N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département d'Informatique

MEMOIRE

Présenté pour l'obtention du diplôme de Master

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Systèmes d'information avancées

Par :

ABDELOUAHAB

Soumia

SUJET

Processus de classification supervisée de textes arabes par la méthode K PPV Application aux articles de presse

Soutenu publiquement le : **devant le jury composé de :**

Mr.	Université de M'sila	Président
Mr. MAHDJOUBI Roussafi	Université de M'sila	Rapporteur
Mr.	Université de M'sila	Examinateur
Mr.	Université de M'sila	Examinateur
Mr.	Université de M'sila	Examinateur

Promotion : 2011 /2012

TABLE DES MATIERES

CHAPITRE 2 : CLASSIFICATION AUTOMATIQUE DES DOCUMENT

CHAPITRE 3 : PROCESSUS DE LA CLASSIFICATION DES DOCUMENTS

INTRODUCTION GENERALE

Les sciences de l'information et de la communication ont connu durant ces dernières décennies une évolution technologique remarquable après l'apparition de l'internet. Cette dernière a contribué énormément à la globalisation du monde et à la réalisation du « village planétaire » par le biais du nombre croissant d'informations mises en ligne chaque jour, le développement continu des infrastructures de communication, ainsi que la progression constante du nombre de personnes connectées au réseau mondial. Selon [S1], le nombre de sites en ligne en 2010 a dépassé le nombre de 206,026,787 dont le contenu textuel est écrit en plusieurs langues (anglais, français, russe, chinois, arabe, persan, hébreu, etc.).

L'organisation de toute cette immense et gigantesque ressource est donc indispensable. Ainsi, les techniques de la fouille de textes par le biais de l'apprentissage artificiel, dont la classification automatique des documents fait partie, s'avèrent être pertinentes et très efficaces.

Depuis les années 1960, les chercheurs se sont intéressés à la question de la classification automatique des documents. La majorité des travaux a été réalisée sur des documents écrits en caractères latins (français, anglais, espagnol, etc.). En revanche, très peu de travaux se rapportent à la classification automatique des documents écrits en caractères arabes malgré la richesse morphologique de cette langue. Pour construire un modèle efficace de classification automatique des documents à catégoriser, le traitement automatique de ces derniers est un élément essentiel. Lors de ce traitement, la richesse morphologique de la langue arabe entraine des difficultés qui empêchent les chercheurs d'adopter directement les méthodes et les résultats obtenus par les travaux portant sur la classification automatique des documents écrits en caractères latins sans mettre en question leur validité, d'où la nécessité de tester la fiabilité et l'efficacité des solutions existantes sur des documents écrits en caractères arabes avant de pouvoir tirer des conclusions décisives.

La classification automatique des documents passe par une suite d'opérations d'analyse et de prétraitement sur les données textuelles afin de les préparer aux algorithmes d'apprentissage. Notre travail consiste ici, à développer une application qui permet de voir et mettre en évidence chacune des étapes de ce processus et ses résultats intermédiaires. L'algorithme choisi pour ce traitement est celui de la méthode kppv appliqué à un corpus d'articles de presse en langue arabe.

Ce travail s'articule ainsi, autour de cinq chapitres organisés comme suit :

Le premier chapitre donne une présentation générale sur : le processus d'extraction de connaissances à partir de données, la fouille de données et la fouille de textes accompagnée de quelques techniques qui lui sont liées.

Le deuxième chapitre donne une présentation générale sur la catégorisation automatique des textes, les différentes méthodes utilisées et ce termine par la présentation de quelques travaux antérieurs.

Le troisième chapitre est consacré à l'explication détaillée des différentes étapes du processus de catégorisation automatique de documents.

Le quatrième chapitre donne en détails le processus d'extraction des racines des mots qui est basée sur l'analyse morphologique des textes en langue arabe.

Le cinquième chapitre décrit le fonctionnement de l'application suivi par une phase d'évaluation des résultats obtenus à l'aide du classifieur implémenté afin de décider de sa fidélité.

Enfin, nous clôturerons par une conclusion qui mettra le point sur l'essentiel de ce travail et ses perspectives.

