ACCES A UNE INFORMATION CIBLEE, SEMANTIQUE ET STRUCTUREE

Joël MARCO, Josiane MOTHE, Maryse SALLES
joel.marco@irit.fr, josiane.mothe@irit.fr, Maryse.Salles@univ-tlse1.fr

Université de Toulouse, Toulouse
IRIT, CNRS UMR5505
Toulouse

Mots clefs :
Recherche d'information sémantique, accès ciblé à l'information, interface, évaluation.
Keywords:
Semantic information retrieval, targeted information access, interface, evaluation.

Résumé

Dans cet article, nous proposons une approche qui considère l'utilisateur à travers les aides qui peuvent lui être apportées afin de cibler l'information potentiellement intéressante pour lui. Cette approche repose sur des espaces de recherche qui restreignent l'espace de consultation à un domaine d'intérêt. Cet espace est constitué d'un ensemble de documents et d'une ontologie de domaine modélisant la terminologie de ce domaine. Le ciblage de l'information recherchée est alors considéré dans les différents composants du processus de recherche implanté dans un moteur de recherche : l'indexation et la formulation de la requête, l'unité documentaire considérée et la présentation des résultats. Un panel d'utilisateurs a utilisé cette interface dans des tâches de recherche d'information et en a indiqué les avantages et onconvénients.

1. Introduction

Les moteurs de recherche fournissent aux utilisateurs un ensemble de documents supposé répondre à leurs besoins en information. Le travail présenté ici a pour but d'améliorer l'accès aux informations contenues dans ces documents.

Les moteurs du web tels que Google © ou Yahoo! © permettent de retrouver des informations sur pratiquement n'importe quel sujet à partir d'une requête formulée via quelques mots. Ils restituent une liste de références, ordonnée par pertinence supposée, avec un texte résumé du contenu (« snippets ») prévu pour aider l'utilisateur dans son choix d'accès à l'information. Ce type de recherche et de résultats est peu adapté à un accès à des éléments précis dans les documents. Par exemple si la réponse à la requête se trouve dans un paragraphe issu d'un document long, l'utilisateur n'est pas aidé pour y accéder.

Un autre inconvénient des moteurs généralistes est le risque d'ambigüité des termes. Le choix par le moteur de restituer un document est basé sur la similarité entre les chaines de caractères issus des documents et celles issues de la requête. L'utilisateur n'est pas aidé dans la formulation de son besoin sous forme de requête ni dans la compréhension de la raison de la restitution des documents par le moteur. D'autre part, le fait de ne pas prendre en compte la sémantique des termes dans la requête ou dans les documents amène du bruit dans les réponses du système qui rend plus difficile la sélection par l'utilisateur des documents réellement potentiellement utiles.

Dans cet article, nous proposons une approche qui considère d'avantage l'utilisateur à travers les aides qui peuvent lui être apportées afin de cibler l'information potentiellement intéressante pour lui. Cette approche repose sur des espaces de recherche qui restreignent l'espace de consultation à un domaine d'intérêt. Cet espace est constitué d'un ensemble de documents et d'une ontologie de domaine modélisant la terminologie de ce domaine. Le ciblage de l'information recherchée est alors considéré dans les différents composants du processus de recherche implanté dans un moteur de recherche : l'indexation et la formulation de la requête, l'unité documentaire considérée et la présentation des résultats.

L'*indexation* est un des composants fondamentaux du processus de recherche. Il s'agit d'associer des termes aux contenus des documents. Lors de la recherche les termes de la requête seront mis en correspondance avec les termes indexant les documents. Lorsque l'indexation est automatique, les termes sont généralement issus des contenus des documents eux-mêmes (Salton, 1971), (Baeza et Ribeiro, 1999), sans que leur sémantique ne soit prise en compte. Alternativement, les termes peuvent être issus d'une terminologie ou d'une ontologie (Aussenac et Mothe, 2004), (Khelif et Dieng-Kuntz, 2004). L'utilisation de ressources terminologiques dans l'indexation vise à se rapprocher des principes de l'indexation manuelle par son traitement sémantique du texte. C'est ce second type d'approche que nous avons choisi de suivre. En effet, la prise en compte de la sémantique des termes aide le ciblage de l'information, en désambigüisant les termes utilisés par la requête et lors de l'indexation. L'indexation des contenus peut être complétée par l'annotation via des métadonnées pour la recherche d'information (Kiryakov *et al.*, 2004). Le langage d'indexation et d'annotation est directement en lien avec le langage d'interrogation qui pourra être en texte libre (Gandhe *et al.*, 2006), (Hallet *et al.*, 2006) ou exprimé au travers une interface graphique (Russel *et al.*, 2007).

