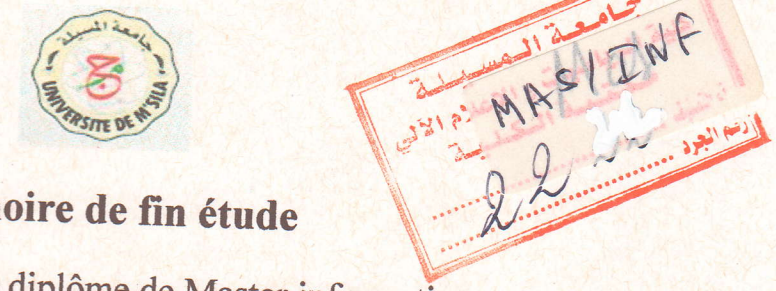
mas/inf 22

République algérienne démocratique populaire

Ministère De L'Enseignement Supérieur et Recherche scientifique

Université de m'sila

Département de l'informatique et mathématique



Mémoire de fin étude

Pour l'obtention de diplôme de Master informatique

Option : système d'information avancée

Thème

Etude des algorithmes de classification supervisés KNN et Naïve Bayes pour la classification des documents textuels

Réalisé par :

YOUCEFI Asma

Encadré par :

Gasmi Abdel Al kader

Bouamama salim

Promotion : 2010/ 2011

Sommaire

Introduction

La recherche 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 au plan international, le développement de l'infrastructure de communication et de l'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.

Parmi les nouvelles technologies qu'on tend à appliquer aux textes, la *fouille de données*, qui ne se limite pas au traitement des données structurées sous forme de tables numériques mais offre des moyens d'investigation des corpus en langage naturel.

Le domaine de la fouille de textes (text 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.

La préparation des données est une étape importante, si ce n'est primordial, du processus d'extraction de connaissances à partir de données. En schématisant, il s'agit de définir au mieux les éléments traités et la représentation utilisée pour l'apprentissage. La qualité (du modèle) de prédiction dépend donc grandement de la qualité de la préparation effectuée en amont.

La préparation consiste à homogénéiser les données et à les disposer en tableau lignes/colonnes, car il s'agit presque toujours de la structure la mieux adaptée à l'exploitation des données. Formellement, chaque ligne/colonne peut être considérée comme un objet vecteur ayant un nombre fixe de composantes. Ce vecteur ligne/colonne sera vu comme un objet mathématique que l'on pourra manipuler selon ses propriétés.

Nous nous intéressons ici plus particulièrement aux algorithmes d'apprentissage et nous avons utilisé et comparé deux algorithmes : les K plus proches voisins (KNN) et

l'algorithme NAIVE Bayes. Pour pouvoir utiliser de tels algorithmes, il est nécessaire de transformer les données, initialement en format texte, en une représentation numérique. Une fois ce prétraitement terminé, nous pouvons effectuer la classification à l'aide de nos algorithmes ; ensuite nous ferons une étude comparative entre les deux algorithmes.

Notre travail s'intéresse à l'analyse de données textuelles pour les préparer à la classification par les techniques de fouille de données et faire une comparaison entre les deux algorithmes. Il comporte quatre chapitres et une brève conclusion avec une perspective.

Le premier chapitre est une présentation générale de la fouille de données (Data Mining), et la fouille de textes(text mining).

Le deuxième chapitre présente la notion de la classification et la catégorisation des textes.

Le troisième chapitre détaille les deux algorithmes KNN et NAIVE BAYES.

Le quatrième chapitre est consacré à l'implémentation.

Conclusion et perspectives

Les techniques du data mining ont pour objectif de satisfaire la nécessité de classer des nouveaux documents. Ce projet a pour objectif principal d'étudier et d'implémenter deux algorithmes de Data Mining. Ces algorithmes dits des algorithmes de classification supervisée.

Apres avoir réalisé les mesures expérimentales, et interprété les résultats donnés par KNN et ceux donnés par NAIVE Bayes. On montre tantôt que Naïve Bayes est plus performant et plus efficace que KNN tantôt c'est l'inverse.

L'algorithme NAIVE Bayes est simple, en vue des résultats obtenus, on peut conclure que l'algorithme Naïve donne de bons résultats avec un bon taux de précision, l'apprentissage est rapide, permettant de traiter des données volumineuses, il convient donc particulièrement bien au problème de la fouille de textes.

Il serait maintenant intéressant de poursuivre cette recherche, qui a permis de découvrir les avantages de ces algorithmes dans le cadre de la catégorisation de texte. Nous espérons que cette contribution pourra ouvrir de nouvelles perspectives à d'autres collègues et d'ouvrir un champ de recherche.

A mon avis je prépose de faire encore une étude approfondie et utilise un grand nombre de test pour trancher et décider.

