VICOTEXT : Un Outil de Visualisation Automatique et Dynamique de Connaissances Textuelles

Abdenour Mokrane

Groupe Connaissance - LGI2P - Ecole des Mines d'Alès, http://www.lgi2p.ema.fr
Parc scientifique Georges Besse - Site EERIE - F-30035 Nîmes cedex 1
Tél : +33 4 66 38 70 94 Fax : +33 4 66 38 70 74
abdenour.mokrane@ema.fr

Résumé

Les outils de visualisation de connaissances textuelles contribuent à l'efficacité et la pertinence des processus mis en œuvre en extraction de connaissances en offrant aux utilisateurs des représentations intelligibles de contenus textuels et facilitant l'interaction. Nous présentons dans ce papier l'outil VICOTEXT (VIsualisation de COnnaissances TEXTuelles) permettant une visualisation automatique et dynamique des connaissances textuelles d'une base documentaire d'une communauté d'utilisateurs travaillant sur une thématique donnée. Le fonctionnement de VICOTEXT est basé sur une approche originale d'extraction et de classification des connaissances textuelles prenant en considération les co-occurrences contextuelles et le partage de contextes entre les différents termes représentatifs du contenu.

Mots clés : Connaissance textuelle ; Hyper-carte d'information ; Terme représentatif.

1 Introduction

La quantité de documents et d'informations disponibles de nos jours, entraîne une « surinformation » de l'utilisateur final (organisation ou individu) qui n'est donc plus capable d'analyser ou d'appréhender ces informations dans leur globalité. Avec le Web, les documents textuels non structurés sont devenus prédominants. L'information utile étant enfouie dans le texte. Face à cette quantité d'information croissante que les entreprises sont contraintes de gérer, la visualisation de l'information devient une activité importante de la gestion de connaissances [4, 5]. Cette démarche d'extraction et de visualisation de connaissances textuelles est un processus de cartographie sémantique qui permet de passer des données à la carte.

Dans cet article nous présentons l'outil VICOTEXT de visualisation automatique et dynamique de connaissances textuelles d'une base documentaire d'une communauté d'utilisateurs travaillant sur une thématique donnée. Il permet une navigation automatique par le contenu, via les hyper-cartes d'information. VICOTEXT intègre des techniques issues de l'ingénierie de connaissances et la fouille de textes. Dans la section 2, nous présentons le fonctionnement général du système VICOTEXT, la section 3 illustre l'environnement et l'application de VICOTEXT sous forme d'une démonstration logicielle. Enfin, la section 3, conclue ce papier et présente les perspectives de recherche et développement associées.

2 VICOTEXT

2.1 Description

L'outil VICOTEXT a été développé en java, dans le cadre du projet CARICOU [1, 3] (CApitalisation de Recherches d'Informations de COmmunautés d'Utilisateurs) mené par le LGI2P. Cet outil permet une visualisation automatique du contenu d'une base de documents textuels, permettant ainsi une navigation automatique par le contenu via les hyper-cartes d'information. Le processus de visualisation des connaissances textuelles, via VICOTEXT comporte trois étapes, à savoir :

- Prétraitement linguistique et statistique
- Extraction et classification des termes représentatifs.
- Réalisation des hyper-cartes d'information.

2.2 Fonctionnement

La première étape concerne le prétraitement statistique et linguistique. Etant donné que nous ne nous intéressons pas ici au traitement automatique du langage naturel (TALN), nous utilisons pour le prétraitement linguistique l'analyseur de la société Synapse (http://www.synapse-fr.com). Le prétraitement concerne le nettoyage du corpus des mots vides et la détection des contextes des termes [1]. Un contexte correspond à une phrase.

Le prétraitement statistique concerne le calcul de la distribution des co-occurrences contextuelles et les matrices de co-occurrences contextuelles brut et réduite. La seconde étape concerne l'extraction et la classification des termes représentatifs du contenu. Les prétraitements linguistiques et statistiques ainsi que l'approche d'extraction des termes représentatifs du contenu sont

détaillés sur [1, 2, 3]. La classification des termes représentatifs est basée sur une mesure de similarité prenant en considération d'une part les fréquences des co-occurrences contextuelles des termes, d'autre part, les contextes partagés entre les termes représentatifs du contenu. La troisième étape concerne la visualisation. Chaque classe de termes représente une carte d'information, les termes sont représentés par des sommets de tailles variables dépendant de leurs poids, les liens entre les termes modélisent les poids liés aux mesures de similarité entre les termes. Au départ les différentes classes de termes sont représentés par un terme central (thème), la taille maximale d'une classe étant fixé à N, le processus de classification est relancé dès le dépassement de cette variable. Chaque terme non final (thème) est lié à une carte d'information détaillant le thème. Les cartes d'information s'auto-organisent d'une manière automatique et dynamique autour d'un thème central suite à une interaction ou à un souhait de l'utilisateur. La section suivante éclaircie le fonctionnement de VICOTEXT et illustre sa capacité de supporter la visualisation automatique et dynamique de différentes connaissances textuelles.