L'*unité documentaire* correspond à l'unité qui est indexée et restituée à l'utilisateur lors d'une interrogation. L'unité documentaire généralement retenue est l'unité imposée par l'auteur du document (par exemple la page web). Alternativement, la structure du document peut être utilisée pour changer la granularité

d'accès (Corral et Mothe, 1995). Cette approche est par exemple utilisée dans la recherche XML comme dans INEX[1]. C'est également cette approche que nous avons retenue dans la mesure où elle permet un accès plus ciblé à l'information par la définition de granules documentaires de taille limitée.

Enfin, une dernière composante correspond à la *présentation des résultats*, c'est-à-dire la manière dont ceux-ci vont être restitués à l'utilisateur. Les moteurs de recherche restituent généralement des listes ordonnées de documents mais ne proposent que peu d'aide à l'accès à l'information au sein d'un document. Certains travaux de recherche ont été menés pour améliorer cet accès, que ce soit par la représentation textuelle avec la mise en valeur des termes de la requête (Google cached) ou par la représentation 2D ou 3D (Bonnel *et al.*, 2005). Nous apportons également des solutions sur cet aspect dans cet article.

La suite de l'article est organisée comme suit. La section 2 présente des travaux de la littérature sur la recherche d'information sémantique, les granules documentaires et la présentation des résultats. Dans la section 3 nous présentons notre approche ainsi que les fonctionnalités correspondantes dans le prototype que nous avons développé. Cette présentation est centrée sur l'aide apportée à l'utilisateur et au ciblage de l'information. Dans la section 4 nous présentons un premier cadre expérimental d'évaluation. La dernière section conclut l'article et présente les perspectives de ce travail.

2. Travaux reliés

2.1. RECHERCHE SEMANTIQUE

L'utilisation de concepts hiérarchisés est une alternative à celle des mots directement issus des contenus des documents pour l'indexation et la recherche. Il s'agit d'un premier pas vers l'indexation et la recherche sémantique dans la mesure où l'ambiguïté des termes peut être levée via les concepts et leur organisation hiérarchique (Aussenac et Mothe, 2004). Lors de la recherche, la présentation d'une hiérarchie de concepts utilisée pour l'indexation peut aider l'utilisateur à formuler sa requête. Cette hiérarchie peut en outre permettre de mettre en œuvre des mécanismes de reformulation de requêtes par exemple par ajout des termes définis proches des termes de la requête par la hiérarchie de concepts. Ainsi, IRAIA (Mothe *et al.*, 1993) permet une indexation et un accès à des informations via des hiérarchies de concepts, qui chacune représente une facette des documents. Les documents gérés sont dans le domaine économique et les facettes correspondent à la géolocalisation, la date, les industries concernées et les indices économiques. Mukherjea (1995) présente également un système dans lequel les documents sont présentés via des hiérarchies de concepts. Cat-a-cone (Hearst et Karadi, 1997) utilise le même type de présentation : les documents sont associés à des catégories hiérarchiques dans lesquelles l'utilisateur peut naviguer pour accéder aux documents.

Les ontologies peuvent être vues comme un autre moyen de structurer le langage d'indexation. Dans (Ka)[2] (Benjamins, 1999), les pages web sont annotées manuellement par des concepts issus d'une ontologie. Lors de la recherche, la requête initiale est reformulée par ajout des concepts liés aux termes de la requête. Meat Annot (Khelif et Dieng-Kuntz, 2004) propose une interface d'assistance à l'indexation basée sur l'utilisation d'ontologies. Elle permet de valider des termes fournis par des outils syntaxiques puis générer des annotations RDF. OntoExplo (Hernandez *et al.*, 2007) propose de représenter les documents à la fois à travers leur contenu via une ontologie de thème qui regroupe les concepts du domaine étudié et à travers des métadonnées correspondant

[1] INitiative for the Evaluation of XML Retrieval, http://inex.is.informatik.uni-duisburg.de/

à une ontologie de la tâche. Une interface permet de consulter les documents associés aux concepts de l'ontologie de thème, éventuellement à la lumière des valeurs des métadonnées associées. C'est cette dernière approche qui sert de point de départ aux travaux présentés dans cet article.

2.2. UNITE DOCUMENTAIRE

Généralement, l'unité documentaire traitée (unité indexée et restituée) dans les systèmes de RI est le document tel que définit par son auteur (page web, article, etc.). D'autres approches se sont intéressées à des granularités plus fines. Par exemple, l'essor du langage SGML a conduit à des propositions permettant d'effectuer des recherches sur des parties de documents sémantiquement cohérentes car issues d'un marquage structurel du document (Wilkinson, 1994), (Corral et Mothe, 1995). Ces recherches se sont poursuivies avec XML, en particulier dans le cadre de INEX (2003), (Hubert *et al.*, 2007), (Lalmas et Tombros, 2007). Outre une recherche sur des unités plus petites et donc un meilleur ciblage de l'information, la prise en compte de la structure des documents permet d'affiner l'ordre dans lequel les éléments doivent être restitués à l'utilisateur et de naviguer dans le document. Dans le cadre de documents non structurés, ce sont des fenêtres de texte (Salton *et al.*, 1994) qui sont considérées ou des passages comme dans le programme d'évaluation TREC, TExt Retrieval Conference qui s'est intéressé à une granularité fine : la phrase (Harman, 2003).

