Visualisation globale de collections de documents sous forme d'hypercube

Le système DocCube

J. Mothe[1,2], C. Chrisment[1], J. Alaux[1]

(1) Institut de Recherche en Informatique de Toulouse, 118 Rte de Narbonne, 31062 Toulouse CEDEX 04

(2) Institut Universitaire de Formation des Maîtres, 56 av. de l'URSS, 31400 Toulouse

mothe@irit.fr

RÉSUMÉ. *L'objectif de cet article est de proposer une approche exploratoire pour des collections de documents basée sur des techniques visuelles. Cette approche permet de caractériser le sous-ensemble de documents répondant aux centres d'intérêts de l'utilisateur. Ceux-ci sont exprimés via des concepts hiérarchisés. La sélection des documents est réalisée via un ensemble de hiérarchies de concepts sur la collection qui représente la connaissance du domaine. L'originalité de l'approche réside dans l'utilisation du concept de cube de données présent dans les approches de type OLAP (On-Line Analytical Processing) avec des collections de documents. Chaque hiérarchie de concept est utilisée comme axe du cube et peut être explorée selon différents niveaux d'abstraction, reprenant ainsi les principes de forage des systèmes décisionnels. L'implantation de cette approche a été réalisée dans le système DocCube qui offre les fonctionnalités des hypercubes comme dans les systèmes OLAP et pour lequel nous en illustrons certaines.*

ABSTRACT. *This paper introduces a visual method in order to explore a document collection. In this approach, a domain is represented under the form of a set of concepts hierarchies that structures the domain knowledge. Users express their points of interest through these hierarchies. One key points of the approach is the use of a data cube representation (which can be found in OLAP systems). Each concept hierarchy is used as an axis of the cube and can be browsed at different levels of abstraction. This approach has been implemented in DocCube. This system offers the functionnalities of data cubes in OLAP systems.*

MOTS-CLÉS : *exploration d'information, système OLAP, analyse multi-dimensionnelle, recherche d'information*

KEYWORDS: *information mining, OLAP systems, multi-dimensionnal analysis, information retrieval*

1. Introduction

L'exploration de collections de documents est une fonctionnalité classique en recherche d'information. L'introduction d'un facteur d'échelle dans le dimensionnement des collections à explorer, notamment dans le contexte du Web à conduit au développement de nouvelles approches. Si l'on considère le contexte du Web, l'exploration est possible via les moteurs de recherche. Deux types de moteurs de recherche sont disponibles sur le Web. Les moteurs de recherche par mots clés permettent à l'utilisateur d'exprimer son besoin d'information via une requête en langage libre qui peut être combinée, dans le cas des recherches *avancées,* avec les opérateurs dits booléens. Google, Altavista fonctionnent sur ce principe. Ces moteurs se basent sur un index qui mémorise la relation entre les termes d'indexation et les références (URL) des pages matérialisant les documents. Les termes d'indexation sont directement issus des contenus des documents. La difficulté majeure de ce principe de recherche réside dans le fait que l'utilisateur ne possède aucune information sur la relation entre le langage d'indexation et son langage d'interrogation. Il n'a donc aucune explication sur la réponse du système à sa requête. Au contraire les moteurs de type répertoire (comme Yahoo) permettent d'assister l'utilisateur dans l'expression de son besoin d'information. L'utilisateur navigue dans l'espace d'information défini par une hiérarchie thématique. La hiérarchie est créée manuellement et l'association des documents aux nœuds de la hiérarchie est, en fonction des moteurs soit manuelle, soit automatique. Ce principe de recherche permet à l'utilisateur de formuler son besoin en prenant en compte le langage d'indexation ou la classification pré-établie des documents. Cependant, même dans ce type de système, l'aide à la navigation reste limitée. En effet, même si à chaque étape l'utilisateur limite l'espace à consulter, il n'a pas d'information sur la taille de cet espace. Pourtant, les professionnels de l'interrogation considèrent cette information comme nécessaire pour reformuler l'expression du besoin, en particulier pour décider si la requête doit être plus large (c'est à dire basée sur des termes plus génériques) ou plus spécifique. Par ailleurs, l'utilisation de hiérarchies de termes pour aider l'utilisateur devrait se (re)développer dans le contexte du Web (Shiri, 00). En effet, l'utilisation de hiérarchies de concepts ou d'ontologies a par exemple été retenue dans certains projets autour du Web sémantique afin de structurer le Web (Berners-Lee, 01).