3 Illustration

Nous illustrons dans cette section l'outil VICOTEXT sur une base documentaire portant sur la thématique voyages. FIG.1 illustre l'environnement du système VICOTEXT et la carte d'information associée au contenu global de la base documentaire. FIG. 2 illustre la carte d'information autour du thème santé et FIG. 3 illustre l'auto organisation de la carte d'information autour du thème jeu suite aux différentes interactions de l'utilisateur.

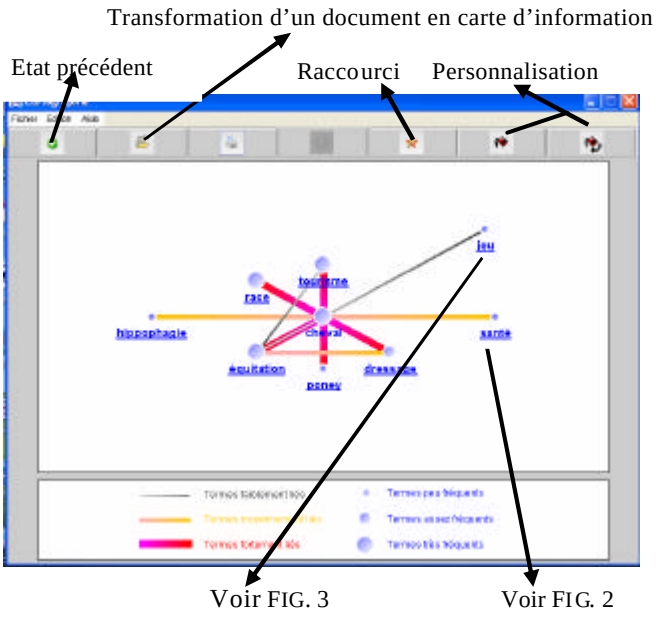

FIG. 1 - *Environnement VICOTEXT et carte d'information du contenu global de la base documentaire*

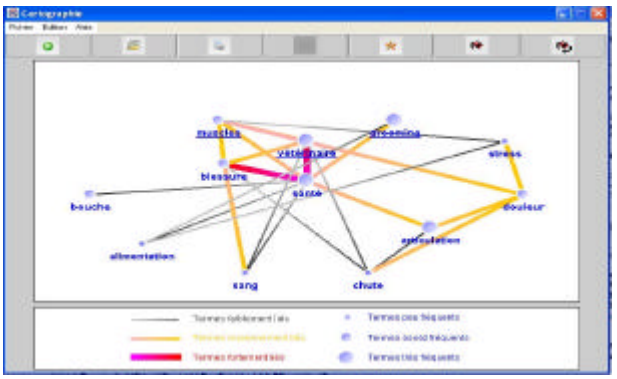

FIG. 2 - *Carte d'information autour du thème santé*

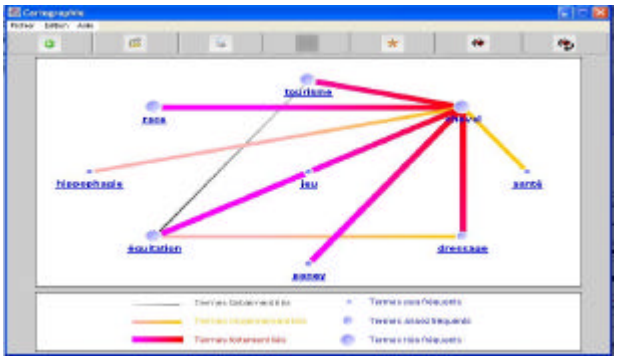

FIG. 3 - *Auto-organisation de la carte autour du thème jeu*

4 Conclusion

Nous avons présenté dans ce papier un outil basé sur une approche originale de visualisation automatique et dynamique de connaissances textuelles. Il serait intéressant d'intégrer l'outil à un système de consultation documentaire basé sur le partage de connaissances et l'échange d'expériences.

Références

[1] A. Mokrane, R. Arezki, G. Dray et P. Poncelet. Cartographie automatique du contenu d'un corpus de documents textuels. *JADT'04, Le poids des mots*, vol 2, pages 816-823, Louvain-la-Neuve : Presses Universitaires de Louvain, mars 2004.

[2] A. Mokrane et R. Arezki. Méthodologie de modélisation du contenu global d'un corpus documentaire par un graphe de liens sémantiques. *Actes des GRM'03*, Paris, décembre 2003.

[3] R. Arezki, A. Mokrane, G. Dray, P. Poncelet et D.W. Pearson. Modélisation dynamique et temporelle de l'utilisateur pour un filtrage personnalisé de documents textuels. *Actes des EGC'04, RNTI*, vol 2, pages 479-484, Clermont-Ferrand : Cépaduès Editions, janvier 2004.

[4] W. Chung, H. Chen and J. Nunamaker. Business intelligence explorer : A knowledge map framework for discovering business intelligence on the Web. *Proceedings of the 36 Hawaii International Conference on System Sciences (HICSS'03)*, Hawaii, 2003.

[5] T. Poibeau. *Extraction automatique d'information : du text mining au Web sémantique*. Paris : Edition Lavoisier, 2003.