2.3. PRESENTATION DES RESULTATS ET AIDE A LEUR EXPLORATION

Fréquemment, la présentation des résultats d'un moteur de recherche, s'effectue sous forme d'une liste de références à des documents, avec une possible utilisation de filtres sur les différentes métadonnées. Cela est par exemple le cas des moteurs Yahoo ©, Google © ou encore Stuff I've Seen (Dumais *et al.*, 2003) qui permet un accès rapide vers des documents consultés auparavant. La navigation par facettes, dans les interfaces web (Hearst, 2008) et les interfaces mobiles (Karlson *et al.*, 2006) permettent à l'utilisateur de naviguer en appliquant des couches successives de filtres. Alternativement, des approches graphiques permettent d'obtenir des représentations plus démonstratives de l'ensemble des éléments retrouvés ainsi que leurs liens (TouchGraph[2], Constellations[3] de Exalead, Gephi[4]). D'autres études s'intéressent à l'utilisation des tags (Zelevinsky et *al.*, 2008), (Shoy et Lui, 2006) ou encore à l'utilisation des topic maps (Rittershofer, 2005).

3. Propositions pour un ciblage des contenus pour la recherche d'information

Nous préconisons une approche permettant à l'utilisateur de cibler au mieux son besoin d'information, et cela à différents niveaux :

- Au niveau des contenus recherchés : il s'agit de permettre une recherche sémantique, c'est-à-dire une recherche basée sur le sens des mots plutôt que sur leur forme (morphologie). Cette recherche sémantique passe par l'utilisation d'une ontologie de domaine à la fois pour l'indexation et pour la recherche. Cette ontologie représente donc le langage d'indexation et de recherche. En complément, l'utilisateur peut inclure dans sa requête des restrictions sur les informations recherchées en s'appuyant sur les métadonnées associées comme la date de publication de l'information, son auteur, etc.

[2] http://www.touchgraph.com/navigator.html
[3] http://demos.labs.exalead.com/constellations/
[4] http://gephi.org/

- Au niveau de l'unité documentaire recherchée : l'utilisateur peut n'être intéressé que par une partie d'un document. Pour rendre possible une recherche par partie, nous proposons un découpage des unités documentaires initiales en unités plus fines. L'ensemble de ces unités seront indexées et retrouvées par le système. En fonction des demandes de l'utilisateur, il pourra être orienté vers une unité de petite taille sans avoir lui-même à retrouver l'information recherchée dans le document entier.

- Au niveau de la présentation des résultats : il s'agit d'aider l'utilisateur à accéder directement aux éléments d'information utiles via une mise en valeur de certains termes. Ces termes correspondent aux concepts de l'ontologie de domaine ayant permis de retrouver l'unité d'information. Cette mise en valeur permet de repérer plus rapidement l'information recherchée lors de la consultation.

3.1. RECHERCHE SEMANTIQUE

Espace de recherche

Afin de comprendre ce qui va suivre, il est indispensable de définir la notion d'espace de recherche. Un espace de recherche est l'ensemble des éléments nécessaires à l'établissement d'une recherche par l'utilisateur. Il est composé d'une ontologie de domaine et d'un corpus ou ensemble de documents. Une ontologie de domaine est elle même constituée d'une ontologie de thème, décrivant le langage utilisé par le processus d'indexation et de recherche d'information et d'une ontologie de la tâche, décrivant les métadonnées liées aux documents (Hernandez *et al.*, 2007). Ces métadonnées peuvent être de plusieurs types : titre, auteurs, date de création, etc. et sont extraites du schéma de représentation Dublin Core[5] (DCAPs). Les métadonnées utiles dépendent de l'espace de recherche. Nous utilisons un profil d'application (Duval *et al.*, 2002), c'est-à-dire un ensemble de règles et de métadonnées pour chaque espace de recherche ; ce profil est associé à l'ontologie de tâche.

Ajout d'un document

Un document est associé à un espace de recherche lors de son ajout.

Le document est ensuite indexé par rapport à l'ontologie de thème correspondant au domaine. Le moteur d'indexation extrait les syntagmes du texte à indexer et les compare aux termes nommant les concepts de l'ontologie. Lorsqu'une correspondance peut être faite, le concept de l'ontologie reconnu est retenu comme terme d'indexation.

