# Visualisation automatique du contenu d'une base de documents textuels via les hyper-cartes d'information

Abdenour Mokrane, Pascal Poncelet, Gérard Dray

Groupe Connaissance et Systèmes Complexes
LGI2P – Site EERIE – EMA
Parc scientifique Georges Besse, F-30035 Nîmes cedex 1
Tél : +33 (0)4 66 38 70 94 Fax : +33 (0)4 66 38 70 74
Email : {abdenour.mokrane, pascal.poncelet, gerard.dray}@ema.fr

## Résumé

De nos jours, les utilisateurs (entreprises, organismes ou individus) se trouvent submergés par la quantité d'information et de documents disponibles. Avec le Web, ce phénomène est de plus en plus important, les utilisateurs ne sont donc plus capables d'analyser ou d'appréhender ces informations dans leur globalité. Les documents textuels non structurés sont devenus prédominants et les informations utiles étant enfouies dans les textes, il devient indispensable de proposer de nouveaux systèmes pour extraire et visualiser de manière automatique l'information contenue dans les corpus de documents textuels. Nous proposons dans cet article le système VICOTEXT de visualisation automatique et dynamique du contenu d'une base de documents textuels via une navigation par les hyper-cartes d'information. 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. VICOTEXT permet d'aider une communauté d'utilisateurs travaillant sur une thématique donnée dans ses consultations documentaires en lui proposant des hyper-cartes d'information sur le contenu global de la base documentaire ou de chacun des documents. Nous illustrons VICOTEXT sur une base de documents textuels thématique.

## Mots-clés

Fouille de données textuelles, Cartographie d'information, Visualisation de connaissances textuelles, Hyper-carte d'information, Partage de contextes.

# 1. Introduction

Chaque jour, en particulier en raison de l'essor des communications électroniques, le nombre de documents disponibles croît de manière exponentielle et l'utilisateur (entreprise, organisme ou individu) se trouve submergé par la quantité d'information disponible. Ces utilisateurs ne sont donc plus capables d'analyser ou d'appréhender ces informations dans leur globalité. Avec le Web, les documents textuels non structurés sont devenus prédominants et les informations utiles étant enfouies dans les textes, il devient indispensable de proposer de nouveaux systèmes pour extraire et cartographier de manière automatique l'information contenue dans les corpus de documents textuels.

Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent à la fouille de corpus de documents textuels volumineux [Poi03 ; Bes01 ; Che&al01 ; He&al01]. Ces travaux ont donné naissance à des systèmes de catégorisation, voire de cartographie de documents tel que Kartoo [Chu&al03] ou Mapstan [Spi02]. Ces outils retrouvent des liens entre les différents documents ou sites Web et représentent ces liens sous forme de carte de navigation. Cependant les informations du contenu du corpus documentaire ou de chacun des documents sont peu représentées.

Nous proposons dans cet article le système VICOTEXT (VIsualisation de Connaissances TEXTuelles) de visualisation automatique et dynamique du contenu d'une base de documents textuels via une navigation par les hyper-cartes d'information. 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. VICOTEXT permet d'aider une communauté d'utilisateurs travaillant sur une thématique donnée dans ses consultations documentaires en lui proposant des hyper-cartes d'information sur le contenu global de la base documentaire ou de chacun des documents.

L'article est organisé de la manière suivante. La section 2 présente le système VICOTEXT et l'algorithme générale de cartographie du contenu. L'approche d'extraction des termes représentatifs du contenu et l'algorithme associé ainsi que les différentes analyses linguistiques et statistiques sont détaillés dans la section 3. La section 4 présente notre méthodologie de construction des hyper-cartes d'information et de navigation. Nous illustrons le système VICOTEXT à la section 5. La section 6 expose les travaux de fouille de texte liés à notre problématique de cartographie du contenu. Enfin, la section 7 conclue ce papier et présente les perspectives de recherche associées.

# 2. Le système VICOTEXT

Le système VICOTEXT, est développé en java dans le cadre du projet CARICOU (CApitalisation de Recherches d'Informations de COmmunautés d'Utilisateurs) du LGI2P. VICOTEXT se charge de la cartographie et la visualisation automatique et dynamique du contenu d'un corpus de documents textuels d'une thématique donnée. Il est composé de deux sous systèmes, le premier sous-système est dédié aux analyses linguistiques et statistiques pour l'extraction des termes représentatifs du contenu et des relations sémantiques ainsi que leurs distributions dans le contenu (Cf. section 3 - Extraction des termes représentatifs du contenu et leurs associations ). Le second sous système exploite les données textuelles analysées par le premier sous système dans l'objectif de construire des hyper-cartes

d'information et de navigation (Cf. section 4.1 - définition d'une hyper-carte d'information) permettant la visualisation du contenu. La figure 1 illustre l'architecture fonctionnelle du système VICOTEXT. Voir algorithme 1 pour la description de l'algorithme général de cartographie et de visualisation du contenu.

