Une approche de catégorisation structurelle de documents numériques pour une meilleure exploitation du patrimoine juridique décisionnel

Jin YAO (1), Jacques MADELAINE (1), Khaldoun ZREIK (2)

(1) DoDoLa - GREYC, Université de Caen, France
(2) CITU - Paragraphe, Université de Paris 8, France

Mots-clés : catégorisation de documents semi-structurés, extraction de connaissance, recherche d'information, patrimoine juridique décisionnel

Keywords: semi-structured document clustering, knowledge discovery, information retrieval, decision support for legal heritage

Résumé :

Le patrimoine de document juridique (loi, jurisprudence, brevet) s'est bien approprié l'univers de numérisation pour permettre une diffusion et une exploitation accrues des informations juridiques par des applications diverses. En conséquence, l'usage des bases documentaires juridiques partageables est devenu de plus en plus ouvert et fréquent favorisant ainsi un débit d'alimentation « semi-automatique » assez important. Constat 1 : par semi-automatique, on entend un processus de dépôt direct des documents dans des bases contrôlées par des SGBDs qui exigent une intervention humaine réduite surtout au niveau de l'indexation et de la classification. En effet, ce sont les modèles de documents (leurs structures logiques et physiques modélisées par le langage de balisage) qui assurent un rôle important dans les processus d'indexation et de gestion. Donc ces modèles incorporent indirectement connaissance et savoir-faire. Constat 2 : devant une telle masse de données « très souvent textuelles », il devient indispensable d'adopter aussi une approche pour gérer les documents électroniques juridiques en tant que supports de connaissance et de savoir faire. Ceci nous mène vers des problématiques de recherche d'information et d'extraction de connaissance. Ces deux constats nous conduisent à formuler une hypothèse de classification automatique qui tiendra compte de connaissance et de savoir-faire incorporés dans les structures des modèles de documents électroniques juridiques. Aussi on constate que ces connaissances ou savoir-faire ne sont pas toujours explicites dans les corps de documents. Cela nous dirige vers une approche de catégorisation pour extraire des catégories décisionnelles. Cet article présente une méthode de représentation de document semi-structuré

permettant d'analyser précisément les connaissances et le savoir-faire incorporé dans les contenus et les structures du document. Les expériences sur un corpus juridiques réel montrent que la prise en compte à la fois du contenu et de la structure conduit à une amélioration remarquable de qualité des catégories décisionnelles.

Abstract :

The legal document (law, case law, patent) uses commonly scanning facilities for dissemination and exploitation of legal information through various applications. Thus, the use of legal documentary databases has become more and more open and frequent, leading to a fairly important "semi-automatic" feeding mode. Observation 1: we intend to make a "semi-automatic" process to deposit directly documents in databases controlled by DBMS, including indexing and classification with a limited human intervention. In fact, it is the documents templates (the logical and physical structures modelled by the markup language) that take an important place in the process of indexing and management. Then the templates incorporate indirectly the knowledge and the expertise. Observation 2: in the presence of such a mass data (very often textual), it becomes essential to adopt an approach to manage the electronic legal documents as carriers of knowledge and expertise. This shifts the problem to domains of information retrieval and knowledge discovery. These two observations lead us to formulate an hypothesis for automatic classification that considers the knowledge and expertise incorporated in the structures of the legal electronic documents. This is motivated as we find that the knowledge or expertise are not always explicit in the document body. That pilots us to an approach of categorization to discover decision-making clusters. This article presents a representation method for semi-structured document who allows to analysis very precisely the knowledge and expertise incorporated in both contents and structures of document. The experiments upon a real legal corpus show that incorporation of content and structure produces a remarkable improvement of the quality of decision-making clusters.