Les valeurs des métadonnées associées à l'ontologie de la tâche sont également renseignées lors de l'ajout du document. Les métadonnées sont complétées conformément à la spécification donnée dans le profil d'application associé à l'ontologie de la tâche. Certaines métadonnées sont renseignées automatiquement via l'application de techniques d'extraction d'information, d'autres sont renseignées par l'utilisateur.

[5] www.

Recherche

Lors de la recherche, le choix de l'espace de recherche permet de réaliser un premier ciblage de l'information. Les autres ciblages sont réalisés par l'intermédiaire des ontologies de thème et de tâche.

Dans un contexte de RI, du fait des nombreux sens que peut avoir chaque terme en fonction du contexte dans lequel il est utilisé, les requêtes soumises par les utilisateurs sont souvent ambigües du point de vue du système. Au contraire, dans notre approche, lorsque l'espace de recherche a été spécifié, les requêtes sont construites en utilisant uniquement les concepts de l'ontologie de thème. La requête est automatiquement désambiguïsée puisque l'utilisation d'une ontologie permet de fournir un langage commun non ambigu pour formuler un besoin d'information (Aussenac et Mothe, 2004). D'autre part, l'ontologie utilisée pour la construction de la requête est la même que celle utilisée pour l'indexation des documents constitutifs du corpus.

L'ontologie de la tâche permet à l'utilisateur de définir les valeurs des métadonnées qui l'intéressent.

L'utilisation d'un espace de recherche permet ainsi de cibler rapidement l'information par le choix du corpus de documents, la désambiguïsation des requêtes de l'utilisateur grâce à l'utilisation de l'ontologie de thème, et le filtrage de l'information à restituer via les métadonnées relatives à chaque document.

Implantation

Afin de valider ces propositions, nous avons développé un prototype intégrant chacune des solutions que nous présentons.

Qu'il soit rédacteur ou lecteur, l'utilisateur doit d'abord choisir un espace de recherche qui définit l'ensemble des documents déjà disponibles et l'ontologie de domaine associée.

Lorsqu'un document est ajouté à un espace de recherche, le système génère un formulaire correspondant au profil d'application associé à l'ontologie de tâche. Les valeurs extraites automatiquement des documents sont présentées à l'utilisateur qui peut modifier et compléter le formulaire. Les informations sont stockées dans une base de données. Le système indexe ensuite le document en s'appuyant sur l'ontologie de thème. Les syntagmes du document sont extraits des documents puis comparés avec les termes nommant les concepts de l'ontologie de thème. L'association entre les concepts de l'ontologie et les termes des documents est stockée ainsi que la position des termes dans le document. La phase d'indexation n'est visible qu'au travers de l'affichage d'une trace d'exécution informant l'utilisateur des concepts et métadonnées insérés dans la base de données ainsi que du bon déroulement de l'indexation.

Un utilisateur lecteur peut utiliser le prototype pour effectuer ses recherches. La première action que l'utilisateur est invité à effectuer est de sélectionner l'espace de recherche qu'il souhaite consulter. Les ontologies de thème et de tâche lui sont alors présentées (voir figure 1). L'ontologie du thème (partie droite de l'écran) est représentée par un arbre de concepts, que l'utilisateur peut librement parcourir. La sélection d'un concept permet à l'utilisateur d'accéder aux documents qui référencent ce concept. L'utilisateur a aussi la possibilité de rechercher un concept en utilisant l'outil de recherche plutôt que le parcours de l'ontologie du thème. L'ontologie de la tâche, décrivant les différentes métadonnées est représentée dans le panneau de gauche (voir figure 1). L'utilisateur peut parcourir cette ontologie, sélectionner les valeurs qui correspondent à ses recherches et accéder aux documents associés.

