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: MEDJAHED Assia

SUJET

LA RECONNAISSANCE EFFICACE DE LA LANGUE DANS UN CORPUS DE TEXTES MULTILINGUE

Soutenu publiquement le : / /2014 devant le jury composé de :

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

Promotion : 2013 /2014

TABLE DES MATIERES

CHAPITRE 2 :CLASSIFICATION AUTOMATIQUE DE DOCUMENT MONO ET MULTILINGUE

CHAPITRE 3 :LA RECONNAISSANCE DE LA LANGUE :PRINCIPE ET METHODES

CHAPITRE 4 : EXPERIMENTATION ET RELULTATS OBTENUS

INTRODUCTION GENERALE

A cause de l'expansion massive du réseau mondial Internet, et le grand flux d'information échangé entre des milliards d'utilisateurs, la recherche de l'information pertinente est devenue une tache très compliquée qui requière la possession de moteurs de recherche très puissants comme c'est le cas de Google. L'information cherchée peut ne pas exister dans la langue de l'utilisateur, mais dans une langue différente, ce qui complique encore la recherche et diminue sa précision.

Identifier la langue d'un texte veut dire attribuer ce texte à la langue dans laquelle il est écrit. Donc, il s'agit d'un genre de classification automatique ou les classes sont des langues (Ar, Fr, An, …). Une véritable reconnaissance de la langue d'un texte n'est pas possible si on considère seulement le mot comme unité d'information, cela peut être possible pour certaines langues comme le français ou l'anglais, mais très difficile pour certaines d'autres langues comme l'arabe, l'allemand ou le chinois. L'approche de découpage de textes en n-grammes caractéristiques représente une solution alternative très efficace dans ce domaine.

Plusieurs algorithmes ont été proposés pour identifier la langue du texte. La plupart de ces algorithmes sont basés sur la notion de distance ou de similarité. L'idée principale est de chercher le texte, parmi l'ensemble d'apprentissage, qui soit le plus proche en distance du texte à identifier la langue et de lui attribuer la même langue. Le choix d'une telle distance est une difficulté commune pour ces algorithmes. Pratiquement il existe plusieurs métriques ou pseudo-distances, notamment : la distance de Beesley, la distance de Cavnar et Trenkle, la distance de Kullbach-Leibler, la distance (χ^2), …etc.

Notre travail consiste donc à développer un système automatique permettant de reconnaitre les langues d'une collection multilingue de textes en utilisant les algorithmes d'apprentissage. C'est une phase très importante qui facilitera l'extraction de l'information pertinente cachée dans le texte et par conséquence de le catégoriser (l'affecter à une catégorie) avec moins d'erreur. L'algorithme d'apprentissage choisi dans notre système est l'algorithme des k-plus proches voisins (k-nearest neighbors KNN) qui base essentiellement sur le calcule de la distance entre le nouveau texte à identifier la langue et

chaque texte de la base d'apprentissage. Ce choix est justifié par la facilité d'implémentation de cet algorithme et l'exactitude de ses résultats.

Pour l'organisation du mémoire, nous avons opté pour la structure suivante :

Un premier chapitre donnant une présentation générale du processus d'extraction de connaissances à partir de données, de la fouille de données et la fouille de textes, ainsi que quelques techniques liées.

Le deuxième chapitre expose d'une manière générale la classification automatique des textes, les différentes méthodes utilisées et les domaines d'applications.

Le troisième chapitre explique le principe et les méthodes de l'identification de la langue.

Le quatrième chapitre expose l'architecture du logiciel conçu et son fonctionnement, ainsi que son implémentation et quelques exemples de démonstration.

Une conclusion générale qui résumé ce qui a été fait, les connaissances acquises à travers la réalisation du projet, les difficultés rencontrées, et enfin quelques perspectives pour des travaux futurs.