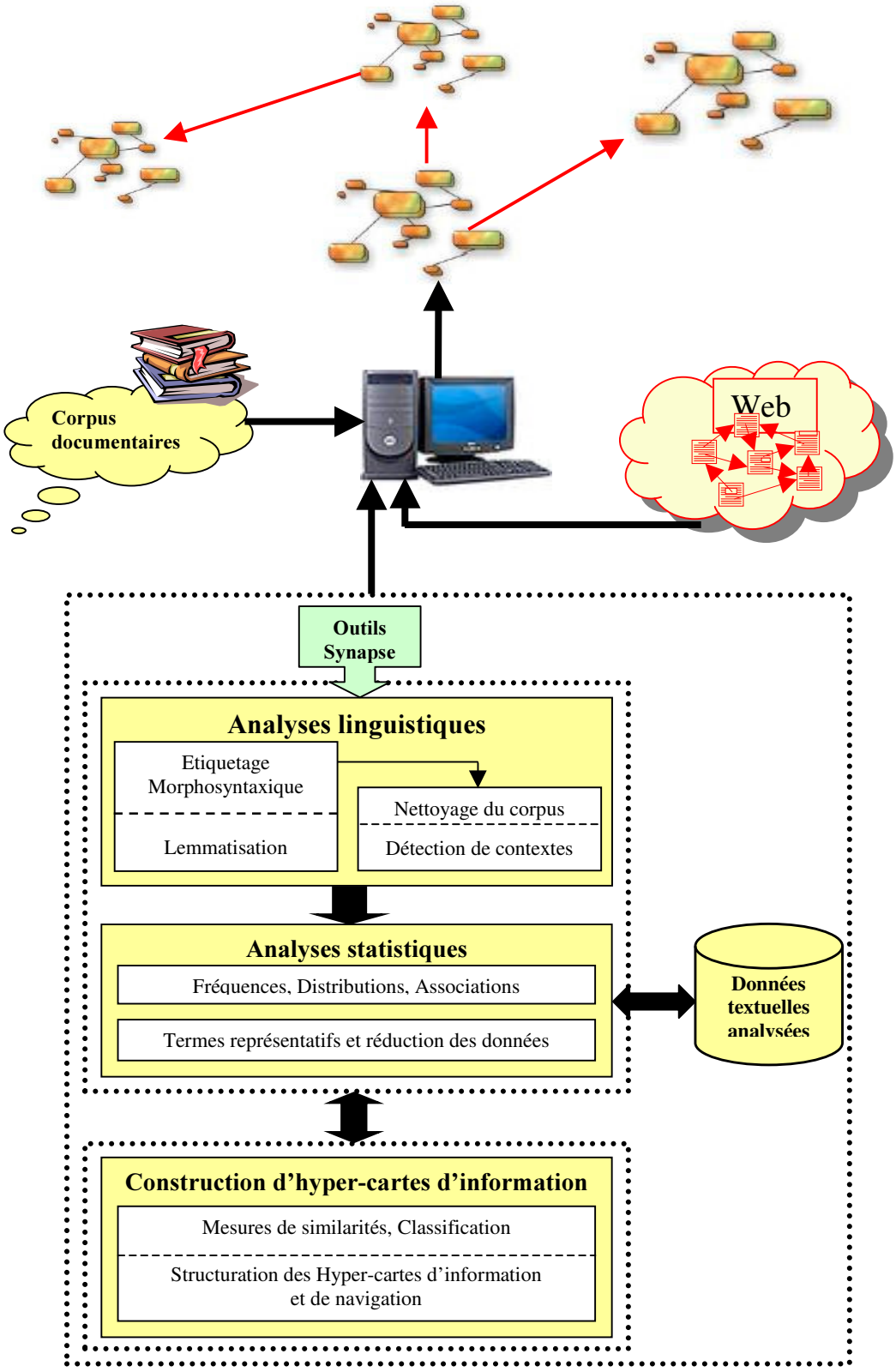


**Figure 1**. *Architecture fonctionnelle du système VICOTEXT*

***Algorithme 1: Visualisation automatique de contenus textuels***

Entrée: Corpus de documents textuels

Sortie: Hyper-cartes d'information et de navigation

DEBUT

Pour chaque documents

(1) analyse morphosyntaxique

(2) élimination des mots vides (garder noms, verbes et adjectifs)

(3) détection des contextes (un contexte = une phrase)

(4) analyse statistique des données (Distribution des termes et leurs associations )

A partir des données textuelles analysées

(1) Extraction des termes représentatifs du contenu

(2) Structuration des données (Données réduites)

