Automatic Keywords Extraction
Using Combined Methods

A Dissertation
Submitted to the Department of Computer Sciences in the
University of Technology as a Partial Fulfillment of the
Requirements for the Degree of Doctor of
Philosophy in Computer Science

By
Rafiq Abdul-Rahman A. Al-Hashimi

Supervised By
Prof. Dr. Hilal H. Saleh & Dr. Ahmed T. Sadiq

2006
Abstract

The popularity of the web and the large number of documents available in electronic form bring the advantage of reaching the information sources in a cheap, quick way and relevant. Keywords are useful tools as they give the shorter summary of the document.

Keywords are useful for a variety of purposes including summarizing, indexing, labeling, categorization, clustering, and searching.

The task of the proposed system, 'Automatic Keywords Extraction', (AKWE) is to extract keywords from the document abstract. The proposed system solves this problem through many statistic and linguistic approaches by three stages. The entered document is firstly pre-processed to remove noisy data, word tagging, and word stemming. Secondly to give candidate keywords, three extracting approaches presented in the proposed system, N-gram approach to extract uni-gram, bi-grams and tri-grams; part-of-speech approach (POS) that extracts phrases which match a set of patterns, and NP-chunk which extract noun phrases. The proposed system uses a scoring system to give a weight for each candidate keyword depending on many features (Term frequency, Inverse Document Frequency (IDF), Word tag, KDF, Font, position and existence in the title). The system combines these three approaches to give more important keywords; in other words the keywords that majority voted by the three approaches and the highest keyword weights are selected.

The proposed system uses document classification as a subsystem that classifies the document in order to find out the informative keyword than the trivial words that are frequently used in the class.
In the last stage (post-processing stage) the system excludes noise (negative) keywords and reduce them through using more than one filter by removing subsumed keywords that exist within other keywords, exists in non-keywords list, removing replicated once and mix keywords intersect with others in more that one word.

The proposed system also presents a new approach to use rules mined from extracted keywords database to improve the accuracy of keyword extraction, i.e. it integrates data mining with keyword extraction.

From the implementation and the proposed system experiments good results are obtained. About 74% for automatic keyword extraction was achieved compared with keywords extracted manually by document author. The system attained more than 93% of text classification accuracy.
الملخص

لا يمكن قراءة النص بشكل طبيعي من الصورة المقدمة.
قد وجدت القوة في الأغراض الأخرى تمامًا. إن هذه الكلمات ككلمة، بائها، بعضاً مع تناقص.

قد توفر النظام هذا جدًا طريقة المحكمة، والإعداد، والبحث، وقرارات تحسين وفاعلية، وعمل النظام. (هذا المفتاح) يمكنه سلامة والتطبيق الفعال.

تير 74% من الأوراق نفذ DNA ونفذ T2 في 55 دقيقة إنها نتاحة. وجدت تير 2! الأوراق، ونفذ DNA ونفذ T2، ونفذ DNA ونفذ T2. هذه T2 ك05% ثانية.
استخلاص الكلمات المفتاحية آلياً
بدمج طرق متنوعة

أطروحة مقدمة إلى قسم علوم الحاسبات في الجامعة التكنولوجية وهي كجزء من متطلبات نيل شهادة دكتوراه في علوم الحاسبات

تقدم بها
رفق عبد الرحمن الهاشمي

بإشراف
الدكتور احمد طارق صالح

الأستاذ الدكتور هلال هادي صالح

2006