تأثير مربع كاي المطور (كطريقة اختيار الخصائص) على مصنفات النصوص العربية

إعداد
هديــل نعيم الشاعـــــر

إشراف

الاستاذ الدكتور محمد عبد الله عطير

الملخص

يمكن تعريف تصنيف النص على أنه طريقة توزيع النصوص إلى مجموعات محددة مسبقًا إعتمادا على محتوياتها. في السنوات القليلة الماضية ، تضاعف حجم المعلومات في مجالات المتنوعة عبر الإنترنت ، مما جعل تصنيف النصوص أحد أهم القضايا ، حتى مع صعوبتها. يُستخدم تصنيف النصوص بشكل كبير في العديد من التطبيقات ولأهداف مختلفة.

إن الاستخدام الكبير والموسع للإنترنت ، لا سيما في العالم العربي ، فضلاً عن العدد الهائل للوثائق والصفحات التي توفرها باللغة العربية ، زادت من الحاجة إلى وجود أدوات مناسبة لتصنيف هذه الصفحات والوثائق حسب أصنافها و مواضيعها .

إن الهدف من هذه الرسالة هو دراسة تأثير طريقة مربع كاي المحسنة (impCHI) على أداء ستة مصنّفات نصية عربية مشهورة هي Random Forest :، Decision Tree، Naïve Bayes، Naïve Bayes Multinomial، Bayes Net، و Artificial Neural Networks.التقنيات المقترحة جميعها تعتبر جديدة ومهمة للغاية لتحسين تصنيف النصوص العربية ويمكن اعتبارها أساسًا واعداً لمرحلة تصنيف النص لأنها تساهم في تصنيف النصوص إلى فئات محددة مسبقًا.

تضم قاعدة البيانات التي استخدمت في هذه الرسالة 9055 وثيقة عربية تم جمعها من مصادر عربية مختلفة. وبناءً على محتواها ، تم تقسيم هذه الوثائق إلى اثني عشر فئة مختلفة. و فيما يتعلق بالأداء ، فقد تم استخدام اربعة معايير لتقييم الأداء: F-measure، recall،precision ، و Time build model. أظهرت نتائج التجارب أن استخدام مربع كاي المحسّن يعطي نتائج تصنيف أفضل من طريقة كاي التقليدية و ذلك مع جميع المصنفات التي تم دراستها، و ذلك طبقا إلى جميع معايير الأداء المستخدمة.

The Effect of Improved CHI Square (As a Feature Selection Method) on Arabic Text Classifiers

Prepared by:
Hadeel N. Alshaer

Supervised by:
Prof. Mohammed A. Otair

Abstract

Text classification could be defined as the way of allocating text into predefined groups according to its contents. Over the past few years, an increase emerged in the volume of information in the varied fields on the Internet, thus making classification of texts one of the most important, yet challenging. Text classification is commonly employed in numerous applications and for different objectives. The extensive and broad use of the Internet, particularly in the Arab world, as well as the massive number of the documents and pages which are provided in the Arabic language raised the need for having suitable tools for classification of these pages and documents by their main categories. The aim of this thesis to study the effect of Improved CHI (impCHI) Square on the performance of six well-known classifiers: Random Forest, Decision Tree, Naïve Bayes, Naïve Bayes Multinomial, Bayes Net, and Artificial Neural Networks.

These proposed techniques are quite important for improving classification of Arabic documents and can be regarded as a promising basis for the stage of text classification because it contributes to classification of the texts into predefined categories. The dataset employed in this thesis comprises 9055 Arabic documents that were collected from various Arabic resources. Based on their content, these documents were divided into twelve categories. Four performance evaluation criteria were used: the F-measure, recall, precision and Time build model. The experimental results show that the use of impCHI square gives better classification results than the normal CHI square method with all studied classifiers, in terms of all used performance criteria.