1 Introduction

Dans le domaine du document juridique, chaque sous-domaine spécifique (brevets, jurisprudences, par exemple) respecte pratiquement la même structure de rédaction. Ceci peut expliquer l'usage réussi des langages de balisage comme XML (*eXtensive Markup Language*) pour la gestion et l'archivage de documents dans ce milieu. Ainsi nous travaillons sur un ensemble de documents juridiques en format XML, qui représente une base de données semi-structurées.
Partant du fait que les travaux en fouilles des données et la recherche d'information ont montré l'efficacité d'extraction d'information et de recherche d'information à partir des données fortement structurées (cas des bases de données relationnelles), nous supposons que les

informations incorporées dans la structure de documents semi-structurés peuvent aider à mieux catégoriser ces derniers afin d'améliorer la recherche d'information ou la découverte de nouvelles connaissances.
Un des objectifs indirects de cette étude est de présenter la structuration de document comme une démarche anticipative pour la gestion de patrimoine « numérique » dans le domaine de droit. Le document juridique dont la forme doit respecter des règles de rédaction strictes, nous semble fortement intéressant comme objet de recherche.
Nous proposons une méthode de catégorisation structurelle pouvant regrouper automatiquement et efficacement les documents similaires dans les mêmes classes sans aucune connaissance du domaine juridique a priori. Cette méthode d'apprentissage automatique, non-supervisé, peut être considérée comme faisant partie d'un processus de prétraitement de documents en vue de recherche ou d'extraction d'information.
Nos travaux de recherche ont montré que les caractéristiques du domaine juridique, fortement structuré, présentent un facteur favorisant l'extraction d'information à partir de la structure. L'extraction d'information dans notre projet est limitée à l'extraction de catégories décisionnelles pouvant aider le juriste à prendre des décisions en classant un cas d'étude ou de procès en cours.
Nous allons présenter la démarche du travail dans la première section. Dans la deuxième section, nous présentons le modèle de représentation que nous avons retenu. Puis nous détaillons des expériences effectuées sur la jurisprudence du Conseil Constitutionnel français. Avant la conclusion, nous analyserons les résultats des expérimentations.

2 Démarche

Les documents électroniques semi-structurés utilisent des balises XML ayant des propriétés structurelles. Cette opportunité a offert de nouveaux défis à l'apprentissage automatique, et particulièrement à la catégorisation. Plusieurs approches et méthodes ont été proposées à ce propos et peuvent être réparties en deux catégories :
Dans la première catégorie, les travaux ne considèrent que la structure du document. [i] adoptent une approche de traitement de signal pour catégoriser les documents. Les balises XML sont ainsi représentées comme une série temporelle. Et la similarité entre les documents est calculée en analysant des coefficients de transformation de Fourier. [ii] et [iii] proposent d'analyser directement la structure du document XML qui est représentée sous la forme d'un arbre de balises. La catégorisation par la structure du document permet de réduire la structure hétérogène d'un semble de documents. L'inconvénient principal de cette approche réside dans la complexité polynomiale des algorithmes utilisés.

La deuxième catégorie tient compte à la fois du contenu et de la structure d'un document XML. Dans [iv] [v] [vi], l'arbre du document XML est transformé en un sac de chemins, un sac de mots ou un sac mixte de chemins et de mots. Pour représenter l'ensemble de ces descripteurs linéaires, ils adoptent le modèle vectoriel proposé par Salton [vii]. [viii] ont étendu le modèle vectoriel en combinant le contenu, les éléments et les hyperliens dans le document XML. Ces travaux ont montré qu'une approche de catégorisation par l'information de contenu et l'information de structure donne une meilleure précision de regroupement si la structure de la collection en question est homogène.
Nous nous intéressons à découvrir la connaissance et le savoir-faire menés par le contenu et la structure du document. Nos expériences précédentes [ix], [x] montre que l'hétérogénéité de la structuration du document général affecte peu la qualité de la catégorisation thématique. Dans cet article, nous nous concentrons sur les documents juridiques à structuration homogène. Nous réalisons un processus heuristique pour comparer au fur et à mesure les différents descripteurs de document semi-structuré : d'abord, le descripteur de mot classique est utilisé ; ensuite, les descripteurs de structure seule sont examinés ; à la fin, le contenu et la structure hiérarchique du document sont pris en compte globalement. En comparant les résultats de trois séries de catégorisation, nous pouvons explorer le savoir-faire de la structure pour le prétraitement de patrimoine de documents juridiques.

