Indexation de co-occurrences guidée par la structure des documents et contrôlée par une ontologie et l'exploitation du corpus

Pierre Pompidor, Boris Carbonneill, Michel Sala

Indexation de co-occurrences guidée par la structure des documents et contrôlée par une ontologie et l'exploitation du corpus

Pierre Pompidor * — Boris Carbonneill ** — Michel Sala *

(*) LIRMM, UMR Université Montpellier II – CNRS, 34090 Montpellier
(**) Société C6, Cap Oméga, 34070 Montpellier

RÉSUMÉ. Confronté à la problématique de l'indexation de très grands corpus documentaires d'entreprises, nous avons mis au point une méthode simple mais efficace (en terme de temps de calcul et de volumétrie), permettant de filtrer par document les co-occurrences les plus représentatives de ceux-ci. Nous nous plaçons dans un contexte de co-occurrences pour deux raisons. D'une part les requêtes portant sur des corpus spécialisés, et composées par des experts, s'appuient sur peu de termes précisément choisis dont nous indexons les associations, et d'autre part car cela facilitera la construction de cartes sémantiques de navigation dans les concepts du corpus. L'axe principal de ce travail est la prise en compte de la structure des documents en validant les contenus des paragraphes par ceux de leurs titres. Notre méthode s'appuie sur des mesures tf.idf successives effectuées dans le contexte d'un document et non d'un corpus, sur les contenus des paragraphes auxquels sont intégrés progressivement la hiérarchie des titres les introduisant. Puis nous exploitons simultanément une ontologie de contrôle et les requêtes des utilisateurs comportant les termes précédemment discriminés pour valider par le théorème de Bayes, les associations sémantiques ainsi déterminées.

ABSTRACT. This paper addresses the problem of indexing very large enterprise corpuses. We have designed a simple yet efficient (especially in terms of computation time and the size of the generated results) method allowing to filter, on a per-document basis, the most representative co-occurrences of the documents. The reason for using co-occurrences is twofold. First, queries composed by experts on specialized corpuses rely statistically on few, carefully chosen terms, for which we index the associations. Second, such co-occurences facilitate the construction of semantic maps used to navigate the concepts of the corpus (this part is not described in this article). Our main approach is to take into account the structure of the documents by validating the content of the paragraphs by their titles. Our method starts with successive tf.idf measures of paragraph contents taken in the context of a document (and not of a corpus), to which we progressively integrate the hierarchy of their introducing titles. We then simultaneously exploit a control ontology and the user queries containing the terms that we discriminated in the first step in order to validate, using Bayes' theorem, the semantic associations contained in a paragraph given the terms of its title.

MOTS-CLÉS : Indexation incrémentale et rapide de très grands corpus, Exploitation de la structure des documents, Contexte de co-occurrences, Théorème de Bayes

KEYWORDS : Very large and fast corpus indexing, co-occurrences' context, Bayes' theorem

1. Introduction

Cet article se place dans le champ de l'indexation de grands corpus spécialisés (comprenant plusieurs dizaines de milliers de documents). En effet, nous travaillons en collaboration avec la société C6 qui propose des solutions de gestion électronique de documents, notamment pour la constitution de dossiers d'autorisation de mise sur marché de médicaments. Or ni la recherche « full text » qui génère énormément de bruit, ni l'exploitation de méta-données trop formelles (et rarement présentes), ne satisfont les utilisateurs (experts) de ces corpus. Par ailleurs le « coût » d'une indexation classique (tant en terme de temps que de volumétrie) est trop élevé pour des entreprises dont les besoins ne sont pas prioritairement ciblés sur la recherche d'information. Nous avons donc choisi deux orientations permettant de réduire les coûts d'indexation, en terme de volumétrie et en temps de calcul, (en sachant bien entendu que l'ordre d'indexation sera très important) :

- seul un nombre très limité d'associations sémantiques est retenu par document (nous préférons le terme d'association sémantique à celui de co-occurrences qui lui peut recouvrir des collocations ne faisant pas sens) ;
- l'insertion d'un nouveau document dans le corpus ne nécessite pas la ré-indexation des documents déjà indexés.

Pour arriver à ces fins, nous proposons un nouveau processus d'indexation d'associations sémantiques entre termes, assujetti à trois niveaux de contrôles successifs :

- les lemmes candidats aux associations sont ceux qui discriminent le mieux les paragraphes du document par rapport au document lui-même, et non pas ceux qui classiquement discrimineraient le mieux les documents par rapport au corpus, cette discrimination s'appuyant sur la structure du document par l'intégration progressive du contenu de la hiérarchie de titres aux contenus des paragraphes ;
- les associations sémantiques candidates sont basées sur les lemmes discriminés lors de la première étape, puis pondérées suivant une ontologie du domaine (ou à défaut un thésaurus généraliste), et les requêtes des utilisateurs du système de gestion documentaire ;
- enfin les associations candidates sont progressivement filtrées par rapport au contenu du titre introduisant la section du document dans laquelle elles apparaissent, et cela grâce au théorème de Bayes qui les contextualise.

L'indexation étant purement incrémentale, nous ne pouvons prétendre à des résultats aussi bons que ceux qui seraient donnés par une indexation globale. La cohérence globale de l'indexation d'un document dans le corpus est assurée indirectement par les requêtes des utilisateurs, requêtes elles-mêmes analysées par rapport à la proximité de leurs termes dans une ontologie support.

Beaucoup de travaux ont porté ces dernières années sur les indexations supportés par des réseaux bayésiens, s'appuyant sur la structure de documents semi-structurés (principalement formatés en XML), et principalement dans des buts de

classification [Piwowarski, 2002] [Denoyer, 2004] [Denoyer et al., 2004] [Bratko et al., 2004].

Par contre, aucun travail n'a porté simultanément sur des très grands corpus de documents textuels, (la société C6 traite des corpus contenant jusqu'à un million de pages, une taille commune étant de 250000 pages), nécessitant une indexation purement incrémentale. Nous menons cette indexation en utilisant l'heuristique de l'analyse du contenu des titres pour associer un contexte sémantique à une section du document.

Par la suite, pour illustrer notre propos et en se focalisant sur un corpus recélant un grand nombre de documents centrés sur la toxicité des champignons, nous prendrons comme exemple le fragment de texte ci-après. Les termes en italique sont ceux inconnus de l'ontologie support qui sera présentée plus loin.

La toxicité des champignons

La toxicité des champignons ingérés est d'autant plus grave que l'apparition des symptômes est tardive. Ces symptômes sont le fait de deux syndromes principaux : les syndromes *phalloïdien* et *orellanien*.