Notre approche vise à aider à la structuration de collections de documents recouvrant différents domaines. Une des originalités de notre approche est que la connaissance d'un domaine est décrite selon différentes hiérarchies de concepts (HC) qui structurent l'espace d'information. L'utilisateur navigue dans l'espace d'information au travers de ces HC. L'intérêt pour l'utilisateur réside dans le fait qu'il ne perd jamais le contexte sémantique lors de sa recherche et ne peut pas rencontrer par hasard des données qui appartiennent à un autre contexte (Englmeier, 01). L'instanciation de la HC, c'est à dire l'association des documents aux différents concepts, est automatique. Elle est basée sur une indexation en langage contrôlé. Il s'agit là d'un élément important de notre approche qui permet d'homogénéiser les

langages d'interrogation et d'indexation. En effet, les HC correspondent au langage d'interrogation puisque l'utilisateur navigue selon leur structure pour accéder aux documents mais constituent également le langage contrôlé d'indexation. La distorsion entre ces deux langages, qui est à l'origine à la fois des difficultés de formulation de requêtes pour l'utilisateur et de son incompréhension face à des résultats de recherche, disparaît. Un autre élément clé de notre approche réside dans le fait que nous ne permettons pas simplement un accès direct à l'information pertinente pour l'utilisateur par rapport à son besoin, mais nous lui fournissons également une vue globale – par niveau d'abstraction- de l'espace d'information qu'il consulte, afin de faciliter ses choix de consultation. Cette vue globale repose sur les principes d'analyse multidimensionnelle. Elle permet à l'utilisateur d'appréhender une collection de documents en ayant déjà une information générale sur son contenu avant d'accéder le contenu des documents.

Cet article est organisé comme suit. La section 2 rappelle l'architecture des systèmes de découverte de connaissances à partir des bases de données. Il s'agit là d'un point de départ pour appréhender la découverte de connaissances à partir des documents décrite dans la section 3. La section 4 présente notre modèle de connaissances intégré au système DocCube. Les deux sections suivantes illustrent le fonctionnement de DocCube: la section 5 précise le principe de rattachement des documents aux hiérarchies alors que la section 6 décrit l'interface de DocCube.

2. Découverte de connaissances à partir de bases de données

Les principes de découverte de connaissances à partir des bases de données ont été introduits afin d'aider les décideurs dans l'analyse des informations issues des sources transactionnelles. Des techniques automatiques sont proposées pour inférer des nouvelles connaissances, potentiellement utiles, à partir de gros volumes de données. Ces connaissances correspondent à des modèles ou des relations à priori inconnus mais qui existent dans les données. Un processus de découverte de connaissances comprend trois étapes principales (Chaudhuri, 97), (Fayyad, 96):

La sélection de données et leur pré-traitement : L'objectif de cette étape est d'obtenir des données dans un format facilement et rapidement exploitable pour la découverte de connaissances. Elle consiste en différentes tâches comme la collecte, l'homogénéisation, le nettoyage et la réduction des données sources. Les données résultant de ces différentes tâches sont typiquement mémorisées dans une structure appelée entrepôt de données. Ainsi, un entrepôt de données correspond à une collection optimisée pour supporter les opérations d'aide à la prise de décision et les traitements OLAP (On-Line Analytical Processing) (Widom, 95).

L'analyse et l'exploration de données : Cette étape a pour objectif principal d'analyser les données issues de l'étape précédente afin d'en induire des relations inconnues mais pourtant utiles (Fayyad, 96). Deux types d'analyse de données peuvent être distingués :

- l'analyse multidimensionnelle : les mesures (données) analysées sont vues à différents niveaux d'abstraction, à la demande de l'utilisateur. Les outils

d'analyse multidimensionnelle ou systèmes OLAP proposent généralement une interface sous forme d'un cube ayant pour axes les dimensions. Les dimensions choisies dépendent des données manipulées. Ces dimensions sont généralement hiérarchiques et des opérateurs permettent facilement de changer le niveau de ces dimensions pour observer les mesures à différents niveaux d'abstraction. Les mesures représentées en fonction de différentes dimensions qui peuvent être "forées", soit vers le haut, pour accroître le niveau d'agrégation et voir les données de façon globale, soit vers le bas, pour réduire le niveau d'agrégation et voir les données détaillées.

