REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département d'Informatique

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Systèmes d'Informations Avancés

Par: DAHOUMI Fares

SUJET

Identification de la langue et catégorisation thématique des textes d'un corpus multilingue en utilisant les algorithmes : NB, SVM

Soutenu publiquement le : / /2013 devant le jury composé de :

.....	**Université de M'sila**	**Président**
Mr. KADRI Said	**Université de M'sila**	**Rapporteur**
.....	**Université de M'sila**	**Examinateur**
.....	**Université de M'sila**	**Examinateur**

Promotion : 2012/2013

Table des matières

INTRODUCTION GENERALE

La recherche en informatique accorde ces dernières années, beaucoup d'importance au traitement des données textuelles. Ceci pour plusieurs raisons : un nombre croissant de collections mises en réseau et distribuées sur le plan international, le développement des infrastructures de communication et d'Internet. Les traitements manuels de ces données s'avèrent très coûteux en temps et en personnel, ils sont peu flexibles et leur généralisation à d'autres domaines est presque impossible ; c'est pour cela que l'on cherche à mettre au point des méthodes automatiques.

Le domaine de la fouille de textes (Texte Mining) s'est développé pour répondre à volonté à la gestion par contenu des sources volumineuses de textes. A l'heure actuelle, de nombreux logiciels de classification de textes sont disponibles, ils ont fait l'objet de publications et leurs champs d'application s'élargit de jour en jour. En général, ces systèmes sont basés sur des algorithmes d'apprentissage automatique (approche statistique, approche syntaxique et approche connexionniste).

Nous nous intéressons ici plus particulièrement aux algorithmes d'apprentissage et nous avons utilisé l'algorithme de Support Vector Machines (SVM). Pour pouvoir utiliser de tels algorithmes, il est nécessaire de transformer les données, initialement en format texte, en une représentation numérique. Nous avons choisi pour ce faire, la méthode de sélection des termes les plus pertinents. Une fois ce prétraitement terminé, nous pouvons effectuer la classification à l'aide de ces algorithmes.

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

Le premier chapitre donne une présentation générale des disciplines suivantes : L'intelligence artificielle (IA), le Data Mining, la fouille de textes ainsi que quelques techniques qui leur sont liées.

Le deuxième chapitre donne une présentation générale sur la apprentissage automatique et la catégorisation des textes, ainsi que les différentes méthodes utilisées dans la littérature.

Le troisième chapitre est consacré à l'explication détaillée des différentes d'approches d'identification de la langue et Les algorithme de catégorisation de textes.

Le quatrième chapitre donne en détails la méthode de classification de textes appelée SVM ou Support Vector Machines.

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 mesurer sa performance.

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

Bibliographie

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

[3] Christopher D.Manning & Hinrich Schutze, Foundation of statistical natural language processing. MIT press, 1999.

[4] Pierce, J.R, an introduction to information theory symbols, signals and noise, Dover publications, 1980.

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

[5] M.Boudjemia et A.bekri, Application d'un algorithme évolutif au problème de la classification non supervisee. Mémoire Ingénieur, 2007.

[6] M.J.A. Berry et G. Linoff, Inter Editions, Data mining: techniques appliqués au marketing, à la vente et au service client, 1997.

[7] P. Naïm et M. Bazsalicza, Editions Eyvolles, Data mining pour le web, 2001.

[8] Ben Messaoud, Data Mining, institut universitaire de technologie lumière, licence C.E.STAT, laboratoire ERIC, 5 avenue Pierre Mendés France 69676 Bron Cedex, 2007.

[9] Gilbert, Data Mining : une nouvelle façon de faire de la statistique, chaire de statistique appliquée, conservatoire national des arts et métiers, 292 rue Saint Martin, 75003 Paris, 2006.

[10] Pascal Vincent, Modèles à noyaux à structure locale, Thèse présentée à la Faculté des études supérieures en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.) en informatique , Université de Montréal, 2003.

[11] Jalam, Radwan , Apprentissage automatique et catégorisation de textes multilingues, Thèse de doctorat, Université Lumière Lyon 2, 2003.

[12] ABIDI Karima, la catégorisation de texte multilingue, mémoire de magistère d'informatique, école supérieur d'informatique, 2011.

[13] Sebastiàn Pena Saldarriaga, Approches textuelles pour la catégorisation et la recherche de documents manuscrits en-ligne, thèse de doctorat, Université de Nantes, 2010 .

[14] Nicolas Béchet, Extraction et regroupement de descripteurs morphosyntaxiques pour des processus de Fouille de Textes, thèse de doctorat, Université des Sciences et Techniques du Languedoc, 2008.

[15] Gardarin Georges, Huaizhong Kou, Karine Zeitouni DocCat : un composant logiciel de catégorisation de documents et de marquage sémantique , XML Laboratoire PRiSM, Université de Versailles-Saint-Quentin 45 , 2003.

Conclusion générale

L'utilisation des langues comme moyen de communication à travers le support informatique a été longtemps appréhendée avec beaucoup d'hésitation par la communauté scientifique du monde où cet outil trouvera beaucoup d'utilisations importantes. En effet, les langues rencontrent plusieurs difficultés qui s'y rattachent, notamment : le problème de l'ambiguïté des expressions, le problème de reconnaissance des formes fléchies, par exemple la langue arabe étant fortement flexionnelle, le problème d'absence de travaux publiés sur l'extraction de l'information exprimée en langue arabe à travers l'utilisation de modèles statistiques du langage. S'ajoute à tout 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 tous ces problèmes, nous avons abordé ce domaine en espérant qu'on arrive à trouver des solutions faisables à chacun des problèmes précités.