Le syndrome *phalloïdien* et le syndrome *orellanien* des champignons à lamelles

Les amanites

La reconnaissance des amanites passe principalement par l'identification d'un anneau et d'une volve. Les principales espèces responsables du syndrome *phalloïdien* sont *amanita phalloides*, *amanita verna* et *amanita virosa*.

Les cortinaires

Les cortinaires sont caractérisés par la présence fréquente d'une *cortine*. C'est un genre qui comprend une très grande variété d'espèces dont certaines sont comestibles, d'autres d'une grande toxicité.

Comme dans la très grande majorité des documents des corpus manipulés par la solution de gestion documentaire de la société C6, cet exemple fait apparaître des nouveaux termes inconnus dans l'ontologie et qui devront y être intégrés (ici des noms latins de champignons et de syndromes, souvent des molécules). Intuitivement en lisant ce texte, nous percevons que les mots essentiels du texte sont ceux relatifs au thème général du texte : *champignon*, *toxicité* ; aux noms des syndromes toxiques *syndrome phalloïdien* et *orellanien* ; aux noms vernaculaires des espèces citées (*amanite*, *cortinaire*) ; à l'anatomie des champignons : *lamelle*, *anneau*, *volve* et *cortine* ; et à un moindre degré :aux noms latins des espèces (*amanita phalloides…)* (ces noms seraient pertinents si le paragraphe relatif aux cortinaires en contenait).

Les termes non pertinents sont ceux relatifs :à la classification des champignons (*espèce*, *genre* et *variété*), et plus encore, ceux structurant leurs descriptions (*reconnaissance*, *identification* et *présence*). Le but de notre méthode est de pouvoir indexer les trois paragraphes de ce texte par des associations pertinentes (comme *champignon – toxicité* ou *anneau – volve (sachant/dans le contexte des amanite)*) en ignorant celles qui ne le sont pas (comme *reconnaissance – identification*).

2. Contextes de co-occurrences et associations sémantiques

Notre processus d'indexation s'intéresse moins aux termes qu'aux associations sémantiques entre termes. Déjà en effet, sur un moteur de recherche généraliste, ce sont les requêtes comprenant deux mots qui sont les plus nombreuses, (mots par requêtes sur AOL.com en août 2006 : 1 mot : 27,5% ; 2 mots : 29% ; 3 mots : .18,7% ; 4 mots : 11,1% ; …), et ce phénomène s'accentue sur les moteurs de recherches des bases documentaires spécialisées utilisées par des experts exprimant des requêtes précises. Les moteurs de recherche généralistes privilégient la confiance (par exemple mesurée par le « pagerank ») à la fréquence des occurrences des termes recherchés. Bien entendu, dans une base de données documentaire spécialisée, cette mesure de confiance n'est plus pertinente. Par ailleurs, les moteurs de recherches adaptables à des corpus spécifiques comme Apache Lucene [LUCENE] se focalisent sur des calculs des fréquences optimisés par différentes variantes de la formule de discrimination *tf.idf* (décrite plus avant). Or simplement présenter à l'utilisateur les documents qui maximisent la fréquence de deux termes sans tenir compte de leurs co-occurrences soulève les problèmes suivants :

- la fréquence de l'un peut-être disproportionnée à l'autre, ou leurs présences peuvent être dissociées dans deux parties distinctes du document (et donc ces termes ne sont pas réellement co-occurrents) ;
- et plus globalement n'étayent pas la construction de cartes sémantiques en créant des liens sémantiques duaux.

Depuis quelques temps déjà, des travaux sont menés sur le calcul des fréquences de co-occurrences et de la définition des contextes sous-jacents (documentaire, positionnel ou syntaxique), et améliorés en considérant les dépendances syntaxiques entre les unités linguistiques [Besançon, 2002]. Nous nous plaçons dans un contexte documentaire, la phrase étant notre unité de base pour la génération des associations candidates qui sont ensuite contextualisées sur les lemmes des titres introduisant les paragraphes dans lesquels se situent ces phrases (en utilisant le théorème de Bayes).

3. Objectifs et étapes du processus

Nos objectifs globaux sont :

- de discriminer les lemmes par la mesure *tf.idf* appliquée dans le cadre de chaque document en utilisant la structure de celui-ci, (hiérarchie des titres organisant les sections du document, une section étant une sous-arborescence de paragraphes) ;
- d'indexer les paragraphes par un nombre réduit d'associations sémantiques composées d'au moins un lemme discriminant, en éliminant les associations sémantiques vides ou pauvres de sens par l'exploitation d'une ontologie et des requêtes des utilisateurs-experts ;
- et dans le futur, d'insérer de nouveaux termes dans l'ontologie initiale pour la spécialiser, voire de créer des versions propres à chaque corpus.

Il faut donc retenir que les corpus manipulés étant particulièrement volumineux, et la recherche d'associations sémantiques augmentant significativement la complexité des calculs, la sélection d'association sémantique est opérée document

par document, le maintien de la cohérence globale étant dévolue à l'ontologie révisée par l'exploitation effectuée par les utilisateurs sur le corpus.

Les **problèmes de volumétrie** sont cruciaux dans des corpus de centaines de milliers de pages, chaque document atteignant fréquemment les 3000 mots, la suppression des mots creux en éliminant à peu près les trois quarts. En moyenne, chaque document comprend une cinquantaine de paragraphes organisant chacun une quinzaine de phrases, chaque phrase conservant en moyenne 5 termes après la phase de pré-traitement. La conservation globale de toutes les co-occurrences possibles reviendrait donc (toujours en moyenne) à conserver 10 co-occurrences par phrase, soit 150 co-occurrences par paragraphe et donc 7500 co-occurrences par documents ce qui en nombre de termes multiplierait par 5 son volume !