- l'exploration d'information : elle fait référence à la découverte automatique de règles et au domaine de l'apprentissage automatique. Différentes fonctions d'exploration de données ont été définies dans la littérature (Agrawal, 93):

  Classification : l'objectif est de partitionner les données soit en les rangeant dans des classes pré-définies, soit en les regroupant en fonction de leur ressemblance. Les méthodes de classification se basent sur une mesure de similarité et sur la ressemblance des objets à classer.

  Dépendances : l'objectif est de trouver des dépendances entre variables, des relations entre les attributs, des corrélations, des dépendances temporelles, des séquences et des régressions.

Interprétation : La découverte de connaissances ne peut être complète que si les informations inférées peuvent être exploitées par un agent (humain ou automate). Ainsi les informations induites à l'étape précédente doivent être présentées dans un format directement exploitable. Dans les systèmes OLAP, les principes visuels associés aux interfaces graphiques jouent le relais entre les données et l'utilisateur. Dans les systèmes d'exploration en revanche, les résultats peuvent être représentés sous forme de règles.

3. La découverte de connaissances à partir de textes

La nécessité de disposer de systèmes de plus en plus sophistiqués pour manipuler d'immenses masses d'information et pour en extraire les seuls éléments utiles en fonction des besoins de l'utilisateur est totalement d'actualité pour les sources documentaires. La forme des connaissances extraites d'une grande collection de documents peut prendre différentes formes. La classification de documents (Jain, 99) correspond à une première classe d'applications. En effet les techniques de classification de documents offrent un moyen de résumer ou de visualiser de façon globale un ensemble de documents. Cette classification est généralement associée à une représentation graphique, par exemple un arbre, que l'utilisateur peut balayer pour connaître rapidement les différents aspects correspondant à son besoin d'information (Hearst, 97). Dans la majorité des cas, les documents sont considérés comme des ensembles de mots et le regroupement des documents est basé sur cette représentation. Dans cette approche, s'il est possible de caractériser les mots à l'origine du regroupement, il est beaucoup plus difficile de matérialiser la variété d'information contenue dans les documents. En effet, les mots sont considérés

indépendamment de leur rôle sémantique. L'utilisation des méta-informations permet de considérer une partie des variétés sémantiques contenues dans les textes afin de regrouper les documents selon différentes vues (Mothe, 01). Une autre approche basée sur l'analyse des liens référentiels entre pages dans le contexte du Web a permis de définir l'*importance* des pages, les pages qui font autorité (Kleinberg, 99). Dans (Guillaume, 99), les liens entre documents sont définis en fonction des méta informations qu'ils contiennent. Les documents sont ensuite classés par rapport à ces liens et représentés sous forme d'une carte auto-organisatrice de Kohonen.

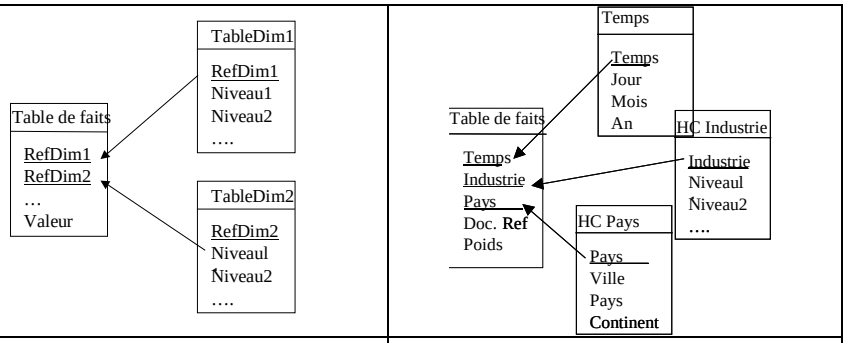
4. DocCube: un modèle pour la recherche et l'analyse de documents