Cependant, et après avoir mené l'étude à sa fin, nous pensons qu'on a quand même pu relever le 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 sous le langage C #.
- ✓ Des nouveaux concepts dans des domaines d'actualité tels que : l'intelligence artificielle, l'apprentissage automatique, le Data Mining, le Text Mining, et les SRI.
- ✓ La classification automatique supervisée de documents qui se situe dans l'intersection des domaines précités.
- ✓ Le traitement automatique du langage naturel TALN (en particulier la langue arabe).

Toutefois, le sujet abordé étant très vaste, il reste beaucoup à faire pour améliorer notre système. A cet effet, nous proposons comme perspectives :

- ✓ Ajouter d'autres langues latines et non latines pour rendre le système multilingue dans le sens propre du mot.
- ✓ Utiliser des corpus d'apprentissage et de test de grandes tailles pour donner plus de crédibilité aux résultats obtenus.
- ✓ Appliquer d'autres algorithmes d'apprentissage (les arbres de décision, les RNA, AdaBoost, K-NN, Rocchio, les algorithmes génétiques) et effectuer des comparaisons entre eux en terme performance et temps d'exécution.
- ✓ Utiliser d'autres formes de documents (HTML, XML, …).

- ✓ Etudier le cas de la classification multi-label et la classification non supervisée (clustering).
- ✓ Toute autre idée jugée utile, réalisable et bénéfique dans ce domaine.

[16] Kotsiantis S. Supervised Machine Learning , A Review of Classification ,2007

[17] Bouckaert, Naive Bayes Classifiers That Perform Well with Continuous Variables, Lecture Notes in Computer Science, Volume 3339,Pages 1089 –1094, 2004.

[18] Romain VINOT, classification automatique de textes dans la catégorie non thématique, thèse pour obtenir le grade de docteur d'école nationale supérieur des télécommunication , Présentée et soutenue le 09 février 2007

[19] Sadik Bessou, Analyse de Données Textuelles pour la Classification Automatique par les Techniques de Text Mining, application à la Langue Arabe, Pour l'Obtention du Diplôme d magister ,universite ferhat abbas – setif, 2007.

[20] http://www.grappa.univ-lille3.fr/polys/fouille/sortie005.html 24/04/2013

[21] Luc Grivel, Outils de classification et de catégorisation pour la fouille de textes, Equipe ISIS, Université de Marne-La-Vallée, 2006.

[22] Ronen Feldman & James Sanger, The text mining handbook, advanced approaches in analyzing unstructured data, Cambridge University Press, New York, USA, 2007.

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

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

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

[26] Philippe Besse, Cours d'Apprentissage Statistique & Data mining , Institut National des Sciences Appliquées de Toulouse— 31077 – Toulouse cedex 4, 2006.

[27] Ludovic Mercier, Les machines à vecteurs support pour la classification en imagerie hyperspectrale implémentation et mise en œuvre, Travail d'Etude et de Synthèse Technique en informatique, 2010.

[28] Reda Jourani, Reconnaissance de visages, pour l'obtention du diplôme des études supérieures approfondies , université Mohammed v-agdal, Rabat, 2006.

[29] Delorme, L'apport de la fouille de données dans l'analyse de texte, Cnam (Conservation national des arts et métiers), Centre Régional de Montpelier , 2002.

[30] **Ricardo** Baeza_yates. & Berthier Ribeiro-Neto, Modern information retrieval. ACM press books, 1999.

[31] Stuart Russell & Peter Norvig ,Intelligence artificielle ,Pearson , 3rd Edition , 2010

[32] Dominique Pastre, cours d' intelligence artificielle, Université Paris 5, 2000

[33] **Meliouh.A,** cours d'intelligence artificielle, 2éme master informatique system d'information, université de M'sila , 2012 .

[34] http://social.msdn.microsoft.com 24/04/2013 .

ملخص :

التصنيف الآلي للوثائق أصبح ضروريا بسبب حجم الوثائق المتبادلة والمخزنة إلكترونيا. إن تعدد الوثائق وتزايد عددها المستمر نتج عنه صعوبة في وضع منهاج ونحن هنا قدمنا طرق التعليم المتمثلة في SVM و Naïve Bayes و التي تسمح لنا بتصنيف مستند جديد انطلاقاً من مستند مصنف مسبقا.

كلمات مفتاحية التصنيف الآلي، طرق التعليم، SVM، Naïve Bayes.

Abstract

Automatic classification supervised document becomes necessary due to the volume of documents exchanged and stored electronically. As there are many documents or their number is constantly increasing. It would be difficult to program in advance of the decision rules to determine the class of a new document. We present learning methods ((SVM) Support Vector Machines And Naïve Bayes) that from documents already classified, used to classify new documents.

Keyword : Automatic classification, supervised, Learning methods, SVM , Naïve Bayes.

Résumé :

La classification automatique supervisée de document devient nécessaire à cause du volume de documents échangés et stockés sur support électronique. Comme les documents sont nombreux ou que leur nombre augmente sans cesse. Il serait difficile de programmer à l'avance des règles de décision pour déterminer la classe d'un nouveau document. Nous présentons donc des méthodes d'apprentissage ((SVM) Support Vector Machine et Naïve Bayes) qui à partir de documents déjà classés ; permettent de classer de nouveaux documents.

Mots clés : La classification automatique, supervisée, méthodes d'apprentissage , SVM, Naïve Bayes.