Bibliographe

[1]. Rachel Konrad, 2001, *Data Mining: Digging user info for gold*, *ZDNET News*, February 7, http://news.zdnet.com/2100-9595_22-528032.html?legacy=zdnn

[2]. The Technology Review Ten, *MIT Technology Review*, January/February 2001.

[3]. http://www.grappa.univ-lille3.fr/polys/fouille/sortie005.html

[4]. http://cedric.cnam.fr/~saporta/DM.pdf

[5]. http://eric.univ-lyon2.fr/~ricco/cours/slides/Introduction_au_Data_Mining.pdf

[6]. http://www.univ-paris1.fr/diplomes/master-droit-du-numerique/bibliotheque-numerique-du-droit-de-ladministration-electronique/tic/informatique/data-mining/data-mining-definition/

[7]. http://www.ultra-fluide.com/ressources/datamining/presentation.htm

[8]. Sadik Bessou, Analyse de Données Textuelles pour la Classification Automatique par les Techniques de Text Mining, application à la Langue Arabe, Thèse de magister, université Ferhat Abbas SETIF (2006 /2007).

[9]. Rémi Gilleron & Marc Tommasi, 2000, *Découverte de connaissances à partir de données*, http://www.grappa.univ-lille3.fr/polys/fouille/sortie005.html

[10]. http://www.madchat.fr/reseau/ids%7Cnids/data_mining-fr.htm

[11]. http://www.commentcamarche.net/forum/affich-7099472-data-mining-et-text-minig

[12]. http://www.datalgo.com/60-text-mining.htm

[13]. http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/fich_art/Patrice/Bellot.pdf

[14]. http://translate.google.com/translate?hl=fr&langpair=en%7Cfr&u=http://www.crisp-dm.org/Process/index.htm

[15]. http://translate.google.com/translate?hl=fr&langpair=en%7Cfr&u=http://www.crisp-dm.org/Process/index.htm

[16]. Radwan jalam, Apprentissage automatique et catégorisation de textes multilingues , thèse de doctorat, université lumière lyon2, 2003

[17]. Sayad M et Abbari A, Modèle discriminant pour la classification de documents XML à l'aide des réseaux bayésiens et le noyau de Fisher, thèse ingénieure, INI, 2009/2010.

[18]. Belacel nabil, Méthodes de Classification Multicritère Méthodologie et Applications à l'Aide au Diagnostic Médical, Thèse doctorat, Université LIBRE DE BRUXELLES, 1999/2000.

[19]. Saeed Raheel, L'apprentissage Artificiel pour la Fouille de Données Multilingues : Application à la Classification Automatique des Documents Arabes, thèse doctorat, Université Lumière Lyon2, 22 octobre 2010.

[20]. Abidi karima, La catégorisation de texte multilingue, thèse de magistère, INI, 2010/2011.

[21]. http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm

[22]. Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés fichiers.

التلخيص

التصنيف الاولي للوثائق اصبح ضروريا بسبب حجم الوثائق المتبادل والمخزين إلكترونيا. ان تعدد الوثائق وتزايد عددها المستمر ، نتج الصعوبة في التخطيط مقدما لقواعد قرار لتحديد فئة مستند جديد .ونحن هنا قدمنا طرق التعليم المتمثلة في(Naïve Bayes et KNN))التي، من خلال مستند مصنف مسبقا, تسمح لنا بتصنيف مستند جديدة. معظم خوارزميات التعلم المراقب تسعى لإيجاد نموذج يفسر الارتباط بين المعطيات الاولى والاصناف. بعد التجارب و ترجمه النتائج المحصل عليها اصبح جلي ان NAIVE Bayes احيانا اكثر فعالية من KNN احيانا العكس .

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 **((KNN) K plus proche voisin et Naïve Bayes)** qui, à partir de documents déjà classés, permettent de classer de nouveaux documents.

Apres avoir réalisé les mesures expérimentales, et interprété les résultats donnés par KNN et ceux donnés par NAIVE Bayes. Tantôt que Naïve Bayes est plus performant et plus efficace que KNN tantôt c'est l'inverse.

Mots clés : documents textuels, catégorisation, sac de mot, KNN, Naïve Bayes.

Abstract

Automatic classification supervised document becomes necessary because of the volume of documents exchanged and stored electronically. As the documents are numerous or their number is growing, it would be difficult to plan in advance of decision rules for determining the class of a new document. We therefore present methods of learning ((KNN) K nearest neighbor and Naive Bayes) which, from documents already classified for classifying new documents. Conducted experiments showed that NAÏVE BAYES classifier had better performance than KNN classifier sometimes KNN had better than Naïve Bayes.

Keywords: textual documents, categorization, Bag-of-words, KNN, Naive Bayes.