Lorsqu'il utilise un système de recherche d'information traditionnel, l'utilisateur interroge la collection en formulant une requête en langage "libre". Le système lui retourne alors la liste des documents qui sont susceptibles de répondre à son besoin, compte tenu des termes communs à la requête et au contenu du document. L'utilisateur doit alors décider quels documents il souhaite réellement consulter ou, après avoir naviguer dans la liste de documents, éventuellement décider de reformuler l'expression de son besoin. DocCube vise à aider l'utilisateur dans ces deux tâches difficiles que sont l'expression du besoin et la décision de consultation ou la navigation dans l'espace d'information. Un des composants de base de DocCube et une de ses originalités correspond à la notion de hiérarchies de concepts qui structurent l'espace d'information. Ces hiérarchies correspondent en fait à différents aspects qui composent chaque domaine. Par exemple, le domaine de la veille scientifique et technique peut être structuré selon l'origine de la technologie (auteur, laboratoire, ville, pays, …), les techniques utilisées, la date, l'objectif de la technologie (Mothe, 01). Le domaine économique peut être structuré autour de hiérarchies comme le temps, le géo-référencement, les indicateurs économiques, les industries. Les documents sont rattachés aux nœuds de ces hiérarchies pour permettre leur accès, comme cela est le cas dans les systèmes de classification. Dans DocCube, nous exploitons également cette structuration des informations selon des hiérarchies pour proposer à l'utilisateur des visualisations globales d'information qui l'aident dans sa recherche et dans l'exploration de la masse d'information dont il dispose. Ces visualisations globales reposent sur une modélisation multidimensionnelle. C'est à dire que l'information est représentée et organisée selon différentes dimensions et que des faits peuvent être analysés de façon interactive. Dans DocCube, les **dimensions** décrivent la connaissance associée au(x) domaine(s) étudié(s). Ces dimensions sont hiérarchiques ; la hiérarchie représente un lien "est un" ou "est plus spécifique que". Les dimensions dépendent du domaine. Par exemple, dans le projet IRAIA (IRAIA) qui concerne la mise à disposition de données économiques pour des analystes économiques, les dimensions qui ont été choisies correspondent aux types d'industrie, aux indicateurs économiques, aux régions et aux dates. Ce sont ces dimensions qui sont également exploitées par DocCube. Chacune de ces dimensions est hiérarchique, mais la profondeur des hiérarchies varie. La mesure (**fait**) que l'utilisateur peut analyser correspond au nombre de documents qui sont associés aux valeurs de la hiérarchie. Comme dans

les systèmes OLAP, la valeur de cette mesure est recalculée en fonction du niveau d'agrégation (i.e. de généralité) auquel l'utilisateur s'intéresse. De plus, contrairement aux systèmes OLAP, le lien avec l'information brute, c'est à dire avec le document ou la page Web, est maintenu. Cela peut permettre un retour aux contenus lorsque leurs représentations synthétiques ne sont pas suffisantes.

4.3. Le modèle de l'entrepôt

Le modèle de l'entrepôt sur lequel repose DocCube contient une table par dimension c'est à dire par hiérarchie de concepts du domaine. Il respecte le modèle dit en étoile dans lequel une dimension est décrite dans une seule table. La table correspondant aux faits qui contient les mesures à analyser, tout en préservant le lien avec les dimensions, est également présente dans le modèle de DocCube. La figure 1 représente le schéma général d'un entrepôt de données ainsi que l'extrait correspondant au domaine économique.



Modèle général de l'entrepôt de données	Modèle de DocCube (domaine économique)

Figure 1 : *Schéma en étoile de l'entrepôt de données*

La table de faits contient, en plus du lien avec les hiérarchies de dimension, une référence vers les contenus de document (*DocRef*) correspondant à l'URL du document et la force du lien entre le nœud représenté et le document (*Poids*).

5. Génération de l'entrepôt de données

5.1. Association des textes aux hiérarchies de concepts

Une hiérarchie de concept (HC) est une arborescence composée de concepts ou entrées, chaque entrée correspondant à un ensemble de termes. L'aide à la construction automatique de hiérarchies de concepts n'est pas abordée dans cet article. Des travaux existent dans ce domaine (Reynaud, 98). L'association automatique des textes à des HC en revanche correspond à une des préoccupations abordées dans ce papier. Cette association peut être vue comme la classification de documents suivant des domaines de connaissances. Elle peut également être vue

comme l'indexation automatique de documents à partir d'un vocabulaire contrôlé issu des HC, l'ensemble des entrées auxquelles un texte est rattaché correspond alors à l'annotation du texte. Cette association automatique permet de créer des contextes de recherche non ambigus et clairement identifiés.