3 Spécificités du document semi-structuré

Conserver le patrimoine exige de ne pas perdre l'information, donc on s'oriente vers une approche de traitement et de prétraitement qui concerne au maximum l'information encapsulée dans un document. Le document semi-structuré propose un modèle hiérarchique qui est généralement considéré comme un arbre. Les travaux existant ont montré que la complexité de catégorisation des arbres est élevée. Nous adoptons une méthode qui transforme une représentation arborescente du document en une représentation vectorielle sans pourtant perdre les informations hiérarchiques de l'arbre.
La figure 1 montre un exemple de document du Conseil Constitutionnel français structuré au format XML. On représente ce document en structure arborescente par des composants linéaires. Chaque composant représentant un type de l'information de contenu ou de l'information structurelle est un descripteur du document. Le modèle de chemins est choisi pour représenter l'information hiérarchique de la structure. Un chemin est une séquence ordonnée d'éléments qui représente une série consécutive de relation parent-enfant. Un chemin complet est une

séquence d'éléments qui commence à l'élément racine et se termine à un élément feuille (voir la figure 2). La longueur d'un chemin est le cardinal de l'ensemble d'éléments dans la séquence. En limitant la longueur d'un chemin complet, on peut créer différents types de sous-chemins. A partir de l'élément racine, après avoir compté *n* éléments, un chemin enraciné de longueur *n* est crée. À l'inverse, un chemin feuillu est crée à partir d'un élément feuille. En attachant le mot contenu dans un élément d'un chemin, on crée un chemin textuel qui comprend à la fois l'information de contenu et l'information de structure

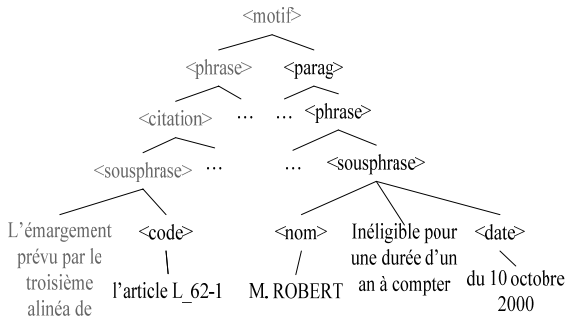

Figure 1. Un morceau d'un document du Conseil Constitutionnel français en XML

Descripteurs	Exemples
chemin complet	\motif\phrase\citation\sousphrase\
|chemin enraciné|=3	\motif\phrase\citation\
|chemin feuillu|=2	\ciataion\sousphrase\
|chemin enraciné&feuillu|=2	\motif\phrase\ \ciataion\sousphrase\
chemin textuel complet	\motif\phrase\citation\sousphrase\alinéa
|chemin textuel enraciné|=3	\motif\phrase\citation\alinéa
|chemin textuel feuillu|=2	\ciataion\sousphrase\alinéa
|chemin textuel enraciné & feuillu|=2	\motif\phrase\alinéa \ciataion\sousphrase\alinéa

Figure 2. Descripteurs structurels du chemin '\motif\phrase\citation\sousphrase\'

Un document peut être représenté par un ensemble de composants de même type (par exemple, les mots, les chemins complets, les chemins textuels enracinés), ou de types différents (par exmple, le mixte de chemin enraciné et de chemin feuillu). Le descripteur de l'approche structurel (le chemin ou le chemin textuel) peut être représenté, comme le descripteur de l'approche de contenu (le mot), dans un vecteur dont chaque dimension correspond à un descripteur. Donc on peut adopter

directement le modèle vectoriel de Salton. Selon l'approche statistique, le nombre d'occurrence peut être un facteur pour calculer l'importance d'un descripteur. Nous utilisons le coefficient TF-IDF pour mesurer son importance. La fréquence d'un descripteur *t* dans un document *d* est définie par l'équation suivante:

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k n_{k,d}}$$

$n_{t,d}$ est le nombre d'occurrence d'un descripteur *t* dans un document *d* ;
$\sum_k n_{k,d}$ est le nombre d'occurrence de tous les descripteurs dans un document *d*.

$$IDF_t = \log(\frac{N}{df_t})$$

N est le nombre total de documents dans la collection ;
df_t est le nombre de documents contenant un descripteur *t*.

4 Expérimentations

4.1 Corpus du Conseil Constitutionnel français

Notre corpus est extrait de la base de documents du Conseil Constitutionnel français qui collecte toutes les publications du Conseil. 2204 documents au sujet de l'élection parlementaire entre 1958 et 2003 ont été sélectionnés. Chaque document décrit des jugements du Conseil sur le contentieux électoral en trois domaines: l'éligibilité de la candidature, le déroulement des opérations et le respect des règles de financement des campagnes. Parmi eux, le contrôle de financement de campagnes couvre une grande partie (53,9%) de la collection. Un document se compose une description des analyses des moyens invoqués, une indication des principes applicables et une réponse à la requête. Deux réponses sont majoritaire : l'inéligibilité de la candidature (49,6% de jugements) et le rejet de la saisie (47,2% de jugements). Donc, nous avons attribué manuellement à chaque document deux types d'étiquette de classes : un sur le sujet du contentieux (« financement » ou « autre »); un autre sur la décision rendue à répondre à la requête (« inéligibilité », « rejet » et « autre »). La structuration de tout le document respecte strictement une règle de rédaction. Autrement dit, les structures de l'ensemble de documents sont homogènes. La figure 1 montre un exemple de la structure du document en XML.

4.2 Prétraitement

Le prétraitement du document consiste à sélectionner les descripteurs pertinents pour la catégorisation. La catégorisation s'appuie sur la comparaison de similarité entre les documents. Plus les documents apportent des descripteurs communs, plus similaires ils sont. Cependant, les descripteurs non contributifs pour la comparaison doivent être éliminés. Par exemple, pour le descripteur de contenu, les mots non significatifs (« le », « de », etc.) sont enregistrés dans une liste (*stoplist*) et sont enlevés avec les chiffres. Les mots sont rendus à leurs formats canoniques en appliquant l'algorithme de *Porter Stemming* [xi] pour réduire le bruit. Les descripteurs couvrant seulement au-delà de 80% des documents dans la collection, et ceux qui se présentent dans quelques documents particuliers (en pourcentage inférieur à 0,5%), sont considérés peu contributifs pour la comparaison de similarité des documents et sont retirés. Avec l'algorithme de prétraitement, 11 types de descripteur sont crées. Chacun est modélisé par une matrice construite de la même façon. Ces matrices sont envoyées à un algorithme de catégorisation hiérarchique.

Algorithme 1: Algorithme de prétraitement

Input: Collection de documents XML : *C*,
Liste de mots vides : *Stoplist*,
Seuils de filtrage : *haut* = 80% et *bas* = 0,5%
Output: 11 matrices correspondent à 11 représentations

```
1 :  Index[mot] = BuildIndex(C)
2 :  Index[mot] = Supprimer(Index[mot], Stoplist)
3 :  for each mot m do
4 :    if m n'est pas un chiffre then
5 :      Index[mot].PorterStemming(m)
6 :    end if
7 :  end for
8 :  MotDescripteur [ ] = MotDescripteurCreation (C, Index, haut, bas)
9 :  for each chemin ch et longueur de chemin chl do
10 :   CheminDescripteur.ch[ ] = CheminCreation (C, ch, chl, haut, bas)
11 : end for
12 : for each chemin textuel cht et longueur de chemin textuel chtl do
13 :   CheminTextuelDescripteur.ct[ ] =
           CheminTextuelCreation(C, DescripteurMot, cht, chtl, haut, bas)
14 : end for
15 : for each descripteur d do
16 :   return MatriceCreation (d, C)
17 : end for
```

