جامعة الموصل
كلية علوم الحاسوب والرياضيات

تصنيف الوثائق النصية بالاعتماد على محلل التشابه المضبب

رشا وائل علي النحاس

رسالة الدبلوم العالي

علوم الحاسوب

بإشراف

الدكتورة آلاء ياسين طاقة

مدرس

١٤٣٥هـ **٢٠١٤ م**

الملخص

تعد عملية تصنيف الوثائق إحدى المواضيع الأساسية في تعدين النص ونظرا إلى الحاجة المتزايدة لإدارة النمو السريع ومعالجته للمعلومات والتطبيقات على الإنترنيت، وما تتضمنه من معلومات نصية، فقد اكتسب تصنيف النص اهتماما خاصا من قبل الباحثين والمعنيين.

تم اقتراح أسلوب التشابه المضبب في تصنيف الوثائق النصية من نوع (TXT) الذي يعتمد على وجود المضبب لوصف العلاقة ما بين الوثيقة والصنف ، وإيجاد مقدار التشابه بالاعتماد على مجموعة قيم المضبب، وتستخدم هذه القيم لتحديد درجة التشابه ما بين الوثائق والأصناف والتي على ضوئها سيتم تحديد الصنف الذي تنتمي اليه الوثيقة.

تم تدريب واختبار المصنف على مجموعة من الوثائق النصية من نوع (TXT) والتي تم تقسيمها يدويا وحسب محتواها إلى ثلاثة أصناف : الصنف الأول ويشمل الوثائق الصحية (Health) التي تضمنت محتوياتها مواضيع تتعلق بصحة الإنسان، والفوائد الصحية، وكان عددها (٤٠ وثيقة), ويشمل الصنف الثاني الوثائق العلمية (Scientific) التي تضمنت مواضيع ذات طابع علمي، مثل الهندسة، والحاسوب، والعلوم وغيرها، وكان عددها (٤٠ وثيقة)، ويشمل الصنف الثالث وثائق رياضية (Sport) تضمنت مواضيع لها علاقة بالرياضة بواقع (٤٠ وثيقة).

قسمت الوثائق إلى مجموعتين: (مجموعة تدريب) و(مجموعة اختبار) بواقع (٦٠) وثيقة للتدريب (٢٠ وثيقة علمية، ٢٠ وثيقة صحية، ٢٠ وثيقة رياضية) و(٦٠) وثيقة للاختبار(٢٠ وثيقة علمية، ٢٠ وثيقة صحية، ٢٠ وثيقة رياضية)؛ استخدمت مجموعة التدريب لتكوين القاموس المستخدم في اختبار المصنف، بينما استخدمت مجموعة الاختبار في اختبار وتقييم أداء عمل المصنف .

تضمن طور التدريب المعالجةُ الأولية لوثائق التدريب، وهي: (عملية التقطيع وإزالة الضوضاء وحذف كلمات التوقف) ثم أجريت عليها عملية اختيار الصفات، وتم اعتماد طريقة المعلومات المشتركة، وكان الناتج هو عبارة عن ثلاثة قواميس، يضم القاموس الأول: الصفات المختارة من وثائق التدريب الصحية، اما القاموس الثاني فيضم الصفات المختارة من وثائق التدريب العلمية، واشتمل القاموس الثالث الصفات المختارة من وثائق التدريب الرياضية.

كما تم اختبار المصنف الذكائي المقترح من خلال استقبال الوثائق النصية من نوع (TXT) ثم بعد مرورها بكافة مراحل المعالجة الأولية تكون الوثيقة ممثلة بمتجه يسمى (متجه الصفات)، تم ادخال متجه الصفات إلى المصنف الذكائي الذي يقوم بتصنيف الوثيقة إلى إحدى الاصناف (العلمية، الصحية، الرياضية) بالاعتماد على اكبر تشابه.

أثبتت نتائج تقييم أداء المصنف باستخدام المقياسين Accuracy و Recall بانها جيدة، إذ إن قيمة Accuracy تساوي (٩٠٪) وقيمة Recall تساوي (٩٠٪).

UNIVERSITY OF MOSUL
COLLEGE OF COMPUTER SCIENCES
AND MATHEMATICS

Textual Documents Classification Based on Fuzzy Similarity Analyzer

Rasha Wail Ali Al-Nahhas

Higher Diploma

Computer Science

Supervised by

Dr. Alaa Yaseen Taqa

Lecturer

2014 A.D. **1435 A.H.**

Abstract

Documents classification is regarded as a basic subject in Text Mining. Because of the increasing need to manage the rapid growth and its ability in dealing with the information and applications of the internet and because it includes textual information , the classification of text acquires a special attention by the researchers .

The Fuzzy Similarity is Proposed in classifying the textual documents of type (TXT) which is based on the fuzzy value to describe the relationship between the document and the type as well as finding the degree of similarity depending on the fuzzy values. These values are used to know the degree of similarity between the documents and the type to which the document is related.

The classifier was trained and tested to deal with a group of textual documents of type (TXT) which were manually classified into three types: the first type is the health type which contains subjects related to the health of human being as well as the health advantages whose number is (40) documents. The second type involves the scientific type which contain the scientific subjects like engineering, computer and sciences whose number is (40) documents. While the third type is the sports type which contain subjects related to sport activities whose number is (40) documents.

The documents are divided into two groups (training group and testing group) the training set consists of (60) documents (20 scientific document, 20 health document, 20 sports document), while the testing set consists of (60) documents (20 scientific document, 20 health document, 20 sports document). The training set was used to build the dictionary used in the training of the "classifier", whereas the testing set was used in testing and evaluating the performance of the classifier work.

The training process includes the preprocessing of the training documents which are (the Tokenization process, Noise removal, Stop word removal) then the

Features Selecting was done by Mutual Information, three dictionaries were resulted from this process, the first one contains the characteristics chosen from the health training documents, the second contains the characteristics chosen from the scientific training, and the third includes the characteristics chosen from the sport training documents.

The proposed intelligent classifier is also tested by receiving the textual documents of type (TXT). Then, after finishing all the preprocessing stages, the document is represented by a vector which consists of values called (Features Vector) and then this vector is entered into the intelligent classifier which classifies the document to one of the types (health, scientific and sports) depending on the highest similarity.

The evaluation of the classifier performance using Accuracy and Recall measurements shows good results where the value of Accuracy equal (90%), while the value of Recall equal (90%).