Chaque texte peut ainsi être associé à différentes entrées d'une même hiérarchie ou de hiérarchies différentes. L'association d'un texte à une entrée d'une hiérarchie repose sur la méthode *Vector Voting* (Pauer,00). Cette méthode se base sur l'extraction automatique des termes de chacune des entrées dans le contenu du texte. L'importance de l'association du texte avec une entrée donnée est calculée par une méthode de vote, qui peut être rapprochée de la méthode HVV (Hyperlink Vector Voting) utilisée pour calculer la pertinence d'une page en fonction des sites qui y référent (Li, 98). Dans notre contexte, plus l'entrée ou une partie de l'entrée est présente dans le texte, plus le lien entre le texte et cette entrée sera fort. Cette technique s'applique à tout type de documents, issus du Web ou non. Dans le cas de textes issus de Web, cette classification permet d'avoir une structuration d'un ensemble de pages.

L'association d'un document à des entrées s'effectue suivant différentes étapes :

- Extraction automatique des termes représentatifs de chaque entrée d'une hiérarchie de concepts et de leur importance dans l'entrée.
- Extraction automatique des termes représentatifs du document et de leur importance au sein du document. Le processus d'extraction est basé sur un ensemble de règles qui utilisent des balises des documents et des expressions régulières. Une fois les balises détectées, des fonctions sémantiques et syntaxiques complètent le processus d'extraction afin de gérer les synonymes et l'élimination de termes non intéressants.
- Pour chaque entrée de la hiérarchie, calcul du score selon une méthode de vote. Le calcul de score peut être basé sur différentes fonctions de calcul qui peuvent faire intervenir des mesures comme l'importance d'un terme dans le document, l'importance d'un terme dans la hiérarchie, la taille du document, la taille de la hiérarchie, le nombre de termes d'une entrée présents dans le document.
- Classement des entrées de la hiérarchie dans l'ordre des scores obtenus, puis sélection de l'ensemble des entrées à associer au document suivant une stratégie définie (par exemple, les entrées ayant obtenu un score supérieur à un seuil donné, ou les *n* premières entrées ayant les meilleurs scores).
- Modélisation des associations entre documents et entrées de hiérarchies sous forme de code XML ajouté au contenu du document. Ce code bien que n'étant pas interprété par les navigateurs actuels peut être exploité par des agents intelligents ou par une application telle que DocCube.

La fonction de vote utilisée pour associer un texte et les entrées d'une hiérarchie est la suivante (Auge,01):

$$Poids(E_H,D,x)=\sum \frac{F(T,D)}{S(D)}.\frac{S(H)}{F(T,H)}.10^{\frac{NT(E,D)}{NT(E)}}$$ si au moins x% des termes de l'entrée apparaissent dans le document,

$$= 0$$ sinon

Avec D, le document, E_H l'entrée E de la hiérarchie de concepts H, $\frac{F(T,D)}{S(D)}$ mesure l'importance du terme T dans le document D, F(T,D) correspond à l'occurrence du terme T dans le document D et S(D) correspond à la taille de D, $\frac{S(H)}{F(T,H)}$ mesure l'importance du terme T dans la hiérarchie H, F(T,H) correspond à l'occurrence du terme T dans H et S(H) correspond à la taille de H ; $\frac{NT(E,D)}{NT(E)}$ mesure le taux de présence de l'entrée dans le texte, NT(E) correspond au nombre de termes de l'entrée E et NT(E,D) correspond au nombre de termes de l'entrée E qui apparaissent dans D.

Cette fonction est issue de tests qui sont décrits dans (Auge, 01). La fonction basée sur x=100 a donné les meilleurs résultats en terme de taux de rappel / précision. Le facteur $10^{NT(E,D)/NT(E)}$ est utilisé pour donner plus d'importance aux entrées pour lesquelles plusieurs termes sont extraits des documents.