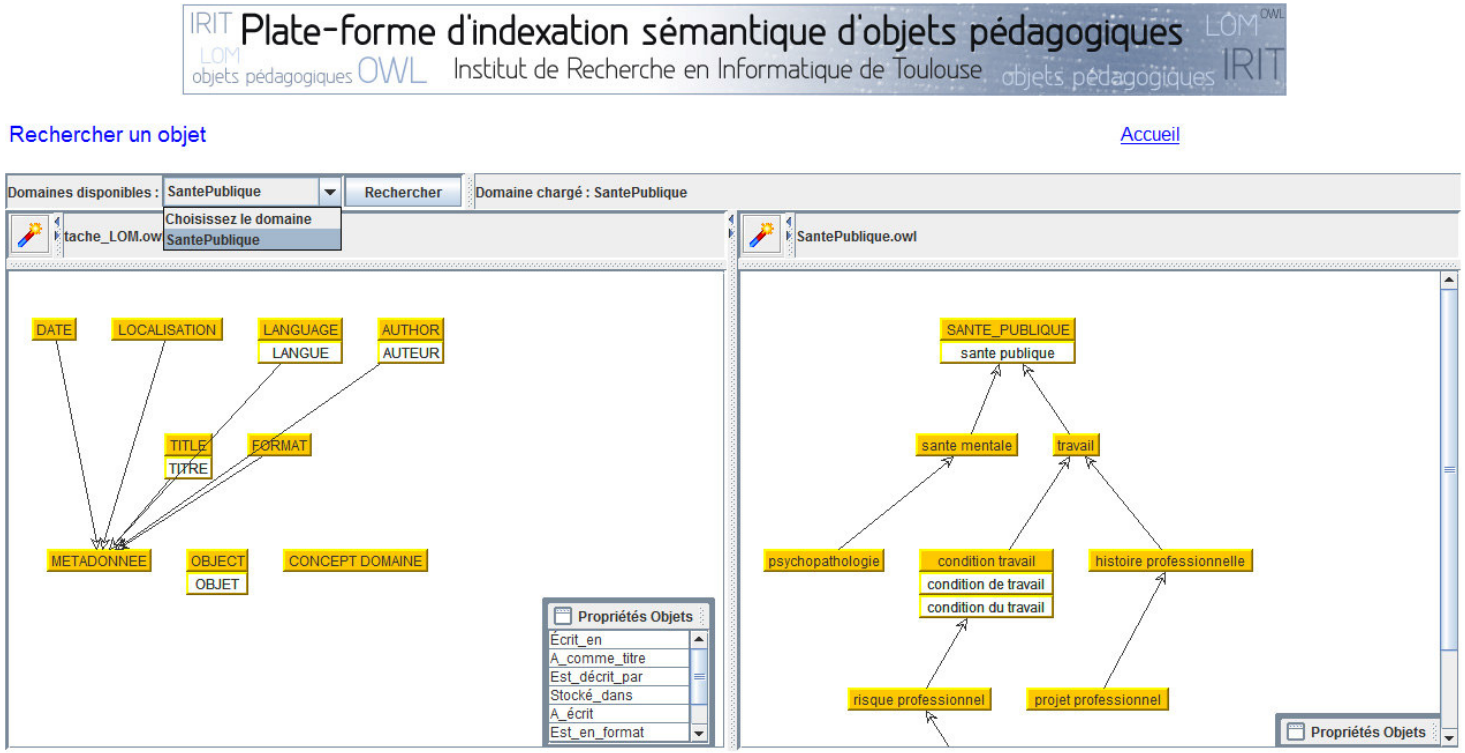

Figure 1 - Espace de recherche et navigation dans les ontologies de thème et de tâche.

3.2. UNITE DOCUMENTAIRE

Proposition

Outre le premier ciblage de l'information par le choix de l'espace de recherche, nous proposons un second niveau de ciblage. Il s'agit de prendre en compte la structure des documents. A la manière d'autres travaux (Chiaramella, 2001), (Hlaoua *et al.*, 2007), nous nous sommes intéressés à des granularités plus fines que le document tel que définit par son auteur. Ainsi, les unités présentées à l'utilisateur en réponse à une recherche pourront être plus focalisées sur son besoin.

Dans notre approche, le document initial est segmenté en plusieurs éléments. Cette segmentation est réalisée soit en fonction de la structure logique du document (en fonction des titres, des sous-titres, des paragraphes pour un document HTML par exemple.), soit en fonction de sa structure physique (un granule par page pour un document pdf par exemple). Cette segmentation du document initial est réalisée lors de son ajout à l'espace de recherche et entraine

une deuxième indexation. Chaque unité est indexée en utilisant les mêmes mécanismes que lors de l'indexation des documents entiers. Les liens existants entre les différents éléments et le document initial, ainsi que leur position sont stockés dans la base de données.

Lors de la recherche, la segmentation du document apporte à l'utilisateur un réel ciblage de l'information. L'utilisateur a la possibilité d'accéder uniquement à la partie du document qui contient l'information recherchée. Lorsqu'il navigue dans l'ontologie du thème, il accède à l'information non pas en la recherchant dans le document complet, tel que vu par son auteur, mais dans l'un de ses segments.

Implantation

Dans le prototype que nous avons réalisé, une première indexation est réalisée sur le document entier. Le document est ensuite segmenté de façon automatique. Une deuxième indexation est alors réalisée sur chaque unité générée lors de la segmentation.

Les différents segments des documents sont accessibles à partir de la navigation sur l'ontologie du thème comme présenté dans la section 3.1. Lorsque l'utilisateur sélectionne un concept répondant à son besoin d'information, le système affiche la liste des documents entiers ainsi que les parties de document référençant ce concept. L'utilisateur peut accéder à l'un de ses segments (voir le panneau de droite de la figure 2).