4.3 Méthode de catégorisation

Un algorithme de partition hiérarchique agglomératif proposé par l'outil CLUTO [xii] est utilisé. Cette méthode traite la catégorisation comme un

processus d'optimisation dont l'objectif est de maximiser une fonction de critères particuliers définies localement sur l'ensemble des solutions de catégorisation [xiii]. Une partition de K-parcours est obtenue via bi-sections répétées. Une bi-section consiste à une application récursive de la procédure d'optimisation de catégorisation de 2-parcours. Voici la fonction de critère utilisée

$$\sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} sim(u,v)} \quad \text{où} \quad sim(u,v) = \cos\theta = \frac{u \cdot v}{\|u\|\|v\|}$$

u et *v* sont deux vecteurs documentaires. Le processus d'optimisation doit maximiser cette fonction. La similarité entre deux vecteurs documentaires est mesurée par le cosinus.

Algorithme 2: Algorithme de catégorisation

Input: Matrice *M*,
Nombre de catégorie souhaitée : *k*,
Fonction de critère : *f*

Output: *k* catégoires : C[ ]

```
1 :  Ctemp = CategorisationInitiale(M)
2 :  i = 0
3 :  while i < k do
4 :    {C0, C1} = PartionEnDeuxCategories(Ctemp, f)
5 :    Ctemp = ChoisirUneCategorie(C0, C1)
6 :    if Ctemp == C0 then
7 :      C[i] = C1
8 :    else
9 :      C[i] = C0
10 :  end while
11 :  return C[ ]
```

4.4 Résultats

La catégorisation est évaluée quantitativement par l'entropie et la pureté ([xiv]). Deux évaluations sont proposées sur le corpus : l'évaluation de catégorisation thématique est une approche traditionnelle ciblée à la recherche d'information ; alors que l'évaluation sur la décision rendue s'adresse à extraire des catégories décisionnelles.
Pour l'évaluation thématique, deux séries sont lancées en différenciant le nombre de catégories. Pour le descripteur « mot », la qualité mesurée par deux coefficients augmente nettement : 26,9% pour l'entropie et 9,3% pour la pureté avec l'augmentation du nombre de catégories. La même tendance est trouvée également pour certains descripteurs. Un constat intéressant est que la qualité de catégorisation pour le descripteur « chemin feuillu » et les descripteurs « chemin textuel feuillu », « chemin textuel enraciné et feuillu », et « mixte de balise seule et mot » restent constant malgré une augmentation du nombre de catégories. L'approche

du chemin textuel permet une meilleure qualité que les deux autres approches quand le nombre de catégories est fixé à 2.