Notre but est de ne conserver en moyenne que *n* (par défaut 3), associations sémantiques pertinentes par paragraphe, en ramenant le poids maximal de l'indexation d'un document à 150 associations sémantiques. Le fonctionnement général du processus d'indexation est illustré par l'algorithme suivant :

* Pour chaque document du corpus :
  Phase d'amorce par indexation simple en utilisant la structure du document :
  Pré-traitement (lemmatisation et suppression des mots creux) (voir section **4.1**)
  Comptage de chaque lemme dans tous les paragraphes
  * Pour chaque paragraphe :
    * Pour tous les lemmes du paragraphe (simples ou agrégés)
      - **Calculs successifs du tf.idf** attribuant un poids à ce lemme **en intégrant progressivement** dans la section le contenant les lemmes des titres introductifs (voir section **4.2**) (*une section est une sous-arborescence de paragraphes, initialement chaque paragraphe feuille forme une section)*
      - **Conservation du maximum de cette mesure** et transformation de celle-ci en une probabilité de « pertinence » (*Ppertinence(lemme))*
      - **Mémorisation des lemmes « probablement pertinents »**, **génération des associations sémantiques** ayant au moins un de ceux-ci (sect. **4.3**)

  Enregistrement des requêtes ciblant ce document dans un entrepôt de données.
  A partir du moment où un nombre suffisant de requêtes concernant le document couvre tous les lemmes probablement pertinents indexés :
  * Pour chaque paragraphe :
    * Pour chaque phrase du paragraphe :
      * Pour toutes les associations sémantiques pré-sélectionnées :
        - **Calcul de la probabilité de pertinence par rapport à l'ontologie support et aux requêtes des utilisateurs** : *Ppertinence(association)* (l'ontologie est préalablement décomposée en composantes) (s. **4.4**)
        - **Mesure de la probabilité de pertinence des lemmes par rapport à cette association** :*Ppertinence(lemme|association)* (section **4.5)**
        * Pour tous les lemmes du titre introduisant le paragraphe :
          Calcul de la probabilité de pertinence de l'association par rapport aux lemmes du titre introduisant le paragraphe par **le théorème de Bayes** (section **4.6**) : *P(association|lemme) = P(lemme|association).P(association) / P(lemme)*
    Indexation du paragraphe par les *n* premières associations

4. Détail des étapes

4.1 *Prétraitement (lemmatisation et élimination des mots creux)*

Nous procédons à une phase classique de lemmatisation des termes, puis à l'élimination des mots creux suivant les critères suivants :

- sont conservés : les substantifs, les groupes nominaux composés d'un nom et d'un adjectif, et ceux composés d'un nom suivi d'un terme inconnu (cette heuristique permet la conservation nombre de collocations pertinentes dans les documents scientifiques, ici *syndrome principal*, *phalloidien*), et tous les groupes de mots inconnus et contigüs (*amanita phalloides*, *amanita verna* et *amanita virosa*) ;
- tous les autres termes sont supprimés.

Nous paramétrons TreeTagger [TreeTagger] pour améliorer ce pré-traitement sur des syntagmes plus complexes. Sur notre exemple, les syntagmes *champignon à lamelles* ou *espèce de champignon* seront prochainement considérés globalement.

4.2 *Calculs successifs du tf.idf attribuant un poids à chaque lemme du document dans une section, et corrélation d'une probabilité de pertinence*

Le texte illustratif comprend plusieurs sections de texte que nous voulons caractériser par des associations sémantiques. Ces sections sont formées initialement par les paragraphes et ne comprennent pas les titres ou sous-titres qui les introduisent. Nous allons opérer cette intégration progressivement.

Après une première mesure *tf.idf* est calculée pour tous les lemmes de chaque paragraphe, une seconde mesure est opérée en absorbant le contenu des titres les plus proches de ces paragraphes qui deviennent des sections, et cela ainsi de suite jusqu'à ce que le titre de plus haut niveau soit absorbé par la section qui lui était la plus proche. Ainsi le paragraphe 2 :

> La reconnaissance des amanites passe principalement par l'identification d'un anneau et d'une volve. Les principales espèces responsables du syndrome phalloïdien sont amanita phalloides, amanita verna et amanita virosa.

est ensuite étendu aux deux sections suivantes :

> Les amanites :
>
> La reconnaissance des amanites passe principalement par l'identification d'un anneau et d'une volve. Les principales espèces responsables du syndrome phalloïdien sont amanita phalloides, amanita verna et amanita virosa.

Puis :

> Le syndrome phalloïdien et le syndrome orellanien des champignons à lamelles :
>
> Les amanites :
>
> La reconnaissance des amanites passe …

Pour identifier les lemmes discriminant le plus fortement les sections par rapport au document, nous utilisons la fonction **Term Frequency Inverse Document Frequency** (*tf.idf*). Issue du monde de la recherche d'information [Salton et al., 1988], elle donne :

– plus de poids aux mots apparaissant souvent au sein d'un même document (ici dans une section, ces mots sont plus représentatifs de celle-ci) ;
– moins de poids aux mots appartenant à plusieurs documents (ici sections) en reflétant le fait que ces mots ont un faible pouvoir de discrimination.

En opérant cette mesure sur des paragraphes qui absorbent progressivement leurs titres, nous détectons les sections (paragraphes+sous-hiérarchie de titres) les mieux discriminées par les lemmes du document. Cette mesure n'est pas une probabilité. Pour calculer la probabilité qu'un lemme soit représentatif d'une section (paragraphe associé à un ou plusieurs de ses titres), l'intervalle entre les mesures *tf.idf* la plus basse et la plus haute, est corrélé à un intervalle de probabilité entre 0,25 et 0,75.

Le tableau suivant (*Figure 1*) présente :

- la liste des lemmes, leurs nombres d'occurrences dans les trois sections et le nombre de sections dans lesquelles ces lemmes apparaissent ;
- les mesures *tf.idf* appliquées au contenu des paragraphes (feuilles, avec titre immédiat, les deux titres les plus proches) et les probabilités corrélées (Px).

Liste des lemmes	Nbr d'occur-rences dans les sections	Nbr de sections ds lesquelles le lemme apparaît (S)	tf.idf niveau 0 N/25 .log(3/S) T0	tf.idf niveau 1 N/29 .log(3/S) T1	tf.idf niveau 2 N/37 .log(3/S) T2	P0 (avec T0)	P1 (avec T1)	P2 (avec T2)
Toxicité	2 - 3 - 3	2 - 2 - 2	0.014	**0,018**	0,014	0,25	**0,393**	
champignon	1 - 2 - 4*****	1 - 1 - 3	0.019	**0,033**	0	0,354	**0,75**	
apparition	1	1	**0.019**	0.016	0.013	**0,354**		
symptôme	2	1	**0.038**	0,033	0,026	**0,75**		
deux	*1*	*1*	*0.019*	*0,016*	*0,013*	*0,35*		
syndrome pr.	1	1	**0,019**	0,016	0,013	*4*		
syndrome ph.	2 - 2 - 4*****	2 - 2 - 3	0.014	0,012	0	**0,354**		
syndrome or.	1 - 1 - 3*****	1 - 1 - 3	**0,019**	0,016	0	**0,25**		
lamelle	0 - 0 - 2*****	0 - 0 - 2			**0,0095**	**0,354**		**0,433**
amanite	1 - 2 - 2	1	0.019	**0,033**	0,026		**0,75**	
reconnaiss.	1	1	**0.019**	0,016	0,013	0,354		
identification	1	1	**0.019**	0,016	0,013	**0,354**		
anneau	1	1	**0.019**	0,016	0,013	**0,354**		
volve	1	1	**0.019**	0,016	0,013	**0,354**		
espèce	2	2	**0.014**	0,012	0,0095	**0,354**		
amanita ph.	1	1	**0.019**	0,016	0,013	**0,25**		
amanita ve.	1	1	**0.019**	0,016	0,013	**0,354**		
amanita vi.	1	1	**0.019**	0,016	0,013	**0,354**		
cortinaire	1 - 2 - 2	1	0.019	**0,033**	0,026	**0,354**	**0,75**	
présence fr.	1	1	**0.019**	0,016	0,013	0,354		
cortine	1	1	**0.019**	0,016	0,013	**0,354**		
genre	1	1	**0.019**	0,016	0,013	**0,354**		
variété	1	1	**0,019**	0,016	0,013	**0,354**		
						0,354		