A partir des différentes structures de données

(1) Calcul de mesures de similarités

(2) Application d'un algorithme de classification

(3) Réalisation des hyper-cartes d'information

FIN

# 3. Extraction des termes représentatifs du contenu et leurs associations

## 3.1 Analyses linguistiques et statistiques

Etant donné que nous ne nous intéressons pas ici au traitement automatique du langage naturel (TALN), nous utilisons l'analyseur de la société Synapse (http://www.synapse-fr.com) qui intègre un étiqueteur morphosyntaxique et un lemmatiseur fonctionnant pour les documents textuels en Français. Via ces deux outils, le corpus de documents textuels est transformé en documents étiquetés et lemmatisés. Les prétraitements linguistiques concernent le nettoyage du corpus des mots vides (articles, pronoms, prépositions, etc.) et la détection des différents contextes. A l'aide des étiquettes, nous conservons les noms, les verbes et les adjectifs. De manière générale, dans les différentes approches existantes, un contexte de co-occurrence peut être une phrase, un paragraphe ou même l'ensemble du document. Etant donné que nous considérons que les éléments pertinents sont généralement proches dans un document, nous considérons dans notre modèle qu'un contexte correspond à une phrase et ainsi la détection des contextes va correspondre à l'annotation des différentes phrases du corpus documentaire.

## 3.2 Définitions

**Co-occurrence contextuelle (*CO*) :** Deux termes *A* et *B* appartenant, en même temps au même contexte, forment une co-occurrence appelée *CO* et notée *{CO : A—B}*.

**Fréquences d'un terme (*FTC* et *FTD*) :** La fréquence *FTC* d'un terme *T* dans un corpus de documents textuels correspond au nombre d'occurrences du terme *T* dans le corpus. La

fréquence *FTD* d'un terme *T* dans un corpus de documents textuels correspond au nombre de documents contenants *T*.

**Fréquences d'une co-occurrence (*FCC* et *FCD*) :** La fréquence *FCC* d'une co-occurrence *CO* dans un corpus de documents textuels correspond au nombre d'occurrences de *CO* dans le corpus. La fréquence *FCD* d'une co-occurrence *CO* dans un document *D* correspond au nombre d'occurrences de *CO* dans *D*.

**Matrice de co-occurrences brute (*MATCO*) :** Soit *N* le nombre de termes d'un corpus documentaire. La matrice de co-occurrence brute notée *MATCO* correspond à une matrice de *N* lignes et *N* colonnes. La ligne *i* de la matrice correspond à un terme *Ti* du corpus et la colonne *j* de la matrice correspond à un terme *Tj* du corpus (*i= 1..N, j= 1..N).*

*Si (i ≠ j)  MATCO (i,j)=FCC* de *{CO : Ti—Tj}  Sinon  MATCO (i,j)=FTC de Ti*

**Matrice de co-occurrences réduite (*RMATCO*) :** Soit *E* l'ensemble des termes d'un corpus documentaire et considérant l'ensemble $R \subset E$, contenant *M* termes. La matrice de co-occurrences réduite notée *RMATCO* correspond à une matrice de *M* lignes et *M* colonnes. La ligne *i* de la matrice correspond à un terme *Ti* de l'ensemble *R* et la colonne *j* de la matrice correspond à un terme *Tj* de l'ensemble *R*. (*i= 1..M, j= 1..M).*

*Si (i ≠ j)  RMATCO (i,j)=FCC de {CO : Ti—Tj}  sinon  RMATCO (i,j)=FTC de Ti*

Les pré-traitements statistiques consistent à calculer les *FTC, FTD, FCC et MATCO* (définies ci dessus) dans l'objectif de sélectionner les termes représentatifs du contenu du corpus documentaire.

### 3.3 Algorithme de sélection des termes représentatifs du contenu

Le choix des termes représentatifs du contenu du corpus prend en considération les co-occurrences contextuelles les plus fréquentes en plus des fréquences de termes, en tenant compte de la distribution *FTD* de chacun des termes, suivant l'algorithme décrit ci-dessous :

***Algorithme2:*** Sélection des termes représentatifs du contenu

Entrée: Matrice *MATCO du corpus,*

Vecteur $Vdist = <(T_1, FTD_1), \ldots (T_i, FTD_i), (T_n, FTD, T_n)>$

Sortie: *E* ensemble de termes représentatifs

DEBUT

1. $E \leftarrow \varnothing$
2. pour *chaque terme* $T_i$ faire

   si $\left( FTC_i / FTD \right) > \alpha$ *alors*  $E = E \cup \{T_i\}$
3. pour *chaque* $\{CO : T_i — T_j\}$ faire

   si *FCC* de $\{CO : Ti—Tj\} > \beta$ *alors* $E = E \cup \{T_i\} \cup \{T_j\}$

FIN

$T_i$ , $T_j$*: Termes du corpus ;* $FTD_i$ *est la fréquence* FTD *de* $T_i$ ; $\alpha$ et $\beta$ *constantes représentantes des seuils de sélection des termes.* Dans notre modèle nous avons choisis les seuils suivants, de façon à mettre au même poids l'apport de l'occurrence des termes et leurs co-occurrences [Mok04&al04 ; Mok&Are03].