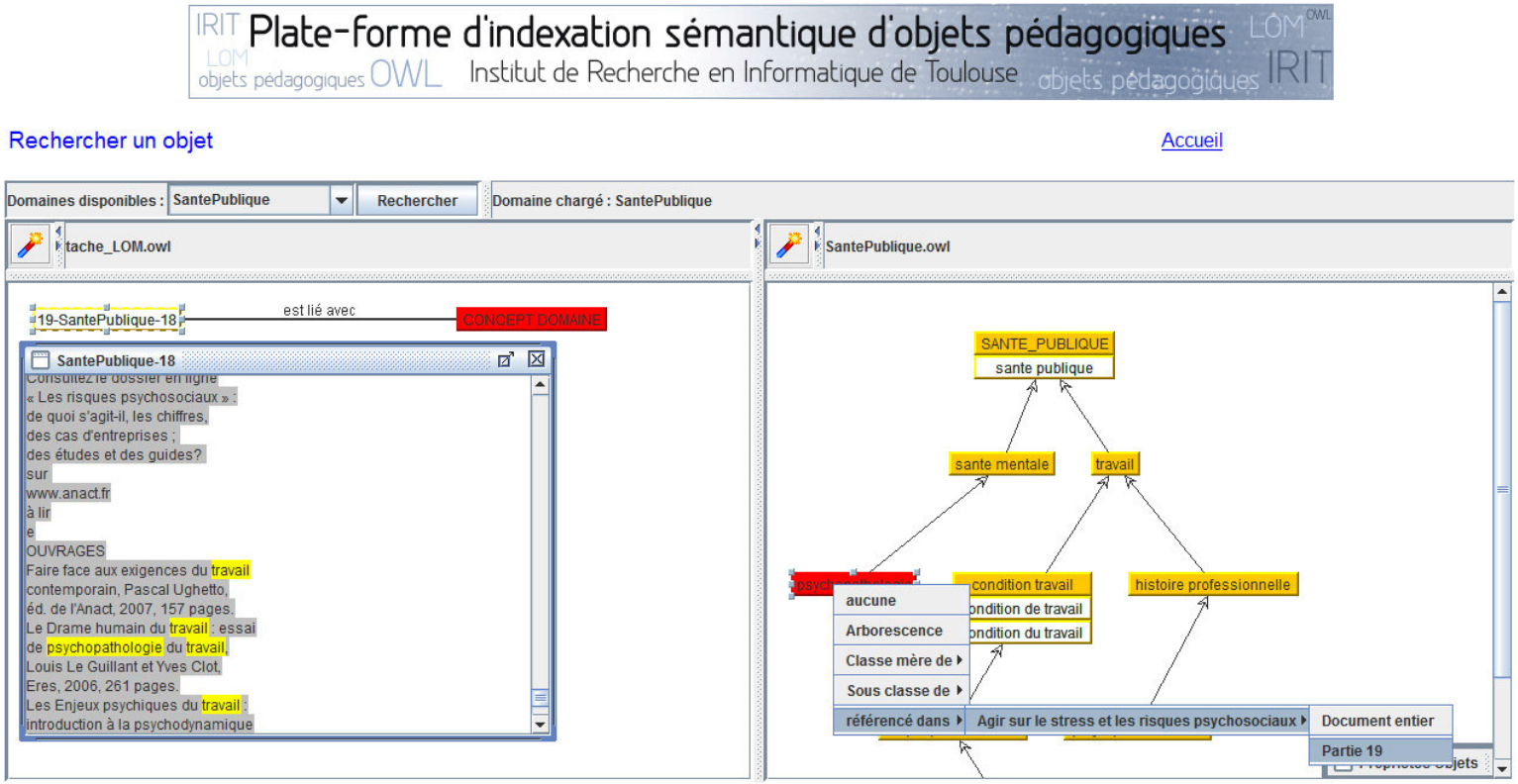

Figure 2 - Segments et Mise en valeur

3.3. PRESENTATION DES RESULTATS ET AIDE A LEUR EXPLORATION

Proposition

Le troisième niveau de ciblage concerne l'aide apportée à l'utilisateur pour faciliter l'accès à l'information qu'il recherche au sein d'une unité documentaire retrouvée. Nous proposons deux solutions pour y parvenir : l'affichage du contexte de l'unité documentaire retrouvée et la mise en valeur des termes relatifs à l'ontologie du thème.

Concernant l'affichage du contexte de l'unité documentaire, il s'agit de permettre à l'utilisateur de situer un segment de texte retrouvé au sein du document entier. En effet, la segmentation permet d'accéder directement à l'information recherchée mais il peut être utile à l'utilisateur de prendre connaissance des parties de textes qui environnent cette information.