Moyenne des probabilités de pertinence : 0.422 ; (*) ces lemmes sont distribués sur les deux sous-sections
Figure 1. Par *lemme, nombres d'occurrences et mesures tf.idf en absorbant les titres les plus proches*

Les lemmes les plus discriminants, et probablement pertinents par rapport aux sections car d'une probabilité supérieure à 0,5, sont : **champignon**, **symptôme**, **amanite** et **cortinaire**. Ils seront donc à la base de l'indexation par les associations

sémantiques initiales car seules les associations comprenant au moins un de ces lemmes seront retenues par un premier filtre.

4.3 *Première indexation des paragraphes sur les termes lemmatisés*

Voici pour la première phrase de chaque paragraphe les lemmes probablement pertinents et les associations sémantiques candidates qui les comprennent :

- La toxicité des **champignons** ingérés est d'autant plus grave que l'apparition des **symptômes** est tardive : *champignon – toxicité ; champignon – apparition ; champignon – symptôme ; symptôme – toxicité ; symptôme – apparition*

- La reconnaissance des **amanites** passe par l'identification d'un anneau et d'une volve : *amanite – reconnaissance ; amanite – identification ; amanite – anneau ; amanite – volve*

- Les **cortinaires** sont caractérisés par la présence fréquente d'une cortine : *cortinaire – présence-fréquente ; cortinaire - cortine*

4.4 Calcul de la probabilité de pertinence d'une association sémantique par rapport à l'ontologie et aux requêtes des utilisateurs

Une probabilité de pertinence va être calculée pour chaque association sémantique en se basant sur une **ontologie support** (à défaut un thésaurus), et sur un **premier contingent de requêtes** effectuées par les experts.

Ces deux outils sont complémentaires, l'ontologie permettant de spécifier la description des champs sémantiques notamment sur les relations de synonymie, de spécialisation et d'agrégation, et les requêtes des utilisateurs-experts ciblant leurs usages fonctionnels.

4.4.1 Une ontologie pour une première probabilité de pertinence de l'association

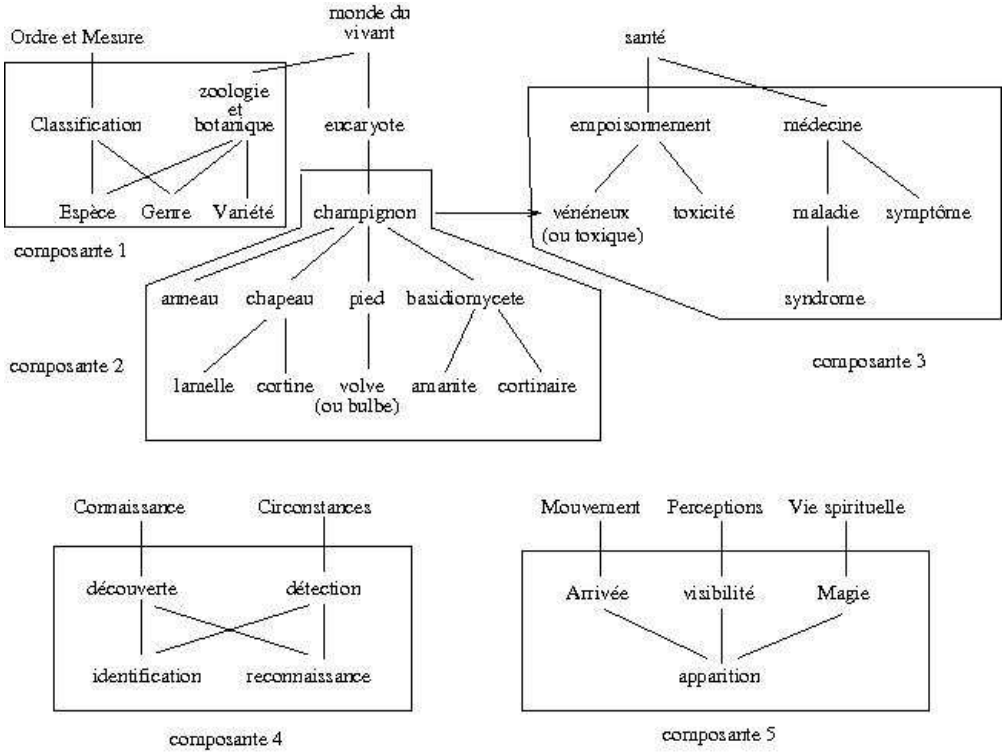
Dans notre processus, les associations sémantiques sélectionnées sont d'abord pondérées suivant un calcul de probabilité étayée par une ontologie support. Si les ontologies sont devenues indispensables à la recherche d'information en permettant de proposer aux auteurs des méta-données pour annoter les documents, d'améliorer les requêtes des utilisateurs par des mécanismes d'extension et des cheminements de recherche, nous les utilisons dans un premier temps pour valider la proximité sémantique des lemmes associés. Notre ontologie, comme souvent dans le domaine de la fouille de texte, se réduit à un « simple » thésaurus explicitant trois relations :

- la relation de synonymie (« *lépiote élevée* » synonyme de *coulemelle*) ;
- la relation d'hyperonymie (*lépiote* hyperonyme de *coulemelle*) ;
- la relation d'association de sens (*amanite* en relation avec *toxicité*) (bien que certaines amanites soient excellentes)

qui ne sont d'ailleurs pas forcément explicitées.