Le seuil $\alpha$ est égal à la moyenne *MoyFTC* des *FTCi* des termes *Ti* pondérées par les *FTD, ce qui se traduit par*

$$MoyFTC = \frac{\sum_{i=1}^{i=M} \left( FTC_i / FTD \right)}{M}$$, avec *M* le nombre de termes du corpus.

Le seuil $\beta$ est égal à la moyenne *MoyFCO* des *FCC* des co-occurrences contextuelles *CO* du corpus. Ce qui se traduit par :

$$MoyFCO = \frac{\sum_{i=1}^{i=N} FCC_i}{N}$$, avec *N* le nombre de co-occurrences contextuelles du corpus.

Après avoir construit l'ensemble *E* des termes représentatifs du contenu, nous procédons à la construction de la matrice de co-occurrences contextuelles réduite *RMATCO* (Cf. définitions) sur l'ensemble *E*. Cette matrice *RMATCO* va nous servir à l'étape de réalisation des hyper-cartes d'information, détaillée dans la section suivante.

## 4. Construction des hyper-cartes d'information

**4.1 Définition :** une hyper-carte d'information est un graphe $G = <X, U>$ où *X* est un ensemble de *N* sommets modélisants *N* termes représentatifs et *U* un ensemble d'arêtes représentant les associations et leurs poids dans le corpus documentaire. Chacun des sommets est représenté par une boule à dimension graphique reflétant le poids du terme et d'un texte représentant le terme. Les sommets du graphe sont des hyperliens (liens dynamiques) à deux fonctionnalités. La première fonctionnalité permet d'auto-organiser (organiser d'une manière automatique) le graphe autour du sommet (terme représentatif) en cliquant sur la boule, la deuxième fonctionnalité permet d'ouvrir une nouvelle hyper-carte d'information suite à un clic sur le texte. Si le texte d'un sommet n'est pas souligné, cela signifie qu'on est devant une feuille ou un terme final (la deuxième fonctionnalité n'est donc pas disponible).

### 4.2 Méthodologie

Afin de réaliser les hyper-cartes d'information permettant la visualisation du contenu, nous appliquons un algorithme de Classification Hiérarchique ascendante (*CAH*) à l'ensemble des termes représentatifs du contenu, à cet effet nous calculons une matrice de similarités textuelles à partir de la matrice *RMATCO*, notre approche de calcul des différentes mesures de similarités prend en considération le partages de contextes et les co-occurrences contextuelles entre les différents termes représentatifs du contenu. Afin de détailler l'algorithme de calcul des mesures de similarité, nous adoptons les définissions suivantes, soit A, B deux termes représentatifs :

$$(A \wedge B) = \{T \in E / \{CO : A—T\} \wedge \{CO : B—T\}\}$$

$$(A \wedge \neg B) = \{T \in E / \{CO : A—T\} \wedge \neg\{CO : B—T\}\}$$

$$(\neg A \wedge B) = \{T \in E / \neg\{CO : A—T\} \wedge \{CO : B—T\}\}$$

$$(\neg A \wedge \neg B) = \{T \in E / \neg\{CO : A—T\} \wedge \neg\{CO : B—T\}\}$$

Cette dernière définition $(\neg A \wedge \neg B)$ n'est pas utilisée dans le cadre de nos calculs.

avec ¬{*CO : B—T*} et ¬{*CO : A—T*} signifie que les couples <*B,T*> *et* <*A,T*> ne forment pas forment pas respectivement des co-occurrences contextuelles.

**Exemple :** soit *A* et *B* deux termes représentatifs du contenu du corpus documentaire et soit l'ensemble des termes représentatifs : *C*, *D*, *E*, *F*, *G*, *H* et *I*. La figure 2 illustre les relations de co-occurrences entre tous ces termes représentatifs : *A*, *B*, *C*, *D*, *E*, *F*, *G* , *H*, *I* .

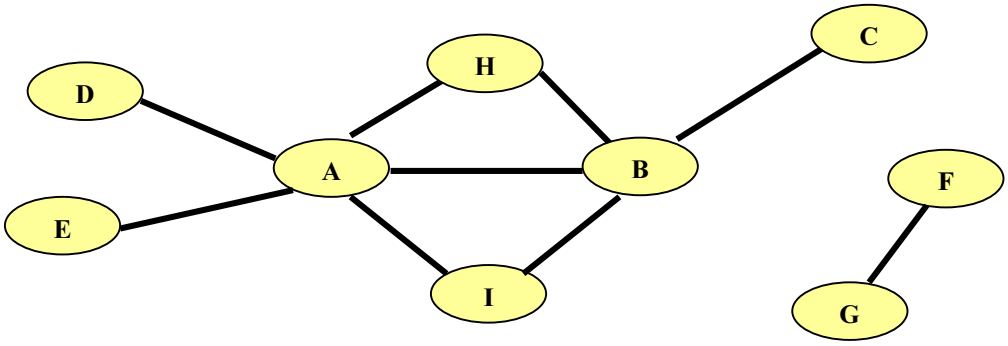


Figure 2. *Exemple d'Illustration*

Pour l'exemple ci-dessus :