La mise en valeur des termes a un objectif différent bien que basé sur le même principe de contextualisation de l'information. Il s'agit d'expliquer à l'utilisateur la raison de la sélection du document ou du segment de document par le moteur. Ainsi, les termes du document qui sont en lien avec les concepts de l'ontologie de thème sélectionnés par l'utilisateur sont mis en évidence dans le texte. L'utilisateur peut se focaliser sur les mots entourant ces termes.

Implantation

Afin de fournir à l'utilisateur une aide à l'accès à l'information qu'il recherche, nous avons décidé de mettre celle-ci en valeur. Pour ce faire, l'application implémente une fonctionnalité d'affichage de contenu. Cette fonctionnalité permet, outre l'affichage du texte, de mettre en surbrillance les concepts du domaine se trouvant dans le document. De plus, s'il s'agit d'un granule, l'on voit apparaître sur la présentation, le texte contenu par ce granule au sein du document entier.

4. 4. Cadre expérimental

Nous avons élaboré un cadre expérimental afin d'évaluer le modèle de représentation des informations que nous proposons. Nous présentons en premier lieu le contexte de l'évaluation puis les différentes données que nous avons obtenues.

4.1. CONTEXTE

Afin de réaliser l'expérience, il a été nécessaire de créer un espace de recherche. L'espace de recherche que nous avons conçu a pour thème la santé publique. Nous nous sommes basés sur le thésaurus de la santé publique[6] pour réaliser l'ontologie du thème. La partie du thésaurus concernant le travail a été complètement intégrée, ainsi que quelques concepts de divers sous-arbres tels que la santé mentale, la sociologie, la justice et la psychologie. L'ontologie de la tâche a été définie par une version simplifiée des métadonnées LOM. Nous avons choisi d'implanter seulement les métadonnées : auteur, date, format, langage, localisation et titre (voir partie de gauche, figure 1). Le corpus est composé de vingt-sept documents, collectés sur internet, en rapport avec le thème de la santé publique. La quantité assez faible des documents du corpus nous permet de simuler l'effet du premier ciblage de l'information qu'est le choix de l'espace de recherche.

L'expérience a eu lieu sur un panel de cinquante-deux étudiants, âgés de vingt à vingt-cinq ans, étant par conséquent relativement habitués aux moteurs de recherche type Google©.Chaque personne a reçu une formation de quelques minutes à l'utilisation du prototype. Nous avons ensuite regroupé les étudiants en binômes pour une durée de deux heures. Chaque groupe a eu pour tâche de répondre à huit questions en utilisant le prototype. Nous leur avons demandé de donner leurs avis sur les avantages et inconvénients du prototype par rapport à d'autres moteurs de recherche qu'ils connaissaient. Enfin, chaque binôme a dû répondre à un questionnaire permettant d'évaluer leur compréhension du fonctionnement du prototype.

[6] http://www.bdsp.ehesp.fr/Thesaurus/Default.asp

4.2. LES RELEVES

Les analyses quantitatives des résultats montrent que les utilisateurs ont du mal à appréhender une nouvelle forme d'interface d'accès à l'information et que la démonstration des fonctionnalités du prototype n'est pas suffisante pour permettre aux utilisateurs une recherche efficace. Le temps passé pour rechercher l'information est largement augmenté par rapport à une recherche par mots utilisée dans les moteurs comme Google.

L'analyse des réponses aux questionnaires montre que :

- les utilisateurs trouvent la recherche longue lorsqu'ils utilisent le prototype (cela est en cohérence avec l'analyse quantitative que nous avons réalisée en parallèle),
- la visualisation de termes associés permet de donner des pistes de recherche. On retrouve ici l'avantage de l'utilisation des thésaurus dans les systèmes documentaires qui les utilisent,
- le ciblage de l'information est vu comme un inconvénient. Les utilisateurs considèrent qu'ils n'ont pas assez d'information lorsqu'ils utilisent le prototype. Cette constatation devra être étudiée plus en détail. En effet, il sera intéressant d'analyser si cette impression de manque d'information a été ressentie pour des requêtes pour lesquelles aucune réponse pertinente n'a été retrouvée ou au contraire s'il s'agit d'une impression globale qui tendrait à dire que les utilisateurs souhaitent que les moteurs restituent beaucoup d'information, même s'ils n'en consultent ensuite que quelques unes.
- les utilisateurs souhaiteraient pouvoir écrire des requêtes plus longues. Dans le prototype, l'interrogation se fait par navigation et un seul terme (qui peut être un mot composé) peut être sélectionné à la fois. Dans ce cas là aussi, il serait intéressant de mettre en corrélation le souhait de pouvoir écrire des requêtes longues et la réalité de la moyenne des requêtes sur internet qui se situe entre 2 et 3 mots. Il est possible que les besoins d'information que nous avons suggérés aux utilisateurs étaient trop précis et qu'ils aient été tentés d'utiliser l'ensemble des mots du sujet dans leurs requêtes.