Cela-dit, il est rare de pouvoir disposer d'ontologies adaptées aux documents manipulés : nous ne disposons souvent que de thésaurus linguistiques trop généralistes qui ne couvrent que partiellement les champs sémantiques des corpus spécialisés (ici la phytotoxicité), ou au contraire de taxonomies trop focalisés...

Considérons les fragments de l'ontologie (*Figure 2*) dont nous disposons, et qui organisent le sens des lemmes de notre exemple. Ces fragments au nombre de cinq sont en fait des composantes du graphe sous-jacent à l'ontologie :



(*) *est appelée ici composante les sous-graphes denses comprenant des nœuds fortement connectés.*
Figure 2. *Fragments de l'ontologie relevant des champs sémantiques couverts par l'exemple*

Nous remarquerons que cette ontologie possède une superstructure qui met en relation toutes les composantes, et que si elle recense en très grande majorité des substantifs, certains adjectifs sont également présents (ici *vénéneux/toxique*). Par ailleurs, même si les ontologies peuvent être manipulées comme des graphes complets, il est naïf de pouvoir mesurer une proximité sémantique entre deux lemmes de l'ontologie en calculant simplement le nombre de pas du chemin le plus court reliant le premier lemme d'une association au second de celle-ci **en dehors d'une composante**, et cela à cause de deux raisons. D'une part, le thésaurus n'est jamais homogène (certains champs sémantiques sont toujours plus détaillés que d'autres), notamment quand il est progressivement complété par du vocabulaire spécialisé, et d'autre part les liens de superstructure sont souvent factices.

Par rapport à notre exemple, notre procédure a donc isolé cinq composantes, soit cinq sous-graphes de nœuds fortement inter-connectés :

- la composante *c1* des termes de classification : *espèce*, *genre* et *variété* ;

- la composante *c2* du terme *champignon* et de ses agrégats (*lamelle*, *anneau*, *volve*…) ou de ses spécialisations (*amanite* et *cortinaire*) ;
- la composante *c3* centrée autour du terme *toxicité* ;
- la composante *c4* centrée autour d'*identification* et de *reconnaissance* ;
- et enfin la composante *c5* centrée autour du terme *apparition.*

(Utiliser les connexions entre ces composantes est dépourvu de signification, même le lien entre les deux composantes *monde du vivant* et *santé* est peu prégnant).

Par rapport à notre document, nous pouvons déjà inventorier les lemmes qui sont potentiellement synonymes (en pouvant être employés l'un pour l'autre **dans ce contexte**). Ce sont ceux, qui au sein d'une composante ont les mêmes pères : *espèce – variété, lamelle – cortine, amanite – cortinaire ... identification – reconnaissance*.

4.4.2 Prise en compte des requêtes des utilisateurs

Les associations sémantiques associées aux paragraphes sont filtrées par rapport aux termes lemmatisés composants les requêtes des utilisateurs (*Figure 3*), dont voici les treize premières comprenant au moins un lemme « probablement » pertinent, et ayant sélectionné le document d'illustration :

amanite – identification	espèce – champignon – toxique
amanite – toxicité	identification – cortinaire – toxique
champignon – empoisonnement	symptôme – empoisonnement – champignon
champignon – syndrome	symptôme – syndrome-phalloïdien
reconnaissance – amanite	variété – champignon – lamelle
champignon – lamelle – toxique	apparition – symptôme – empoisonnement – champignon

En effet, à partir du moment où tous les lemmes probablement pertinents ont été utilisés dans les requêtes des utilisateurs, celles-ci participent aux calculs des probabilités de pertinences des associations sémantiques. Le premier prétraitement consiste à fixer les termes synonymes dans le contexte de notre document : ce sont ceux qui pré-identifiés peuvent être interchangeables dans un nombre significatif de requêtes. Sur les six couples de termes pré-identifiés, les termes *identification* et *reconnaissance* peuvent être substitués dans deux requêtes : ils sont donc confondus. Une analyse morpho-syntaxique préalable à l'exploitation de l'ontologie a aussi permis de confondre les noeuds *vénéneux-toxique* et *toxicité*.

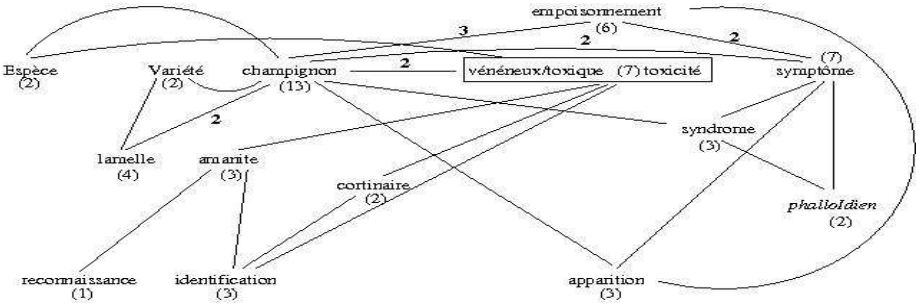


Figure 3. *Graphe des requêtes des utilisateurs (l'arité des nœud étant spécifiée)*

En partant du nœud ayant la plus forte arité (*champignon*) sont progressivement englobés tous les nœuds connexes ayant une arité supérieure ou égale à l'arité moyenne (de 3,9). Est ainsi créé un sur-nœud (*Figure 4*) qui représente la principale composante sémantique fonctionnelle liée au document et qui :

- confond les deux composantes *c2* et *c3* (*monde du vivant* et de la *santé*) ;
- confère une forte probabilité de pertinence (0,75) à toutes les associations formées par tout couple de lemmes pris dans ce sous-graphe ;
- crée une transitivité entre des lemmes joignables par son intermédiaire.

Nous remarquerons que tous les nœuds dont l'arité est supérieure à l'arité moyenne ont été absorbés. Si ce n'était pas le cas, cette opération d'identification de sur-nœuds serait réitérée sur les nœuds restants d'arité supérieure à la moyenne et permettrait d'identifier d'autres composantes sémantiques fonctionnelles.