Conclusion générale

L'utilisation de la langue arabe comme moyen de communication à travers le support informatique a été longtemps appréhendée avec beaucoup d'hésitation par la communauté scientifique, notamment celle du monde arabe où cet outil trouvera beaucoup d'utilisations importantes. En effet, la langue arabe et les différentes difficultés qui s'y rattachent, notamment le problème de l'ambiguïté issue de l'absence des voyelles, le problème de reconnaissance des formes fléchies (la langue arabe étant fortement flexionnelle) et le problème d'absence de travaux publiés sur l'extraction de l'information en langue arabe à travers l'utilisation de modèles statistiques du langage, s'ajoute à cela, la diversité des techniques et méthodes relatives au processus de classification qui pose un problème de choix, tout cela pose un énorme défi difficile à surmonter.

Malgré tout cela, nous avons osé nous aventurer dans ce domaine et on peut dire que, vu les résultats obtenus, nous pensons qu'on a quand même pu relever ce défi et par la même occasion apprendre beaucoup de nouvelles connaissances tout au long de la réalisation de ce travail telles que la programmation objet (C # et Delphi), des concepts telles que la classification automatique des documents et les techniques de recherche de l'information, le traitement automatique du langage naturel (en particulier, la langue arabe).

Toutefois, le sujet étant très vaste, il reste beaucoup à faire pour améliorer ce travail, on peut donc proposer comme perspectives, d'ajouter d'autres langues pour rendre le système multi-langues, d'étendre l'éventail des formes des mots arabes pris en compte par l'analyse morphologique, d'intégrer d'autres techniques et méthodes de classification supervisée, ainsi que toute autre idée jugée utile, réalisable et bénéfique.

BIBLIOGRAPHIES ET WEBOGRAPHIES

Bibliographies

[1] Piatetsky-Shapiro, G. et Frawley, W. J., éditeurs (1991). Knowledge-Discovery in Databases. AAAI Press / MIT Press, Menlo Park (Ca) and Cambridge (Ma).

[2] Makhlouf LEDMI , Classification Automatique des documents XML , mémoire de fin d'études pour l'obtention du diplôme de Magistère en informatique,Ecole Doctorale Sciences et Technologies de l'Information et de la Communication ,Option : Systèmes d'Informations et de Connaissance,2010 .

[03] David Hand, Heikki Mannila & Padhraic Smyth, 2001, Principles of Data Mining, MIT Press, Cambridge, MA.

[04] Stéphane Tufferry, 2002, Fouille de données et scoring, bases de données et gestion de la relation client, Dunod, Paris

[05] David Hand, Heikki Mannila & Padhraic Smyth, 2001, Principles of Data Mining, MIT Press, Cambridge, MA.

[06] Sadik Bessou, Analyse de Données Textuelles pour la Classification Automatique par les Techniques de Text Mining, application à la Langue Arabe, Mémoire de Magister En Informatique , Université de Sétif,2007.

[7] Azizi Nabil , Apprentissage automatique et fusion d'informations Application à l'extraction des connaissances des documents web, Mémoire de Magister En Informatique,Ecole Doctorale Sciences et Technologies de l'Information et de la Communication ,Option Systèmes d'Informations et de Connaissances.

[8] Assila.S, Slimani .N , Classification supervisée de textes arabes par la méthode K PPV, Application au Hadith , Mémoire d'ingénieur d'état en Informatique, Université de M'sila ,2011.

[9] Smyth, 2000, Bref historique sur le data mining.

[10] Manu Konchady, 2007, Text Mining Application Programming, Charles River Media Programming series, USA .

[11] Hearst, M. A et al, 2000 . The debate on automated essay grading, IEEE Intelligent systems (September 2000).

[12] Manu Konchady, 2007, Text Mining Application Programming, Charles River Media Programming series,USA

[13] Saeed RAHEEL, L'Apprentissage Artificiel pour la Fouille de Données Multilingues : Application à la Classification Automatique des Documents Arabes, Thèse de doctorat en Sciences de l'Information et de la Communication Université Lumière Lyon 2 . 22 octobre 2010,

[14] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys.