5.2. Association des textes aux nœuds dans la table de faits

La table de faits mémorise la référence aux documents ainsi que le poids de l'association pour un n-uplet de valeurs issues de chacune des hiérarchies (cf. fig. 1). Ce poids est défini par la moyenne des poids associant chacune des entrées et le texte donné :

$$Poids(N, D) = Moyenne_H(Poids(E_H, D, X))$$

6. L'interface de DocCube

L'interface de DocCube est basée sur une représentation en cube de données pour lequel les axes correspondent aux hiérarchies ou dimensions du domaine. Ces hiérarchies peuvent être balayées afin d'analyser les informations à différents niveaux de détail. Les opérateurs de forage sont disponibles à cette fin. La vue 3D peut être réduite à une vue 2D par l'opération de coupe. L'analyse multi-dimensionnelle du contenu de la collection de documents peut être poursuivie par un accès direct aux documents.

6.1. Choix du niveau d'agrégation

La navigation de l'utilisateur débute par le choix de l'espace d'information ou du domaine. Il a alors accès à l'ensemble des hiérarchies ou dimensions de ce domaine. Pour éviter de perdre l'utilisateur en lui fournissant trop d'information, seul le premier niveau de chaque hiérarchie est présenté. Il s'agit là d'un point de départ commun à la plupart des systèmes basés sur des hiérarchies. A partir des racines des hiérarchies, l'utilisateur peut naviguer jusqu'à obtenir le niveau de détail voulu (fig. 2). Il peut également sélectionner les éléments auxquels il s'intéresse plus spécifiquement.

Figure 2: *Sélection du niveau dans la hiérarchie et des éléments pertinents*

6.2. Représentation 3D

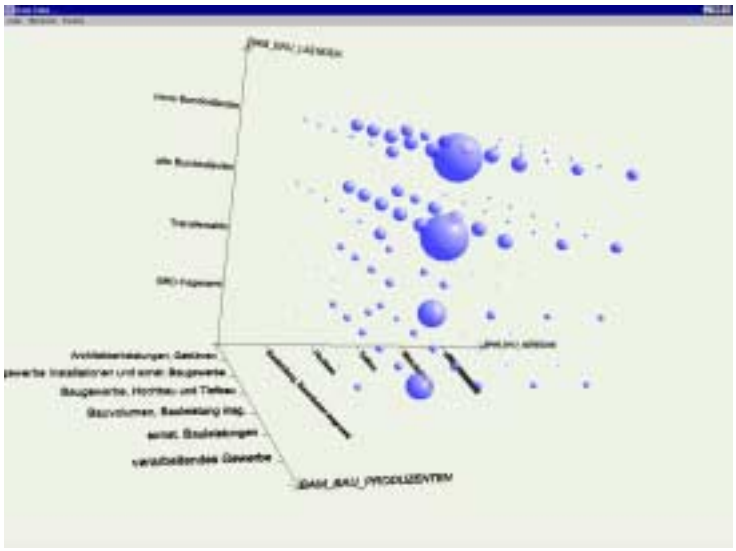


Figure 3*: Représentation 3D de DocCube*

DocCube offre des visualisations globales d'information concernant la collection de documents couvrant le domaine choisi par l'utilisateur. Les axes du cube correspondent aux dimensions, c'est à dire aux hiérarchies de concepts. L'intersection des axes correspond au nombre de documents rattachés à cet axe. Cette information est représentée sous la forme d'une sphère dont la taille est proportionnelle au nombre de documents. Ce nombre est recalculé dynamiquement lorsque l'utilisateur change le niveau d'agrégation.

