REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUE ET DE L'INFORMATIQUE

Département de stic

Mémoire présenté pour l'obtention du diplôme de **master**

Domaine : mathématique et informatique

Filière : Informatique

Spécialité : technologie de l'information et la communication

L'apport de la sélection des termes sur la catégorisation automatique des textes arabe

Préparer par :

HAFIDI Mustapha

supervise par :

Mr.KADRI Said

2014/2015

Tableau de matières

Introduction générale

La quantité d'information disponible sous format électronique sur Internet ou dans l'intranet des entreprises croît de façon extrêmement rapide. Par exemple, le 22 juillet 2002, Google avait recensés 2.073.418.204 pages ; le 16 mars 2003, ce nombre est passé à 3.083.324.652 pages, soit moitié plus en moins d'un an (voir http://www. google.fr). L'utilisateur, submergé par cette masse d'informations, ne se pose plus la question d'accéder à l'information . Son problème devient : comment trouver l'information dont il a besoin, parmi toutes celles qui est accessible. Dans ce contexte, La classification de textes a pour objectif de regrouper les textes similaires, c'est à dire thématiquement proches, au sein d'un même ensemble. L'intérêt d'une telle démarche est d'organiser les connaissances de façon à pouvoir effectuer, par la suite, une recherche ou une extraction d'information efficace. Le volume de documents numériques s'accroissant, des besoins en classification automatique se sont fait ressentir aussi bien sur internet (moteurs de recherche), qu'au sein des entreprises (classement de documents internes, dépêches d'agences, etc.). On distingue dans le domaine de la classification automatique deux types d'approches : la classification supervisée et la classification non supervisée. Ces deux méthodes diffèrent sur la façon dont les classes sont générées. En effet dans le cas de la classification non supervisée, les groupes de documents (classes) sont calculés automatiquement par la machine [SAL 83, IWA 95], tandis qu'ils sont, dans l'approche supervisée [JOA 98b, SEB 02, YAN 99a], définis par un expert. Dans ce dernier cas, il est intéressant de représenter les documents et les classes à l'aide d'un même formalisme et celui généralement utilisé est un espace vectoriel [SEB 02, BES 02]. Dans cet article, nous nous intéresserons à la catégorisation, c'est à dire aux algorithmes d'apprentissage supervisés et plus particulièrement aux méthodes basées sur une représentation vectorielle de documents

Conclusion générale

La classification de textes s'est avérée au cours des dernières années comme un domaine majeur de recherche pour les entreprises comme pour les particuliers.

Ce dynamisme est en partie dû à la demande importante des utilisateurs pour cette technologie. Elle devient de plus en plus indispensable dans de nombreuses situations où la quantité de documents textuels électroniques rend impossible tout traitement manuel. La catégorisation de textes a essentiellement progressé ces dix dernières années grâce à l'introduction des techniques héritées de l'apprentissage automatique qui ont amélioré très significativement les taux de bonne classification. Il reste néanmoins difficile de fournir des valeurs chiffrées sur les performances qu'un système de classification peut actuellement atteindre. Les travaux de recherche dans le domaine se focalisent surtout sur deux aspects : l'efficacité et l'amélioration de performances. Dans ces deux optiques nous avons entamé notre Project de catégorisation et sélection des termes.

Bibliographie

BÉCHET, N. (2008). « Extraction et regroupement de descripteurs morphosyntaxiques pour des processus de Fouille de Textes ». doctorat.

BÉCHET, N. (2008, dec 8). « Extraction et regroupement de descripteurs morphosyntaxiques pour des processus de Fouille de Textes ». doctorat.

C.D.Loupy. (2000). *« Évaluation de l'Apport de Connaissances Linguistiques en Désambiguïsation Sémantique et Recherche Documentaire » .*

F.Sebastiani. (2002). *« Machine learning in automated text categorization » .*

G.Brown, H. (1998). *« The Guru System in TREC-6 » .*

I.Moulinier. (1996). *« Une approche de la catégorisation de textes par l'apprentissage symbolique ».*

interstices. (s.d.). Récupéré sur interstices.info: https://interstices.info/encart.jsp?id=c_41867&encart=3&size=600,500

J. Brank, M. G.-F. (2002). Interaction of Feature Selection Methods and Linear Classification Models. *Workshop on Text Learning.*

J.Clech. (2004). *« Contribution méthodologique à la fouille de données complexes » .*

J.Clech, D. (2004). *« Une technique de réétiquetage dans un contexte.*

Karima, A. (2011). Consulté le 2011, sur http://share.esi.dz/55/1/La%20cat%C3%A9gorisation%20de%20texte%20multilingue.pdf

Lang, K. (1995). *« NewsWeeder : Learning to Filter Netnews » .*

Lewis. (1992). *« An evaluation of phrasal and clustered representations on a text.*

M.McGill, G. &. (1983). *« Introduction to Modern Information Retrieval» .*

P.Hayes, S. (1990). *« Construe/Tis : A system for content-based.*

Pedersen., Y. Y. (1997). Comparative Study on Feature Selection in Text Categorization.

R.Armstrong, D. T. (1995). *« WebWatcher : a Learning apprentice for the World Wide Web » .*

Sebastiani, F. (1999). A Tutorial on Automated Text Categorisation. *ASAI-99.*

Sebastiani, F. (1999). A Tutorial on Automated Text categorsiation .