*(A ∧ B) = {I, H} ; (A ∧¬B) = {D, F} ; (¬A ∧ B) = {C} ; (¬A ∧¬B) = {F, G}*

A partir de ces définitions nous pouvons détailler l'algorithme de calcul des mesures de similarités entre les termes représentatifs du corpus qui prend en considération les contextes partagés partage de contextes partagés entre chaque couple de termes <*A, B*> ; i.e. *(A ∧ B)* ; *(A ∧¬B)* ; *(¬A ∧ B)* en plus des fréquences de co-occurrences *FCC* de *{CO : A—B}*.

***Algorithme3**: Calcul des mesures de similarités entre les termes représentatifs*

*Entrée: Matrice réduite RMATCO, E = {*$T_1, \ldots T_i, \ldots T_m$*} termes représentatifs avec m le nombre des termes représentatifs.*

*Sortie: Matrice de similarités MATSIM*

*Début*

Si *i≠j*

pour i allant de 1 à m faire

pour j allant de 1 à i-1 faire

Calculer $A$ = Card ($T_i \wedge T_j$) ; Calculer $B$ = Card ($T_i \wedge \neg T_j$) ;

Calculer $C$ = Card ($\neg T_i \wedge T_j$)

$Sim1 = A /( A + B + C)$

$S1 = FCC$ de $\{CO : T_i — T_j\} / \sum FCO$ de $\{CO : T_i — T_k\}$, k = 1..m ;

$S2 = FCO$ de $\{CO : T_j — T_i\} / \sum FCO$ de $\{CO : T_j — T_k\}$, k= 1..m ;

$Sim2 = ½ (S1 + S2)$

$MATSIM (i,j) = \alpha * Sim1 + (1-\alpha) Sim2$

*Sinon MATSIM (i,j)=1*

*Fin*

*Card signifie cardinal d'un ensemble,* $\alpha = 0.5$ (mise au même poids l'apport des fréquences de co-occurrences contextuelles et du partage de contextes entre les termes représentatifs du contenu; $\sum FCO$ *de {CO :* $T_i$ *—* $T_k$*)* et $\sum FCO$ *de {CO :* $T_j$ *—* $T_k$*) ne sont pas égaux puisque* $T_i$ et $T_j$ ne partagent pas forcément les mêmes contextes, il est clair que la matrice de similarités est symétrique et donc l'algorithme détaille les calculs sur la moitié de la matrice *MATSIM.*

A partir de la matrice de similarités *MATSIM* nous calculons la matrice de dissimilarités *MATDIS* = 1—*MATSIM,* puis nous appliquons un algorithme de Classification Hiérarchique ascendante (*CAH*) à cette dernière matrice. Le résultat du *CAH* est un arbre de classification. Pour obtenir des classes, la méthode la plus usuelle est de couper l'arbre à un niveau (ou seuil). Ils existent plusieurs méthodes numériques, dépendant de la méthode de classification, qui indiquent le seuil optimum. Etant donner que notre objectif n'est pas de développer des méthodes de classification automatique, nous avons choisis aussi la méthode la plus utilisée pour la coupure de l'arbre (coupure d'arbre à la plus grande rupture). Après la classification, chaque classe de termes représente l'ensemble des sommets d'une hyper-carte d'information. Au départ chaque classe de termes est représentée par un terme central (terme ayant le poids le plus élevé). Chacun de ces termes centraux représente un hyperlien (lien dynamique) lié à l'hyper-carte d'information représentant sa classe d'appartenance. La taille maximale d'une classe étant fixé à N. le processus de classification est relancé dès le dépassement de cette variable. Ce qui permet d'éclater une classe en une ou plusieurs autres classes, i.e. une carte d'information en une ou plusieurs cartes, permettant ainsi une visualisation du contenu à travers une navigation à travers des hyperliens. La section suivante éclaircie le fonctionnement de VICOTEXT à travers une illustration sur une corpus de documents textuels thématique, via des copies écran du prototype actuel.

## 5. Illustration

Nous illustrons dans cette section le prototype VICOTEXT sur un corpus de documents textuels portant sur la thématique voyages et chevaux, à travers des copies écrans. La Figure 3 illustre l'environnement du système VICOTEXT et la carte d'information associée au contenu global. La Figure 4 illustre l'auto-organisation de la carte d'information autour du thème cheval. La figure 5 illustre la carte d'information autour du thème santé suite à une interaction de l'utilisateur lié à un besoin d'approfondir ses connaissances sur le contenu global du corpus concernant le thème santé et les chevaux.