[15] Jalam, R. (2003). Apprentissage automatique et catégorisation de textes multilingues.Thèse de doctorat, Université LUMIERE LYON2.

[16] Romain Vinot , Natalia Grabar & Mathieu Valette, 2003, Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet, TALN 2003.

[17] Zighed, D. A. et Rakotomalala, R. (2000). Graphes d'induction. Apprentissage et Data Mining. Hermes Science Publication, Paris.

[18] René Lefébure, Gilles Venturi, 2001, Data mining. Gestion de la relation client. Personnalisation de sites web. Eyrolles.

[19] Vapnik, V. (1995). The Nature of Statistical Learning Theory.Springer Verlag.

[20] Catanzaro B., Sundaram N., Keutzer K., "Fast Support Vector,Machine Training and Classification on Graphics Processors", In : International ,Conference on Machine Learning, 2008.

[22] Cao, L.J. "Support Vector Machines Experts for Time. Series Forecasting", Neurocomputing, 2003.

[23] Maron. M. Automatic indexing : an experimental inquiry. Journal of the Association for Computing Machinery , 1961.

[24] Belkin, N. J. and Croft, W. B. 1992. Information filtering and information retrieval: two sides of the same coin?. Commun. ACM 35, 12 (Dec.1992).

[25] Liddy et al. 1994] Liddy, E. D., Paik, W., and Yu, E. S. 1994. Text categorization for multiple users based on semantic features from a machine-readable dictionary.

[26] Schütze, H., Hull, D. A., and Pedersen, J. O. 1995. A comparison of classifiers and document representations for the routing problem. In Proceedings of the 18th Annual

international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995).

[27] Bhagwat, D., Eshghi, K., and Mehra, P. 2007. Content-based document routing and index partitioning for scalable similarity-based searches in a large corpus. In Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Jose, California, USA, August 12 - 15, 2007).

[28] Iyer, R. D., Lewis, D. D., Schapire, R. E., Singer, Y., and Singhal,A. 2000. Boosting for document routing. In Proceedings of the Ninth international.Conference on information and Knowledge Management (McLean, Virginia,United States, November 06 - 11, 2000). CIKM '00. ACM Press, New York.

[29] Zhou, D., Burges, C. J., and Tao, T. 2007. Transductive link spam detection. In Proceedings of the 3rd international Workshop on Adversarial information Retrieval on the Web (Banff, Alberta, Canada, May 08 - 08, 2007).

[30] Lefèvre P. La Recherche d'informations, Hermès, 2000.

[31] Delphine Bernhard, 2007, Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique, TALN 2007, Toulouse.

[32] Shuchuan Lo, 2006, Web service quality control based on text mining using support vector machine, Expert Systems with Applications, vol 34, n°1, pp 603-610, Elsevier.

[33] Amy J.C. Trappey, Fu-Chiang Hsu Charles V. Trappey & Chia-I. Lin, 2006, Development of a patent document classification and search, platform using a back-propagation network, Expert Systems with Applications, Elsevier

[34] Laurent Pierron, Coskun Durkal et Jean-Baptiste Chevalier, 2005, Classification, combinaison et regroupements pour séparer les discours de Mitterrand de ceux de Chirac, Atelier DEFT'05, TALN 2005, Dourdan.

[35] Robinson G, 2003, A statistical approach to the spam problem, Linux journal

[36] Patrice Bellot, Marc El Bèze, 2001, Classification et segmentation de textes par arbres de décision. Application à la recherche documentaire, Technique et Science informatiques. Vol..

[37] Jouadi W, Benghezala H, Zrigui M, 2007, La distance intertextuelle pour la classification de textes en langue arabe, CITALA 2007, Rabat, Maroc.

[38] F. Denis, R. Gilleron, A. Laurent, M. Tommasi, Text Classification and Co-Training from Positive and Unlabeled Examples,Proceedings of the ICML 2003 Workshop : The Continuum from Labeled toUnlabeled Data.