Sebastiani, F. (2002). *« machine lerning in automted texte categorization ».* Récupéré sur http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf

Stricker, M. (2000). « Réseaux de neurones pour le traitement automatique du langage : conception et réalisatiion de filters d'informations ». paris, doctorat.

theses. (s.d.). Récupéré sur theses.ulaval.ca: http://theses.ulaval.ca/archimede/fichiers/22376/ch02.html

C.D.Loupy. (2000). *« Évaluation de l'Apport de Connaissances Linguistiques en Désambiguïsation Sémantique et Recherche Documentaire » .*

F.Sebastiani. (2002). *« Machine learning in automated text categorization » .*

G.Brown, H. (1998). *« The Guru System in TREC-6 » .*

I.Moulinie. (s.d.). *Une approche de la catégorisation de textes par l'apprentissage symbolique .*

I.Moulinier. (1996). *« Une approche de la catégorisation de textes par l'apprentissage symbolique ».*

J.Clech. (2004). *« Contribution méthodologique à la fouille de données complexes » .*

J.Clech, D. (2004). *« Une technique de réétiquetage dans un contexte.*

Lang, K. (1995). *« NewsWeeder : Learning to Filter Netnews » .*

Lewis. (1992). *« An evaluation of phrasal and clustered representations on a text.*

M.McGill, G. &. (1983). *« Introduction to Modern Information Retrieval» .*

P.Hayes, S. (1990). *« Construe/Tis : A system for content-based.*

R.Armstrong, D. T. (1995). *« WebWatcher : a Learning apprentice for the World Wide Web » .*

Sebastiani, F. (1999). A Tutorial on Automated Text categorsiation .

Stricker, M. (2000). « Réseaux de neurones pour le traitement automatique du langage : conception et réalisatiion de filters d'informations ». paris, doctorat.

Résumé/ملخص / Abstract

Avec l'avènement de l'informatique et l'accroissement du nombre de documents électroniques stockés sur les divers supports électroniques et sur le Web, particulièrement les données textuelles, le développement d'outils d'analyse et de traitement automatique des textes, notamment la classification automatique de textes, est devenu indispensable, pour assister les utilisateurs, de ces collections de documents, à explorer et à répertorier toutes ces immenses banques de données textuelles. Ainsi la catégorisation automatique de textes, qui consiste à assigner un document à une ou plusieurs catégories, s'impose de plus en plus comme une technologie clé dans la gestion de l'intelligence, les résultats obtenus sont utiles aussi bien pour la recherche d'information que pour l'extraction de connaissance soit sur internet (moteurs de recherche), qu'au sein des entreprises (classement de documents internes, dépêches d'agences, etc.). À l'égard des différentes approches de classification automatique de textes, décrites dans l'état de l'art, se reposant sur une architecture classique basée sur un seul point de vue, nous avons introduit une nouvelle utilisation du classifieur « Naïve Bayes » avec la méthode de sélection des termes " chi2", L'objectif principal de nos travaux, est d'améliorer les performances et l'efficacité du modèle de classification.

Mots clés : classification , catégorisation automatique , textes , Naïve Bayes , méthode de sélection , chi2.

مع تطور المعلوماتية وتزايد الوثائق الالكترونية ، تاتي ادوار التحليل و المعالجة الاوتوماتيكية لهذه البيانات النصية كضرورة حتمية لمساعدة المستخدمين للاستكشاف وفهرسة هده القواعد النصية الضخمة . و في هذا الإطار ، اصبح التصنيف الاوتوماتيكي للنصوص وسيلة من السائل التكنولوجية الرئيسية لإدارة هذا النوع من الإشكاليات ، و النتائج المتحل عليها مفيدة سواء في البحث عن المعلومات او استخراجها . على غرار التقنيات المختلفة المستعملة في هذا المجال قمنا بتصنيف النصوص عن طريق خوارزمية « Naïve Bayes » مع استعمال طريقة " chi2" لتقليل عد كلمات النص بدون التأثير في التصنيف و الهدف من هذه المذكرة هو امكانية تصنيف النصوص باستخدام طرق رياضية و مقارنة النتائج .

المصطلحات الرئيسية : التحليل ، و المعالجة الاوتوماتيكية ، النصية ، خوارزمية ، تصنيف ، طريقة.

With the advent of computers and the increasing number of electronic documents stored on various electronic media and web, especially text data, development of analysis tools and automatic processing of texts, including automatic text classification has become essential to assist users of these document collections, to explore and identify all these huge banks of textual data. And automatic categorization of text, which is to assign a document to one or more categories, is becoming increasingly recognized as a key technology in the management of intelligence, the results are useful both for the search information to extract knowledge or on the Internet (search engines), and at the company (ranking of internal documents, news agencies, etc.). In respect of different approaches to automatic text classification, described in the prior art, relying on a conventional architecture based on a single point of view, we introduced a novel use of the classifier "Naïve Bayes" with the ethod of selection of terms « chi2 » The main objective of our work is to improve the performance and efficiency of the classification model. The reference corpus Reuters will be used to conduct a comparative study of results.

Kay words : *analysis , automatic processing , texts , classification , categorization , Naïve Bayes , chi2.*