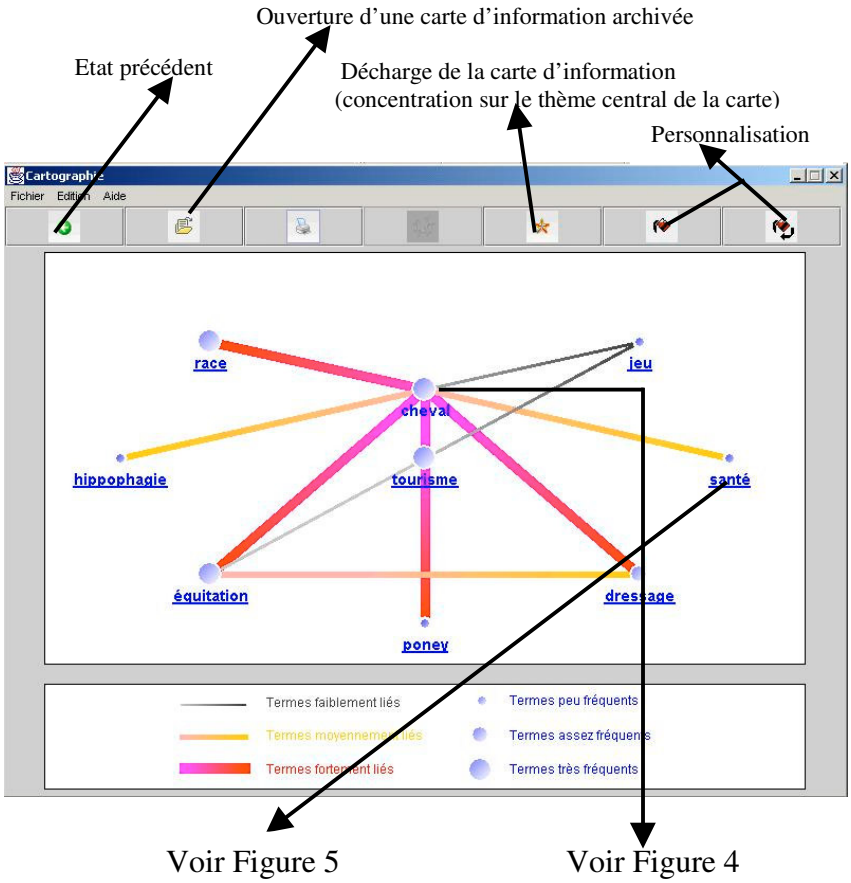


**Figure 3** – *Environnement VICOTEXT et carte d'information du contenu global*

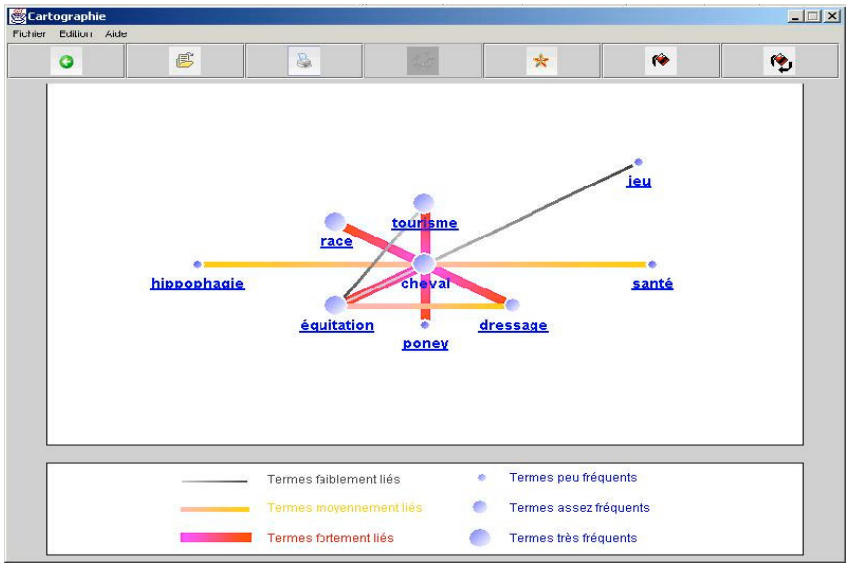


**Figure 4**. *Auto-organisation de la carte autour du thème cheval*
*(Cheval devient le thème central de la carte suite au souhait de l'utilisateur)*

La figure 6, illustre la décharge de la carte d'information de la figure 5 suite à la concentration de l'utilisateur sur le thème santé.