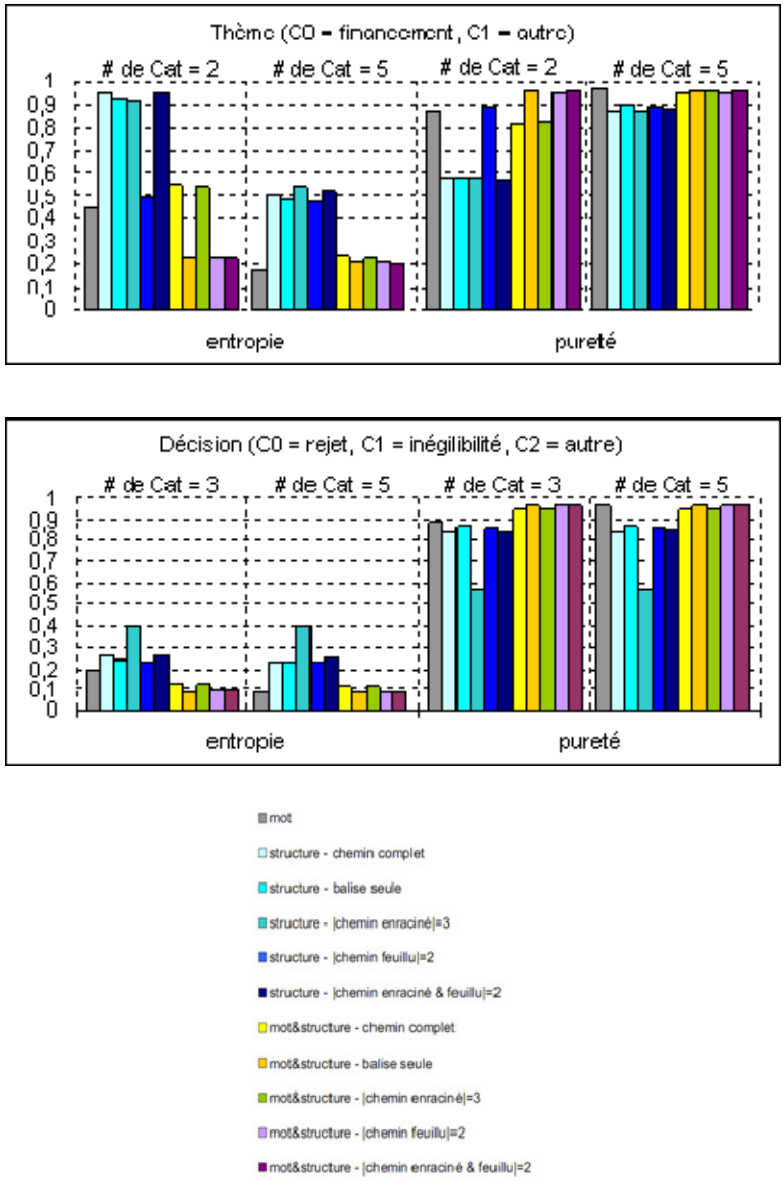

Figure 3. Résultats de l'évaluation thématique et de l'évaluation décisionnelle sur 11 descripteurs de trois approches

Au point de vue traditionnel pour une catégorisation thématique, deux documents proches partagent une partie de mots communs significatifs.

Le vocabulaire du document joue un rôle important dans ce cas. La structure du document n'apporte pas de vocabulaire approprié au thème du document. Pour cela, la qualité de l'approche de la structure seule reste limitée. Cependant, on observe que la structure offre une stabilité considérable. En combinant le mot et la structure, la qualité de catégorisation est nettement augmentée. La qualité brillante du descripteur « mixte de balise seule et mot » implique l'importance du vocabulaire de structure. Les productions de descripteur « chemin (textuel) feuillu » et de « chemin (textuel) enraciné et feuillu » montrent l'importance de l'information hiérarchique de la structure. Le descripteur « chemin (textuel) feuillu » prenant une sous-structure reposant sur les éléments feuilles est plus intéressante que le descripteur « chemin (textuel) enraciné » basé sur l'élément racine et ainsi que le descripteur « chemin (textuel) complet » reflétant la hiérarchie complète.
En ce qui concerne l'évaluation de la décision rendue, on constate que, à l'exception du descripteur « mot », le nombre de catégorie influence peu la qualité de catégorisation. Mis à part l'exception du descripteur « chemin (textuel) enraciné », tous les descripteurs produisent une qualité élevée : la valeur de l'entropie est inférieure à 0,3, et la valeur de la pureté est supérieure à 0,8. Parmi eux, l'approche du chemin textuel produit les meilleurs scores. L'approche du mot mène à une qualité élevée par rapport à l'approche de la structure seule. Même si cette dernière peut conduire à une qualité de catégorisation satisfait. Mais une combinaison du mot *et* de la structure offrent une qualité bonne et stable. Au contraire des résultats de l'évaluation thématique, tous les descripteurs de chemins textuels produisent de bons scores. Ces observations impliquent que le savoir-faire mené par la structuration du document est liée à la décision rendue.
En comparant deux évaluations effectuées, on constate que les résultats de catégorisation décisionnelle sont plus stables quand le nombre de catégories retenues augmente. Nous conclussions que les catégories retenues sont plutôt une partition des jugements qu'une partition thématique.