Conclusion générale

Nous avons présenté dans ce mémoire, un apperçu général sur les approches d'analyse intelligente de documents qui basent essentiellement sur les techniques d'apprentissage automatique, notre but était l'identification de la langue des textes multilingues en utilisant ces approches.

concernant les méthodes d'apprentissage statistique, nous avons présenté et discuter les différents algorithmes de classification en mettant l'accent sur l'algorithme des k plus proches voisins, puisque c'est la méthode choisie dans notre travail.

La démarche générale de l'identification de la langue de est constituée deux étapes, à savoir :

Le pré-traitement du texte : cette phase consiste en l'elimination des éléments inutiles du texte (caractères de ponctuation et mots outils) et la représentation du texte dans un format adapté aux algorithmes d'apprentissage; on utilise souvent la représentation vectorielle.

choix d'un algorithme d'apprentissage : Nous avons choisi l'algorithme de K-ppv, qui est un algorithme simple, permet de traiter des données volumineuses et le plus important, ce qu'il donne de bons résultats.

Comme ce travail était de grande importance, on à rencontré plusieurs difficultés telles que : le problème d'absence des travaux publiés sur les sujets qui concerne la reconnaissance de la langue dans un corpus de textes multilingue, s'ajoute la non suffisance du temps, la difficulté d'obtenir de l'information relative au domaine dans le moment opportun et en plus, le domaine de notre sujet de recherche est très vaste.

Malgré tout cela, nous avons entamer ce domaine, et 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 maitrise de la programmation sous delphi, apprendre des nouveaux concepts tels que : la classification automatique de documents , les techniques d'extraction des textes et la reconnaissance de la langue.

Il serait maintenant intéressant de poursuivre cette recherche jusqu'au bout, d'où on propose comme perspectives, d'élargir notre étude pour traiter d'autres langues et avec

plus des corpus de textes plus représentatifs, d'intégrer d'autres techniques et méthodes de classification supervisée et de continuer à utiliser notre méthode d'apprentissage dans la recherche d'information, ainsi que toute autre idée jugée utile, réalisable et bénéfique.

BIBLIOGRAPHIES ET WEBOGRAPHIES

[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] Biskri I., Rompré L., Laouamer L. & Meunier F. (2006). Classification de documents Multimédias : vers une approche générale. In Actes du colloque JADT 2006. Besançon, France.

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

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

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

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

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

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

[19] SIMON RÉHEL, « Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés », Mémoire présenté à la Faculté des études supérieures de l'Université Laval Québec, Canada, Janvier 2005.

[20] Paradis, F. et J. Nie, Filtering contents with bigrams and named entities to improve text Classification, In AIRS, 2005, pp. 135–146.

[21] Tan, C., Y. Wang, et C. Lee, The use of bigrams to enhance text categorization. Inf. Process, Manage. 38(4), 2002, pp.529–546.

[22] Sami. Laroum, Nicolas. Béchet, Hatem. Hamza, Mathieu. Roche, Classification automatique de documents bruités à faible, Manuscrit auteur, publié dans "RNTI: Revue des Nouvelles Technologies de l'Information 1 (2009) 25, 2009.

[23] Radwan JALAM, "Apprentissage automatique et catégorisation de textes multilingues», Thèse de doctorat, Université Lumière Lyon, 2003.

[24] Amel TERKIA DERDRA, Fatima Zahra BENSFIA, «La Représentation Conceptuelle pour la Catégorisation des Textes Multilingue», Mémoire de mastère, Université Abou Bakr Belkaid– Tlemcen, Algérie, Septembre 2012.

[25] Leonardo Rigutini, Marco Maggini et Bing Liu, « An EM based training algorithm for Cross Language Text Categorization», Université de di Siena, Italie, Université Illinois à Chicago, USA, 2005.

[26] Dunning, T. (1994). Statistical Identification of Languages. Technical Report MCCS 94-273, omputing Research Laboratory.

