karouach, B. Dousset (2001). Visualisation interactive pour la découverte de connaissances : GeoECD, Veille Stratégique, Scientifique et Technologique, Actes I , 301-311.