5 Conclusion

Dans cet article, nous proposons une méthode pour découvrir la connaissance et le savoir-faire du patrimoine de documents juridiques semi-structuré. Les résultats montrent que l'importance de l'information hiérarchique de la structure du document pour stabiliser la partition thématique de documents juridiques et pour l'extraction d'information décisionnelle par catégorie. En comparant avec le modèle classique « sac de mots », on remarque que la représentation tenue à la fois du contenu et

des sous-structures hiérarchiques du document améliore généralement ici la qualité de la tâche de prétraitement de documents juridiques. Et l'amélioration se trouve sous condition que la structuration de tous les documents soit homogène.
Malgré une approche structurelle testée sur un corpus homogène à la structure, notre méthode doit permettre de modéliser les documents à la structuration hétérogène qui est le cas pour la base documentaire hétérogène ou les documents en Web. Ceci doit être développé et testé dans nos futurs travaux.

6 Références bibliographiques

[i] Flesca S., Manco G., Masciari E., Pontieri L., Pugliese A. Detecting Structural Similarities between XML Documents. In *Proceedings of the International Workshop on the Web and Databases (WebDB)*. 2002

[ii] Nierman A., Jagadish H. V. Evaluating Structural Similarity in XML Documents. In *Proceedings of the Fifth International Workshop on the Web and Databases (WebDB 2002)*, Madison, Wisconsin, USA. 2002

[iii] Francesca F. D., Gordano G., Ortale R., Tagarelli A. Distance-based Clustering of XML Documents. In *L. De Raedt et T. Washio (Eds.), MGTS-2003 : Proceedings of the First International Workshop on Mining Graphs, Trees and Sequences*, pp. 75–78. 2003

[iv] Joshi S., Agrawal N., Krishnapuram R., Negi S. A bag of paths model for measuring structural similarity in Web documents. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003

[v] Leung H., Chung F., Chan S.C.F., Luk R. XML Document Clustering Using Common XPath. *In WIRI'05 Proceedings of the 2005 International Workshop on Challenges*. 2005

[vi] Vercoustre A.M., Fegas M., Gul S., Lechevallier Y. A Flexible Structured-based Representation for XML Document Mining. *In: Workshop of the INitiative for the Evaluation of XML Retrieval (2005)*. page 443-457. 2005

[vii] Salton G. *Automatic Text Processing*. Addison-Wesley Publishing Company. 1988

[viii] Yang J., Chen X. A semi-structured document model for text mining. *J. Comput. Sci. Technol*. 17(5), 603–610. 2002

[ix] Yao J. et Zerida N. Rare patterns to improve path-based clustering of Wikipedia articles, In *XML data mining challenge INEX'07*, Dagstuhl, Germany, 2007

[x] Yao J. et Zreik K. La question de la structure dans la catégorisation de documents XML hétérogènes. In *Systèmes Intelligents*, Edited by Mustapha Bellafkih, Mohammed Ramdani, Khaldoun Zreik. ISBN 978-2-909285-53-3, Ed. Europia, Juin 2009

[xi] Porter M.F. An algorithm for suffix stripping. *Program*, 14(3) pp 130−137. 1980

[xii] Karypis G. CLUTO: A Software Package for Clustering High-Dimensional Data Sets. *University of Minnesota, Dept. of Computer Science*, Minneapolis, MN, Nov. 2003. Release

[xiii] Zhao Y. and Karypis G. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, Vol. 10, No. 2, pp. 141 - 168. 2005

[xiv] Zhao Y. and Karypis G. Criterion functions for document clustering: Experiments and analysis. *Technical Report TR #01–40, Department of Computer Science, University of Minnesota*, Minneapolis, MN, 2001.