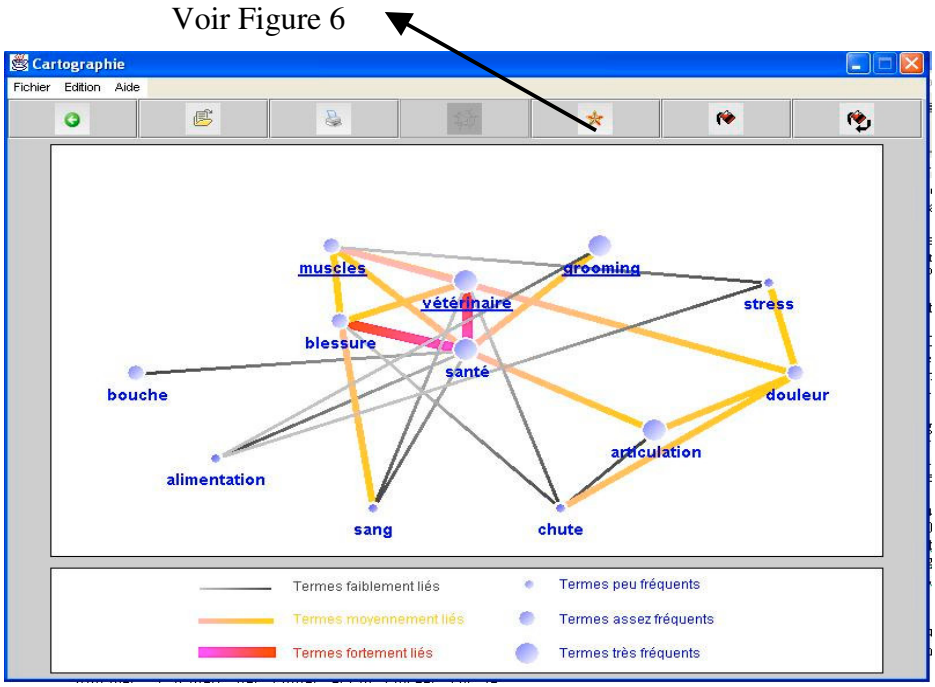


**Figure 5**. *Approfondissement de la carte d'information autour du thème santé*

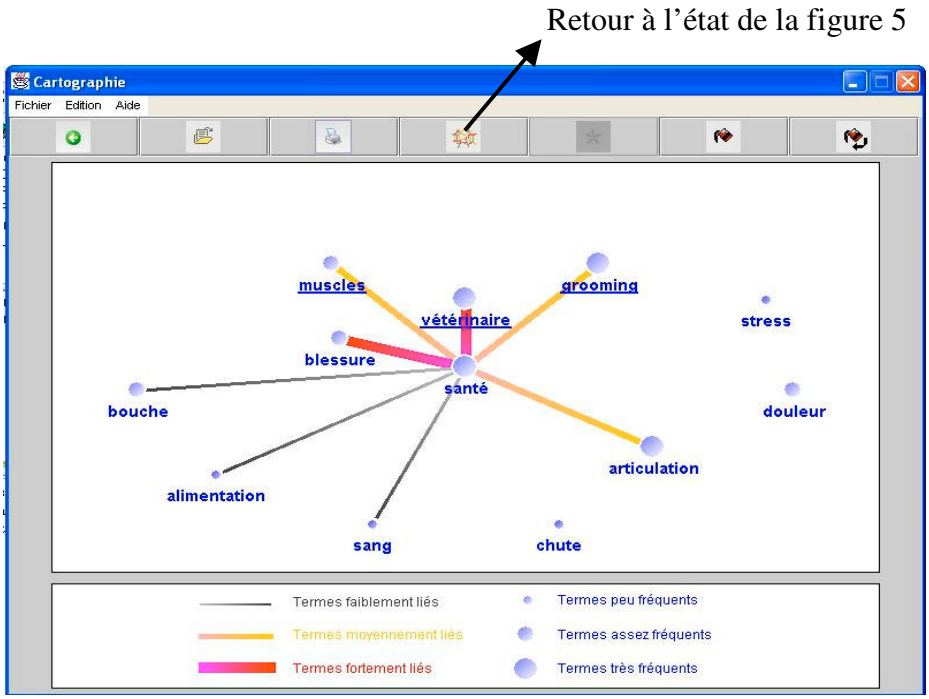


**Figure 6.** Décharge de la carte d'information, concentration de l'utilisateur sur le thème santé

## 6. Travaux connexes

Les outils et les méthodes de fouille de textes permettent l'acquisition, le classement, l'analyse, l'interprétation, l'exploitation et la visualisation systématiques d'informations