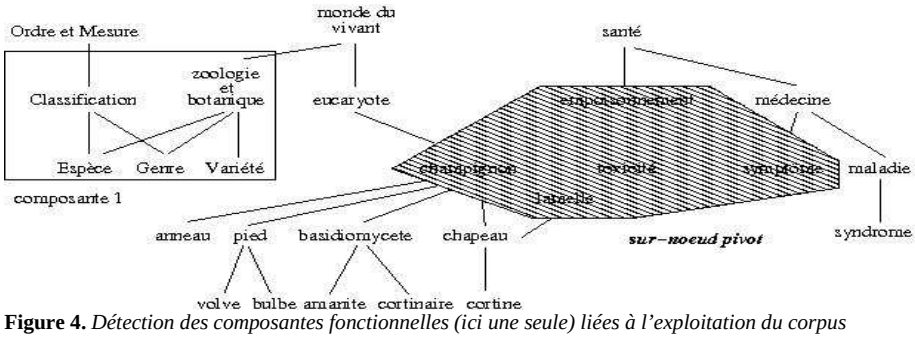


Figure 4. *Détection des composantes fonctionnelles (ici une seule) liées à l'exploitation du corpus*

Voici donc les éléments participant au calcul des probabilités de pertinence associées aux associations sémantiques :

- le nombre de pas minimal qui relie leurs termes dans une composante ;
- la probabilité de proximité sémantique entre deux termes d'une association sémantique corrélée à la distance intra-composante ;
- une maximisation à 0,75 si les deux lemmes sont associés dans la même composante sémantique fonctionnelle (post-exploitation du corpus) ;

(*) 1- (distance/(distance maximale +1)) / facteur de minoration	Distance intra-composante *entre parenthèses (distance max.)*	Probabilité de pertinence par rapport à l'ontologie (*)	Probabilité de pertinence : ontologie et requêtes
champignon – toxicité/symptôme	-	*0,2*	0,75
champignon – apparition	-	*0,2*	0,25
symptôme – toxicité	4 (5)	0,33	0,75
symptôme – apparition	-	*0,2*	0,2
amanite – reconn./ident.	-	*0,2*	0,2
amanite –anneau	3 (4)	0,4	0,4
amanite – volve	4 (4)	0,2	0,2
cortinaire – cortine	-(*cortine* inconnu)	*0,5*	0,5
cortinaire – présence-fr.	-	*0,2*	0,2
Reconnaiss. – identification	1 (2)	0,66	(synonymes)
anneau – volve	3 (4)	0,4	0,4
syndrome-phalloïdien – espèce	-	*0,2*	0,2
champignon – lamelle	2 (4)	0,6	0,75

Figure 5. Probabilités de pertinence des associations sémantiques

Cette probabilité peut être par la suite minorée car elle est vulnérable à l'insertion de vocabulaire de structure entre lemmes peu usité dans les documents (cette minoration dépend de l'homogénéité de l'ontologie). La probabilité d'association de deux lemmes inter-composantes est fixée à la plus basse des probabilités de pertinence calculée entre deux lemmes d'une même composante (ici 0,2 pour l'association entre *amanite* et *volve*)), tandis que la probabilité d'association d'un lemme connu avec un lemme inconnu est fixée à 0,5.

4.5 Association « a priori » des lemmes par rapport aux associations

Nous calculons maintenant la probabilité de pertinence « a priori » des lemmes par rapport aux associations. Ce *prior* sera utilisé dans le théorème de Bayes pour finalement filtrer les associations sémantiques par les titres des sections dans lesquels elles se trouvent. La probabilité qu'un lemme présent dans un titre soit pertinent sachant que ce lemme est associé à une ou plusieurs associations sémantiques se trouvant dans la section qu'il introduit, est calculée comme suit.
Une probabilité de 0,5 correspondrait à ce que le terme ait le même poids dans le document que dans les requêtes. Pour étager ces probabilités entre 0,25 et 0,75, une première étape consiste à rechercher le lemme le plus sur-représenté dans les requêtes par rapport au document, et celui qui l'est le moins.
Dans notre exemple, le lemme le plus sur-représenté est *champignon*. Par rapport aux seules requêtes, sa sur/sous-représentation est de $R = (7/31 / 4/37) = (0,226 / 0,108) = 2,1$. Par contre le lemme *cortinaire* étant présent deux fois dans le document et une seule fois dans les requêtes, sa représentation est $R = (1/31 / 2/37) = (0,032 / 0,054) = 0,6$, puis à une probabilité de P = 0,25 + ((R-0,6)/3), les probabilités étant étagées de 0,75 à 0,25.

Le tableau suivant présente les lemmes des titres, leur fréquence dans les requêtes et le document, et leur probabilité de pertinence par rapport aux requêtes :

lemmes présents dans les titres :	Fréquence du lemme dans les requêtes et dans le document	Probabilité de pertinence suivant les associations définies par les requêtes
Amanite < basid. < champignon	3/31 = 0,097 et 2/37 = 0,054	0.65
Champignon	7/31 = 0,226 et 4/37 = 0,108	0,75
Cortinaire < basid. < champignon	1/31 = 0,032 et 2/37 = 0,054	0,25
Lamelle	2/31 = 0,065 et 2/37 = 0,054	0,45
Syndrome et syndrome phalloïdien	2/31 = 0.065 et 4/37 = 0,027	0,25
Toxicité < empoisonnement	4/31 = 0,129 et 3/37 = 0,081	0 ,58

Figure 6. *Fréquence et probabilité de pertinence des lemmes présents dans les titres*

4.6 ***Calcul final de la probabilité de pertinence d'une association par rapport à une section*** *(P(association|lemme du titre))*

Pour contextualiser chaque association sémantique, nous appliquons le théorème de Bayes qui aura pour but de valider, via des probabilités conditionnelles, la pertinence de l'association sémantique par rapport aux lemmes des titres introduisant la section dans laquelle cette association a été retrouvée.

Si les réseaux bayésiens sont un formalisme de raisonnement probabiliste très utilisé dans la fouille de texte, nous utilisons simplement le théorème de Bayes pour discriminer les associations sémantiques suivant la structure du document. En 1992,

le système de recherche d'information INQUERY [Callan et al., 1992] introduit l'utilisation des réseaux bayésiens dans la RI. Ceux-ci auront pour but de calculer la probabilité qu'une requête soit satisfaite par un document. Mais ce modèle utilisé dans INQUERY fait encore un traitement « plat » des documents, c'est-à-dire que leur structure n'est pas prise en compte et que tous les mots sont traités de la même manière quelque soit l'endroit où ils se trouvent. Une extension de INQUERY proposée par Myaeng [Myaeng et al., 1998] prend en compte la structure des documents en plus de leurs contenus, celle-ci étant représentée par un arbre. Chaque feuille de cet arbre est prolongée par les nœuds termes contenus dans l'élément de structure représenté par cette feuille. D'autres travaux ont appliqué les réseaux bayésiens à des corpus de documents XML [Zargayouna, 2004], mais ces modèles reposent sur des documents XML qui doivent être structurés uniformément.