[39] Song Y., Zhou D., Huang J., Councill I. G., Zha H., C. Lee Giles : Boosting the Feature Space : Text Classification for Unstructured Data on the Web. ICDM 2006.

[40] Escudero, G., Màrquez, L., and Rigau, G. 2000. Boosting Applied toe Word Sense Disambiguation. In Proceedings of the 11th European Conference on Machine Learning (May 31 - June 02, 2000). R. L. Mántaras and E. Plaza, Eds. Lecture Notes In Computer Science, vol. 1810. Springer-Verlag, London.

[41] Suárez, A. and Palomar, M. 2002. A maximum entropybased word sense disambiguation system. In Proceedings of the 19th international Conference on Computational Linguistics - Volume 1 (Taipei, Taiwan, August 24 -September 01, 2002). International Conference On Computational Linguistics.Association for Computational Linguistics, Morristown.

[42] Plantié M., Roche M., Dray G., EGC 2008 : Un système de vote pour la classification de textes d\'opinion, Laboratoire LGI2P, Laboratoire LIRMM.

[43] Khreisat, L. Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study, Proceedings of the 2006 International Conference on Data Mining. Las Vegas, USA, 2006.

[44] Mesleh, A. Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. Journal of Computer Science 3 (6) : 430-435, 2007.

[45] Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In Belkin, N. J., Ingwersen, P. and Pejtersen, A. M., editors Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, pages 37-50,Kobenhavn, DK. ACM Press, New York, US.

[46] Blum, A. et Mitchell, T. Combining Labeled and Unlabeled Data with Co-training. Proceedings of the 11th Annual Conference on Computational Learning Theory, pp. 92-100, 1998.

[47] Deerwester, S. Dumais, S. Landauer, T. Furnas, G. et Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science, 416(6) : 391-407.

[48] Baeza_yates R. & B Ribeiro-Neto, 1999, Modern information retrieval. ACM press books .

[49] Sebastiani, F. (1999). A tutorial on automated text categorization. Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR, 1999, pp. 7-35.

[50] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information Retrieval

Webographies

[S1] http://news.netcraft.com/archives/web_server_survey.html.
[Date de dernière visite : Mai 2010].

[S2] http://www.technique-ingénieur.fr. [Date de dernière visite : Février 2012].

[S3] http ://www.shopwiki.com/.

[S4] http ://www.nextag.com/

[S5] http ://www.scirus.com/,

[S6] http ://www.medstory.com/

[S7] http://www.daviddlewis.com/resources/testcollections/reuters21578/.
[Date de dernière visite : Mai 2012].

[S8] http://davis.wpi.edu/~xmdv/datasets/ohsumed.html .
[Date de dernière visite : Mai 2012].

Résumé :

Le travail entrepris dans le cadre de ce mémoire concerne la réalisation d'un système qui permet de suivre pas à pas le processus de classification automatique supervisée des articles de presse écrits en caractères arabes à partir de catégories déjà connues (Politique, Economie, Religion et Sport) en utilisant l'algorithme d'apprentissage KNN.

Mots clés : Classification supervisée, Algorithme d'apprentissage, Articles de presse, TALN arabe, KNN.

Abstract :

The work undertaken within the framework of this memory relates to the realization of a system which makes it possible to follow step by step the process of supervised classification automatic of the articles of Arab newspaper industry in characters S starting from already known categories (Policy, Economie, Religion and Sport) by using the *algorithm of training KNN.*

Key words: Supervised classification, Algorithm of training, Articles of press, Arab TALN, KNN.

ملخص :

العمل المنجز في إطار هذه المذكرة يتعلق بتصميم برنامج يسمح بمتابعة عملية التصنيف الآلي للنصوص الإخبارية باللغة العربية، بناءا على أنواع معروفة مسبقا وهي الأخبار(السياسية،الاقتصادية،الدينية والرياضية) ، وهذا من خلال استخدام خوارزم التصنيفKNN والذي يعتمد في التصنيف على الحالات المجاورة.

الكلمات المفتاحية : التصنيف الآلي للنصوص،النصوص الإخبارية، خوارزم التصنيف KNN،المعالجة الآلية للغة العربية.