contenues dans des documents textuels [Poi03]. Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining [Blo&Cos00] et du Text Mining, s'intéressent à la fouille de corpus documentaires volumineux [And00 ; Bes01 ; Che&al01 ; Han&Kam00 ; He&al01 ; Tur00]. L'objectif de ces travaux est généralement d'analyser le contenu des documents pour en extraire des termes significatifs ainsi que les liaisons qui peuvent exister entre ces différents termes. Dans ce cadre, les modèles de similarités textuelles et la notion de co-occurrences sont les plus utilisées pour l'analyse du contenu [Poi03]. Dans un contexte proche, celui de la recherche documentaire, la recherche de co-occurrences a également été largement étudiée ces dernières années, elle consiste à rechercher les associations de termes les plus fréquentes dans les documents afin de retrouver rapidement les documents pertinents qui peuvent répondre aux requêtes de l'utilisateur. Dans [Per&al93] cette co-occurrence est utilisée pour la classification des termes selon la distribution de leurs contextes syntaxiques. [Tan&al96] utilise la matrice de co-occurrences pour la désambiguïsation des termes. [Bes&al02] ont proposé un modèle de filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents et la recherche documentaire. Tous ces travaux de fouille de textes, ne prennent pas en considération les co-occurrences dans le choix des termes de l'espace vectoriel ou les termes représentatifs du contenu d'un corpus documentaire. Ce qui implique une pénalisation d'une partie importante des relations de co-occurrences. Cette pénalisation est due au choix de termes basé principalement sur leurs fréquences d'occurrences [Sal&al75 ; Sal&Buc88], ces méthodes ne contiennent pas de stratégie, prenant en considération les associations des termes et leurs occurrences, pour sélectionner les termes représentatifs d'un corpus documentaire d'une thématique donnée. L'application de ces approches pour la cartographie du contenu des documents est donc limitée dans la mesure où elle ne permet pas une extraction des informations pertinentes et une visualisation représentative du contenu du corpus documentaire.

## 7. Conclusion et perspectives

Nous avons présenté dans cet article le système VICOTEXT de cartographie et de visualisation du contenu d'un corpus de documents textuels via les hyper-cartes d'information. Ce système développé en java dans le cadre du projet CARICOU du LGI2P met en œuvre notre 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. VICOTEXT permet d'aider une communauté d'utilisateurs, travaillant sur une thématique donnée, dans ses consultations documentaires en lui proposant des hyper-cartes d'information sur le contenu global de la base documentaire ou de chacun des documents. Nous avons illustré l'environnement du système VICOTEXT et sa capacité de supporter la visualisation automatique et dynamique de connaissances textuelles cohérentes à partir d'une base de documents textuels thématique, via une navigation par les hyper-cartes d'information. Nos travaux en cours portent sur l'extension du système VICOTEXT en lui intégrant des fonctionnalités permettant la détection des intérêts d'une communauté d'utilisateurs, favorisant le partage de connaissances et l'échange d'expériences entre les différents membres de la communauté, dans une tâche de consultation et de recherche documentaire.

## 8. Références

[And00] Andrieu O. « Créer du trafic sur son site Web ». Edition Eyrolles, 2000.

[Bes02] Besançon R. « Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents ». *Conférence TALN*, Nancy, 2002.

[Bes01] Besançon R. « Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes ». PhD thesis, Ecole polytechnique Fédérale de Lausanne, 2001.

[Chu&al03] Chung W., Chen H. and Nunamaker J. « 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.

[Che&al01] Chen H., Fan H., Chau M. and Zeng D. « MetaSpider: Meta-searching and categorization on the Web ». *Journal of the American Society for Information Science and Technology*, vol. 52, pages. 1134 –1147, 2001.

[Blo&Kos00] Kosala R. and Blockeel H. «Web Mining research : A survey ». *SIGKDD Explorations*, 2(1), pages 1-15, 2000.

[He&al01] He X., Ding C., Zha H. and Simon H. « Automatic topic identification using Webpage clustering ». *Proceedings of 2001 IEEE International Conference on Data Mining*, Los Alamitos, CA, 2001.

[Han&Kam00] Han. J. and Kamber. M. «Data Mining : Concepts and Techniques », Morgan Kaufmann Publishers, 550 pages, 2000

[Mok&al04] Mokrane. A, Arezki. R, Dray. G et Poncelet. P. « 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.

[Mok&Are03] Mokrane A et Arezki. R. « Méthodologie de modélisation du contenu global d'un corpus documentaire par un graphe de liens sémantiques ». *Actes des GRM'03*, Paris, 2003.

[Poi03] Poibeau T. « Extraction automatique d'information, du text mining au Web sémantique ». Edition Lavoisier, 2003.

[Per&al93] Pereira, F., Tishby, N. and Lee, L. « Distributional clustering of English words ». In *Proceedings of the 31th Meeting of the Association for Computational Linguistics*, pages 183-190, 1993.

[Spi02] Spinat E. « Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ? » *Colloque Cartographie de l'information : De la visualisation à la prise de décision dans la veille et le management de la connaissance*, Paris, 2002.

[Sal&Buc88] Salton G. and Buckley C. «Term weighting approaches in automatic text retrieval. Information Processing and Management », 24(5) pages 513-523, 1988.

[Sal&al75] Salton G., Yang C.S. and YU C.T. « A theory of term importance in automatic text analysis », *Journal of the American Society of Information Science,* 1975.

[Tan&Iwa96] Tanaka K.and Iwasaki H. « Extraction of lexical translations from non-aligned corpora ». In *Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

[Tur00] Turenne N. « Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles ». Thèse de doctorat, Université Louis Pasteur, Strasbourg, 2000.