La probabilité finale de pertinence d'une association par rapport aux lemmes de son titre est la suivante (lemme ayant la signifiant lemme du titre) :

P(association|lemme) = P(lemme | association) . P(association) / P(lemme)

(P(lemme | association) = P (lemme | associations) * moyenne P(lemme) des associations)

Voici le tableau récapitulant les différents résultats des étapes de notre processus :

Association sémantique candidate	Lemme filtre du titre	P (lemme | ass)	P (asso.)	P (lemme)	P (asso. | lemme)
champ. – toxicité	toxicité	0.58 * 0,422 = 0,245	0,75	0,393	**0,468**
champ. – apparition	toxicité	0.58 * 0,422 = 0,245	0,2	0,393	0,125
champ. – symptôme	toxicité	0.58 * 0,422 = 0,245	0,75	0,393	0,468
champ. – toxicité	champignon	0,75 * 0,422 = 0,317	0,75	0,75	**0,317**
champ. – apparition	champignon	0,75 * 0,422 = 0,317	0,2	0,75	0,085
champ. – symptôme	champignon	0,75 * 0,422 = 0,317	0,75	0,75	0,317
symptôme – toxicité	toxicité	0,58 * 0,422 = 0,245	0,75	0,393	**0,468**
symptôme – apparition	toxicité	0,58 * 0,422 = 0,245	0,2	0,393	0,125
symptôme – toxicité	champignon	0,75 * 0,422 = 0,317	0,75	0,75	0,317
symptôme – apparition	champignon	0,75 * 0,422 = 0,317	0,2	0,75	0,085
amanite – recon./identif.	amanite	0,65 * 0,422 = 0,274	0,2	0,75	0,073
amanite –anneau	amanite	0,65 * 0,422 = 0,274	0,4	0,75	0,147
amanite – volve	amanite	0,65 * 0,422 = 0,274	0,2	0,75	0,073
cortinaire – cortine	cortinaire	0,25 * 0,422 = 0,106	0,5	0,75	**0,071**
cortinaire - présence	cortinaire	0,25 * 0,422 = 0,106	0,2	0,75	0,028
reconnaiss. – identific.	amanite	0,65 * 0,422 = 0,274	- : synon.	0,75	-
anneau – volve	amanite	0,65 * 0,422 = 0,274	0,4	0,75	**0,147**
syndrome-ph. – espèce	amanite	0,25 * 0,422 = 0,106	0,2	0,75	**0,028**
champignon – lamelle	toxicité	0,58 * 0,422 = 0,245	0,75	0,393	**0,468**
champignon – lamelle	champignon	0,75 * 0,422 = 0,317	0,75	0,75	**0,317**

Figure 7. *Récapitulatif des probabilités de pertinence des associations sémantiques*

4.7 *Bilan*

Dressons un comparatif des calculs de probabilité effectués sur l'exemple (où les associations n'emploient que les lemmes des premières phrases de chaque paragraphe), et sur le corpus complet lié à la toxicité des champignons. Ce qui est indexé par les experts (lemmes ou associations) est en gras dans les tableaux :

Classement des lemmes par comptage de fréquence	Classement des lemmes par mesures *tf.idf* maximisées	Classement des associations par rapport à l'ontologie et aux requêtes	Classement des associations suivant le théorème de Bayes
champignon 3 **syndrome ph. 3** **toxicité** 3 symptôme 2 **amanite** 2 **cortinaire** 2 syndrome or. 2 amanite ph. 1 amanita ve. 1 amanita vi. **anneau** 1 apparition 1 **volve** 1 **cortine** 1 espèce 1 genre 1 identification 1 **lamelle** 1 présence fréqu. 1 reconnaissance1 syndrome pr. 1 variété 1	0.75 : **champignon** symptôme **amanite** **cortinaire** 0,435 : --------- **lamelle** 0,393 : **toxicité** 0,354 : --------- apparition syndrome prin. syndrome orel. reconnaissance identification **anneau** **volve** amanita … présence **cortine** 0,25 : --------- **syndrome ph.** Espèce	0.75 : **champignon – toxicité** **champignon – symptôme** **symptôme – toxicité** **champignon – lamelle** 0.5?: **cortinaire–cortine** 0.4 : **anneau – volve** amanite – anneau 0.2 : amanite – volve champignon – apparition symptôme – apparition cortinaire – présence-fr. **syndrome-ph.–espèce** 0 : reconn. – identif.	0.468 (sachant toxicité) : **champignon – toxicité** **champignon – symptôme** **symptôme – toxicité** **champignon – lamelle** 0.317 (sachant champignon) : *mêmes que précédemment* 0,147 (sachant amanite) : **anneau – volve** amanite – anneau 0,125 (sachant toxicité) : champignon – apparition symptôme – apparition 0,085 (sachant champignon) : champignon – apparition symptôme - apparition 0,073 (sachant amanite) : amanite – reconn./identific./volve 0,071 : **cortinaire–cortine** (cortin.) 0,028 : cortinaire–présence (cortin.) **syndrome-ph.–espèce** (amanite) 0 : reconnaissance–identification

Figure 8. *Classement des lemmes et des associations suivant les probabilités calculées*

Voici les associations sémantiques faites par les experts pour chaque section :

La toxicité des champignons :
La toxicité des champignons ingérés est d'autant plus grave que l'apparition des symptômes est tardive. Ces symptômes sont le fait de deux syndromes principaux : les syndromes phalloïdien et …
champignon – toxicité
champignon – symptôme sachant toxicité
symptôme – toxicité sachant champignon

Le syndrome phalloïdien et le syndrome orellanien de certains champignons à lamelles :
Les amanites :
La reconnaissance des amanites passe par l'identification d'un anneau et d'un volve. Les principales espèces responsables du syndrome phalloïdien sont amanita phalloides, …
champignon – lamelle sachant toxicité
anneau – volve sachant amanite
syndrome-phalloïdien – espèce sachant amanite

Le syndrome phalloïdien et le syndrome orellanien de certains champignons à lamelles :
Les cortinaires
Les cortinaires sont caractérisés par la présence fréquente d'une cortine. C'est un genre qui comprend une très grande variété d'espèces dont certaines sont comestibles, d'autres d'une grande toxicité.
champignon – lamelle sachant toxicité
cortinaire - cortine

A cette annotation comparons les taux d'adéquation de nos classements :

- par **comptage** : discrimination de 7 lemmes parmi les 9 pré-sélectionnés dont 5 sont parmi les 7 distingués par le classement :(5/7)*(7/9) = **55,5%**.
- par des **mesures *tf.idf* maximisées** : discrimination de 6 lemmes parmi les 9 pré-sélectionnés dont 5 sont parmi les 6 distingués : (5/6) * (7/9) = **65%**.
- par l'exploitation **de l'ontologie et des requêtes** : ce classement discrimine 7 associations sur 7 pré-selectionnées dont 6 distinguées : (6/7) * (7/7) = **86%** qui est le meilleur taux brut d'adéquation.
- par le **théorème de Bayes** : discrimination de 6 associations sur 7 pré-sélectionnées dont 5 sont parmi les 6 discriminées : (5/6) * (6/7) = **72%**.