[27] Souter, C., Churcher, G., Hayes, P., Hughes, J., and Johnson, S.(1994). Natural Language Identification Using Corpus-Based Models. Hermes Journal of Linguistics, 13 :183–203.

[28] Grefenstette, G. (1995). Comparing Two Language Identification Schemes. In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95), Rome, Italy

[29] Giguet, E. (1998). Méthode pour l'analyse automatique de structures formelles sur documents multilingues. PhD thesis, Université de Caen, France.

[30] Péry-Woodley, M. P. (1995). Quels corpus pour quels traitements atomatiques ? Traitement Automatique des Langues, 36(1-2) :213–232.

[31] Déjean, H. (1998). Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. In Workshop on Paradigms and Grounding in Natural Language Learning, pages 295–299, Adélaide, Australie.

[32] Beesley, K. (1988). Language Identifier : A Computer Program for Automatic Natural Language Identification on On-Line Text. In Proceedings of the 29th Annual Conference of the American Translators Association, pages 47–54.

[33] Cowie, J., Ludovik, E., and Zacharski, R. (1998). An Autonomous,Web-based, Multilingual Corpus Collection Tool. In Proceeding of Natural Language Processing and Industrial Applications, pages 142–148, Moncton, Canada.

[34] Cavnar,W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 161–175, Las Vegas, US. Available from World Wide Web : http://www.nonlineardynamics.com/trenkle/papers/sdair-94-bc.ps.gz.

[35] Sibun, P. and Reynar, J. (1996). Language Identification :Examining the Issues. In Symposium on Document Analysis and Information Retrieval,pages 125–135, Las Vegas.

[36] Benzecri, J. P. (1973). L'Analyse des Données, volume 1. Dunod,Paris.

[37] Rajman, M. and Lebart, L. (1998). Similarités pour données textuelles. In 4th International Conference on Statistical Analysis of Textual Data (JADT'98), pages 545–555, Nice, France.

[38] Kadri Youssef, «Recherche d'information translinguistique sur les Documents en Arabe », thèse en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.) en Informatique, Université Montpellier, France, septembre 2008.

[39] Hadjer SAHNOUN, Kais HADDAR, « Étude comparative des techniques de la traduction automatique et leurs expérimentations sur les entités nommées avec NOOJ », Laboratoire MIRACL, FSS, 2009.

ملخص :

العمل المنجز في إطار هذه المذكرة يتمثل في تطوير نظام آلي يسمح بالتعرّف على لغة نصّ ينتمي إلى مصنّف نصوص متعدّد اللغات و ذلك باستعمال خوارزميات التعلّم الألي التي تعتمد على حساب قياسات التشابه بين النصوص من حيث اللغة أو ما يعرف بحساب المسافات. و قد اخترنا من بين الخوارزميات خوارزمية الجار أو الجيران الأقرب KNN نظرا لسهولة برمجتها و دقة نتائجها.

الكلمات المفتاحية:خوارزمية التعلم ،تحديداللغة ،-Nغرام .

ABSTRACT:

The work realized in this thesis consists in developping an automatic system to recognize the languages of a multilingual collection of texts using learning algorithms and basing on the most known similarity metrics in the field. The learning algorithm adopted in our system is the algorithm of k-nearest neighbor due to its ease of implementation and accuracy of its results.

Key words: Algorithm of training, Language identification, N-gram.

RESUME

Le travail réalisé dans le cadre de ce mémoire de fin d'études consiste à développer un système automatique pour reconnaitre les langues d'une collection multilingue de textes en utilisant les algorithmes d'apprentissage et en basant sur les métriques de similarité les plus connues dans le domaine. L'algorithme d'apprentissage adopté dans notre système est l'algorithme des k-plus proches voisins (k-nearest neighbors KNN) à cause sa facilité d'implémentation et l'exactitude de ses résultats.

Mots clés : Algorithme d'apprentissage, Identification de la langue, N-grammes