6.3. Coupe

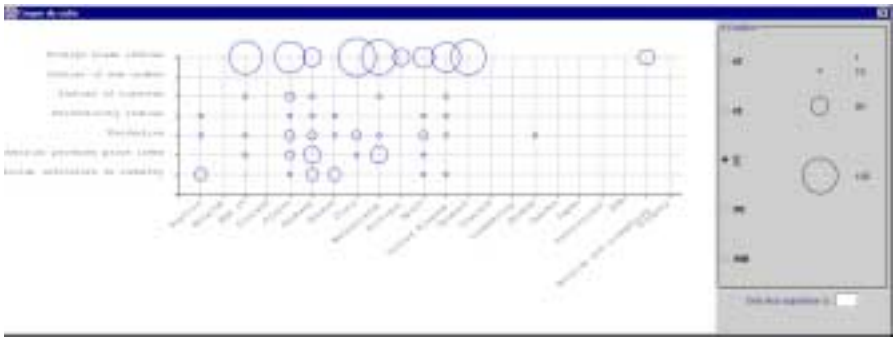


Figure 4: *La coupe dans DocCube*

La fonction de coupe est utilisée lorsque l'utilisateur souhaite fixer la valeur d'une des dimensions et ainsi obtenir une vue 2D des mesures. Dans DocCube, la coupe résulte d'une interaction via le pointeur de la souris. Le résultat d'une coupe est représentée comme indiqué à la figure 4. Le nombre de documents répondant aux critères choisis est affiché via des cercles dont l'échelle peut être modifiée de façon interactive. Cela peut permettre de ne s'intéresser qu'aux éléments prépondérants. Il est également possible d'ignorer certains éléments dans la visualisation (ceux qui correspondent à un nombre trop faible ou au contraire trop important). Cette fonctionnalité existe également dans la visualisation 3D.

6.4. Accès aux documents

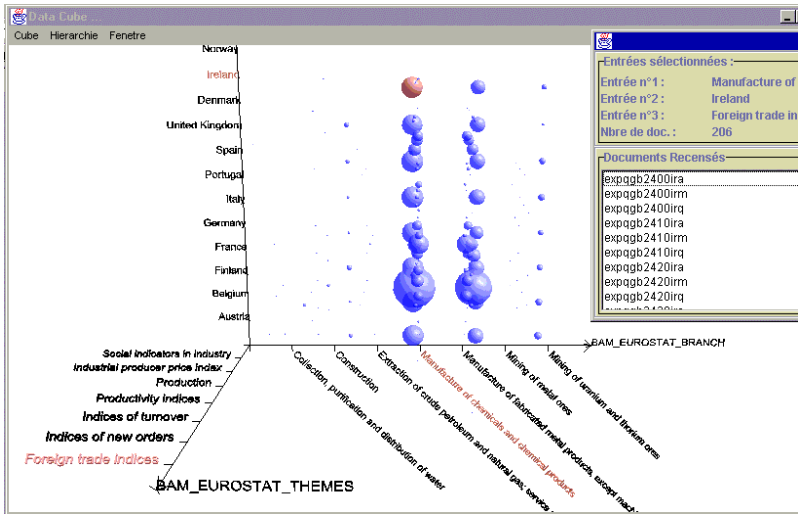


Figure 5*: Accès aux documents via DocCube*

Un besoin d'information peut être directement formulé en sélectionnant un ou plusieurs nœuds dans le cube de la représentation 3D. Après cette opération, la liste

des références aux documents correspondant est affichée. Il s'agit d'une liste ordonnée, par pertinence supposée. Les termes correspondant à la requête sont mis en évidence dans le cube ; ils sont également rappelés dans la fenêtre résultat. Le score obtenu par les documents dépend du poids associé à chaque document et au niveau de rattachement du document dans les différentes hiérarchies.

7. Conclusion

Dans cet article, nous avons présenté une nouvelle approche pour l'exploration de large collections de documents, en particulier issus du Web. Notre approche combine les fonctionnalités des moteurs de recherche et des systèmes OLAP. Comme dans les moteurs de recherche traditionnels, l'utilisateur peut accéder aux contenus des documents à partir d'un besoin d'information. Une des composantes originales de notre approche est que cet accès est guidé par la structure de l'espace d'information que l'utilisateur accède. En effet, chaque domaine est décrit par un ensemble de hiérarchies de concepts que l'utilisateur peut balayer afin de compléter sa connaissance du domaine ou pour spécifier ses centres d'intérêt. Un élément clé de notre approche est la visualisation globale de la collection à différents niveaux d'abstraction ou d'agrégation. Ainsi, l'utilisateur peut avoir une connaissance globale sur le contenu de la collection avant même de commencer à l'interroger. Cette représentation globale d'information peut être utile pour les utilisateurs de différentes façons. Nous donnons des exemples d'usage qui correspondent à des objectifs différents.