Sur le corpus complet, les taux d'adéquation sont respectivement de 51% (comptages simples), 63% (mesures tf.idf maximisées), 83% (exploitation de l'ontologie et d'une indexation manuelle via les requêtes), et 70% (par Bayes).

Nous remarquons que si le résultat donné par l'application du théorème de Bayes est donc moins bon que celui obtenu par l'établissement d'associations sémantiques prenant en compte seulement l'ontologie et l'exploitation du corpus, il permet d'enrichir l'annotation du document **en précisant le contexte de celle-ci** (grâce à la probabilité conditionnelle), et de définir des cartes sémantiques associées au corpus.

Pour systématiser l'évaluation des annotations générées, nous sommes en train d'automatiser la mise en correspondance, par champ sémantique, avec les ontologies extraites de WordNet en appliquant à nos annotations les mesures de similarité sémantique suivant les méthodes proposées par Resnik [Resnik 1995].

En termes de temps de calcul et volumétrie, si le moteur de recherche proposé par la base documentaire utilisé par le logiciel de gestion documentaire de la société C6 (Documentum) est performant, il ne réalise pas une indexation contextuelle contrairement à notre système, qui génère des associations sémantiques jugées pertinentes sans faux positifs par les experts (sur le premier cas de mise en œuvre).

5. Conclusion

Le but de notre travail est de pouvoir offrir à un système documentaire cohérent, une indexation rapide (et donc purement incrémentale), générant une volumétrie d'annotation sémantique raisonnable. Nous avons donc choisi de mettre en œuvre la règle de Bayes pour filtrer les associations sémantiques les plus pertinentes d'une section du document, (une section étant un sous-arbre quelconque de paragraphes), par rapport aux termes du titre l'introduisant. Cette heuristique efficiente en termes de calcul, donne de bons résultats qualitatifs, reconnus par les utilisateurs des corpus, et apporte deux avantages liés à la contextualisation des annotations : la création facilitée de cartes sémantiques et la révision topologique des ontologies sous-jacentes à l'indexation. Cela dit, deux problèmes peuvent être soulevés.

Une première critique tout à fait naturelle de ce travail, est l'effet délétère que pourrait causer des mots de structure par nature vides (comme « introduction », « conclusion »). En fait ce problème est rarement fondé car si ces termes ne se retrouvent que dans les titres et non dans les sections qu'ils introduisent, ils n'apparaîtront pas dans les associations sémantiques retenues pour indexer les paragraphes des documents. Au contraire un mot-concept apparaissant dans un titre va être également sur-employé dans les sections où il apparaît, et donc discriminé.

La seconde critique plus sérieuse, est l'impact de certains biais. Le premier biais est dû au fait que l'analyse des requêtes composées par les utilisateurs et qui modifie la probabilité de pertinence d'une association sémantique par rapport aux lemmes qu'elle contient, n'est faite qu'à partir du moment où un nombre critique de requêtes se trouve dans l'entrepôt de données. Dans l'exemple, cette analyse n'a été déclenchée qu'à partir du moment où tous les lemmes contenus dans les associations sémantiques filtrées par la première étape du processus ont été retrouvés dans les requêtes des utilisateurs. Cette analyse étant amorcée trop tôt, les derniers lemmes intégrés étant défavorisés, son seuil de déclenchement doit donc être affiné.

Le second biais, corrélé au premier, est que l'exploitation d'un corpus n'est ni homogène dans le temps, ni dans le vocabulaire utilisé, certains descripteurs étant plus utilisés que d'autres, par habitude ou par notoriété. Ainsi, peu d'utilisateurs ont interrogé notre corpus avec le lemme « cortinaire », le nom de champignon le plus fréquemment utilisé comme point d'entrée dans la recherche d'information sur leur toxicité, étant « amanite », *même si les experts étaient également à la recherche d'information sur la toxicité des cortinaires*. Cette sous-représentation conduit à la dépréciation des associations fondées sur ce lemme, et doit être corrigée.

6. Bibliographie

Besançon R., Rajman M., « Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents », TALN 2002

Bratko A., Filipic B., « Exploiting Structural Information in Semi-structured document classification », Andrej P., 13th Int. Electrotechnical and Computer Science Conf., ERK'2004

Callan J. P., Croft W. B., and Harding S. M., "The INQUERY Retrieval System", In Tjoa A. M. and Ramos I. editors, *Database and Expert Systems Applications, Proceedings of the International Conference*, pages 78–83, Valencia, Spain. Springer-Verlag, 1992

Denoyer L., "Apprentissage et inférence statistique dans les bases de documents structurés : application aux corpus de documents textuels", thèse de doctorat de l'Un. Paris VI, 12/2004

Denoyer L., Gallinari P., « Bayesian network model for semi-structured document classification», Inf. Processing Management 40(5):807-827, 2004

Myaeng S., Jang D., Kim M., and Zhoo Z., « A Flexible Model for Retrieval of SGML documents », Proc 21st ACM SIGIR, 138-140, Melbourne, ACM Press, New York, 1998

Piwowarski, Gallinari P., « A bayesian network model for page retrieval model in a hierarchical structured collection », XML w. of the 15th ACLM SIGIR Conf., Tampere, Finland 2002

Resnik Ph,«Using Information Content to Evaluate Semantic Similarity in a Taxonony», IJCAI 95

Salton G., Buckley C., "Term-weighting Approaches in Automatic Text Retrieval", *Information Processing and Management* , 24(5), pp. 513-523, 1988.

Zargayouna H., «Contexte et sémantique pour une indexation de doc. semi-structurés» CORIA'04

LUCENE : http://fr.wikipedia.org/wiki/Lucene

TreeTagger : http://ims.uni-stuttgart.de/projekte/corplex/TreeTagger

DOCUMENTUM : http://en.wikipedia.org/wiki/Documentum