5. 5. Conclusion

Dans cet article, nous avons présenté une proposition de modalité d'accès à l'information qui permet un ciblage de l'information via des mécanismes variés qui opèrent dans les différents modules d'un moteur de RI. Nous proposons ainsi l'utilisation d'un espace de recherche, composé d'un ensemble de documents d'un domaine et d'une ontologie de domaine associée. Cette dernière est composée d'une partie relative au thème, l'autre décrivant la tâche au travers des métadonnées utiles pour la tâche. L'utilisation d'un espace de recherche permet de cibler rapidement l'information par le choix du corpus de documents, la désambiguïsation des requêtes de l'utilisateur grâce à l'utilisation de l'ontologie de thème, et le filtrage de l'information à restituer via les métadonnées relatives à chaque document. Le ciblage de l'information est également permis au travers de la prise en compte de la structure des documents (structure logique et structure physique). Enfin, concernant la présentation des résultats, le ciblage de l'information recherchée au sein des documents retrouvés est assuré par une mise en évidence visuelle. Le prototype a été présenté à un panel d'utilisateur qui l'a utilisé après une présentation de ses fonctionalités.

6. 6. Bibliographie

Aussenac-Gilles N., Mothe J., Ontologies as Background Knowledge to Explore Document Collections, RIAO, 2004, p. 129-142.

Benjamins R., Fensel D., Decker D., Gomez Perez A., (KA)2 : building ontologies for the internet : a mid-term report, International Workshop on ontological engineering on the global information infrastructure, 1999, p. 1-24.

Bozsak E., Ehrig M., Handschuh S., Hotho A., Maedche A., Motik B., Oberle D., Schmitz C., Staab S., Stojanovic L., Stojanovic N., Studer R., Stumme G., Sure Y., Tane J., Volz R., Zacharias V., KAON - Towards a Large Scale Semantic Web, EC-Web 2002, 2002, p. 304-313.

Corral M.-L., Mothe J., How to retrieve and display long structured documents ?, Basque International Workshop on Information Technology, BIWIT'95, 1995, p. 10-19.

Dumais S., Cutrell E., Cadiz J. J., Jancke G., Sarin R. et Robbins D. C., Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, p. 72–79.

INEX, Initiative for the Evaluation of XML retrieval http://qmir.dcs.qmw.ac.uk/INEX/.

Hearst M.A., Karadi C., Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, Inter. Conference on Research and Development in Information Retrieval, 1997, p. 246-255.

Hearst M. A., UIs for Faceted Navigation Recent Advances and Remaining Open Problems. Workshop on Computer Interaction and Information Retrieval, HCIR 2008, 2008, Redmond, p. 13-17.

Harman D., Overview of the TREC 2002 novelty track, actes de Text Retrieval Conference TREC 2002, 2003, p. 46-55.

Hernandez N., Mothe J., Chrisment C. et Egret D., Modeling context through domain ontologies, Journal of Information Retrieval, Special issue Contextual Information Retrieval, vol. 10, n°2, 2007, p. 143-172.

Khelif K. et Dieng-Kuntz R., Ontology-Based Semantic Annotations for Biochip Domain. Workshop on Knowledge Management and Organizational Memories, ECAI2004, 2004, p. 54-60.

Mothe J., Hubert G., Augé J., Englmeier K., Catégorisation automatique de textes basée sur des hiérarchies de concepts, Journées Bases de Données Avancées, 2003, p. 69-87.

Mukherjea S., Foley J.D., Hudson S., Visualizing complex hypermedia networks through multiple hierachical views, CHI, 1995, p. 331-337.

Salton G., Allan J., C. Buckley, Automatic structuring and retrieval of large text files, communication de l'ACM, vol. 37, n°2, 1994, p. 97-108.

Wilkinson R., Effective retrieval of structured documents, ACM Research and Development in Information Retrieval, SIGIR'94, 1994, p. 311-317.

A. Karlson, G. Robertson, D. Robbins, M. Czerwinski,et G. Smith. FaThumb: a facet-based interface for mobile search. Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 711–720, 2006.

V. Zelevinsky, J. Wang, and D. Tunkelang. Supporting Exploratory Search for the ACM Digital Library. Workshop on Human-Computer Interaction and Information Retrieval (HCIR'08), October 2008, p. 85 - 88.

Duval E., Sutton S. et Weibel SL. (2002). Metadata Principles and Practicalities D-Lib Magazine 8(4).