Analyse bibliométrique : DocCube offre un moyen efficace de réaliser des études bibliométriques. L'objectif d'une telle analyse est de découvrir les liens ou les relations qui existent dans les informations analysées. Une des applications concerne la veille technologique et scientifique (Karouach, 01). Dans le cas de DocCube, la visualisation globale peut permettre de visualiser des vues de type histogramme à trois dimensions pour découvrir par exemple quels sont les auteurs ou les organisations principaux d'un domaine répartis par rapport aux thématiques du domaine et comment ces contributions au domaine évoluent dans le temps.

Amélioration de requêtes **:** DocCube offre à l'utilisateur un moyen de naviguer dans l'espace d'information qu'il choisit. En connaissant mieux le vocabulaire du domaine, il peut mieux spécifier son besoin. De plus, la visualisation globale d'information lui donne une indication sur le niveau de détail pertinent pour l'accès à l'information. Si trop de documents correspondent à un nœud, l'utilisateur à intérêt à spécialiser sa requête, c'est à dire à la préciser en descendant d'un niveau.

Exploration d'un ensemble de documents retrouvés suite à une requête sur le Web : DocCube peut permettre de structuré un ensemble de documents restitués suite à une requête sur un moteur de recherche. Dans ce cas, les hiérarchies de concepts permettent de structurer l'espace des pages retrouvées. La navigation dans les hiérarchies permet de savoir comment les pages sont dispersées par rapport au domaine, c'est à dire quels aspects sont traités dans les documents retrouvés.

7. Bibliographie

R. Agrawal, T. Imielinski, A. Swami, Database Mining: A Performance Perspective, IEEE transactions on knowledge and data engineering, pp 914-925, Vol.5, N.6, 1993.

J. Augé, K. Englmeier, G. Hubert, J. Mothe, Classification automatique de textes basée sur des hiérarchies de concepts, Veille Stratégique Scientifique & Technologique, 2001.

T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American, 2001, http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html.

S. Chaudhuri, U. Dayal, An overview of data warehousing and OLAP technology, ACM SIGMOD Record, Vol. 26, N.1, pp 65-74, 1997.

F. Crimmins, T. Dkaki, J. Mothe, A. F. Smeaton, TétraFusion: Information Discovery on the InternetIEEE Intelligent Systems & their applications, Vol 14, N 4, pp 55-62, IEEE Computer Society, 1999.

K.Englmeier, J. Mothe, B. Pauer, Users bootstrap seraching the Web through interactive agents supporting best pratice sharing, 9th international Conference Human-Computer Interface, pp 923-927, 2001.

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

D. Guillaume, F. Murtagh, An Application of XML and XLink Using a Graph-Partitioning Method and a Density Map for IR and KD, ASP Conf. Series, V 172, *ADASS VIII*, 1999.

IRAIA, Projet IST, http://iraia.diw.de

A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Computing Surveys, Vol.31, N.3, pp 264-323, 1999.

S. Karouach, B. Dousset, Visualisation interactive pour la découverte de connaissances: GeoECD, Veille Stratégique Scientifique & Technologique, 2001.

J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, Vol.46, N.2, pp 212-235, 1999.

Y. Li, Toward a qualitative search engine, IEEE Internet Computing, pp 24-29, V2(4), 1998.

J. Mothe, Recherche et exploration d'informations – Découvertes de connaissances pour l'accès à l'information, Habilitation à diriger des recherches, Univ. P. Sabatier, 2000.

J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret, Information mining: use of document dimensions in order to analyse a document set, pp 66-77, European Colloquium on Information Retrieval Research, 2001.

B. Pauer, P. Holger, Statfinder, Document Package Statfinder, Vers. 1.8, mai 2000.

C. Reynaud, N. Aussenac-Gilles, F. Tort, A Support to Domain Knowledge Modelling - A case study - Information Modelling and Knowledge Bases IX - H. Kangassalo et P.J. Charrel (eds).- Vol. 45 - Frontiers in AI - Amsterdam, IOS Press - - p. 35-50,1998.

A.A. Shiri, C. Revie, Thesauri on the Web: current developments and trends, Online Information Review, Vol. 24, N. 4, pp 273-279, 2000.

J. Widom, Research problems in data warehousing, International Conference on Information and Knowledge Management, 1995.