

**1. معلومات عامة**

اسم الطالب ـأمجد كميل توفيق ايوب الرقم الجامعي: 51301001015

التخصص:ـ علم الحاسوب الكلية: الأمير عبدالله بن غازي لتكنولوجيا المعلومات

تاريخ الالتحاق بالجامعة: الفصل الاول ـ العام الجامعي ـ2013/2014

اســم المشــرف عــلى الرسالــة:ــد. معاذ رفعات الزغول التوقيع ــــــــــــــــــــــــــــــــــــــ

اسم المشرف المشارك (أن وجد): ـــــــــــــــــــــــــــــــــــــــــــــــــــــــــــ التوقيع ــــــــــــــــــــــــــــــــــــــ

تاريخ إقرار عنوان رسالة الماجستير وتعيين المشرف من قبل مجلس الدراسات العليا: ــــــــــــــــــــــــــــــــــــ

تاريخ تسجيل ساعات الرسالة :ـ ـــــــــــــــــــــــــــــــــــــــــــــــــــــــــــــــ

**2. عنوان الرسالة كما أقر من مجلس كلية الدراسات العليا:**

1. باللغة التي ستكتب بها الرسالة:

**A Fusion-based Extraction of Key Phrases from Abstracts of Scientific Articles**

1. مترجمة إلى العربية أو الإنجليزية:

**استخراج الجمل المفتاحيه المعتمد على اندماج ملخصات المقالات العلميه**

**3- مشكلة الدراسة وأهميتها:**

( يعطى الطالب فكرة عامة عن موضوع الدراسة والأساس النظري الذي يقوم عليه والأسئلة أو الفرضيات التي ستحاول الدراسة الإجابة عليها).

Keywords can be defined as the words that represent the content of the whole text, which can help users to identify what the article is about without having to read the entire document. Filtering documents using “keywords” can be regarded as very short summaries, which may help to save time while searching.

Keywords extraction is an important technology in many areas of information technology and document processing such as document tagging, text categorization, text summarization, and text retrieval. The main aim of any keywords extraction system is to identify the most informative and important words in the document.

The task of keywords extraction application is to automatically identify in a text a set of terms that best describe the document, the simplest possible approach is perhaps to use a frequency criterion to select the “important” keywords in a document. [1] However, this method was generally found to lead to poor results. Consequently, other methods were explored. [2]

Early approaches to automatically extract keywords focus on evaluating corpus-oriented statistics of individual words. Later keyword extraction research applies these metrics to select discriminating words as keywords for individual documents. Corpus-oriented methods also typically operate only on single words. This further limits the measurement of statistically discriminating words because single words are often used in multiple and different contexts. [1] To avoid these drawbacks, this research interests on methods of keyword extraction that operate on individual documents; such document-oriented methods will extract the same keywords from a document regardless of the current state of a corpus.

The main objective of this research is to extract keywords from abstracts using a fusion-based technique from three other original techniques, which all shared the same dataset “Hulth 2003”, as training and testing dataset. The first one is a Rapid Automatic Keyword Extraction, (RAKE). It was founded by Stuart R., Dave E., Nick C. and Wendy C., 2010. RAKE is based on keywords usually containing multiple words, but rarely standard punctuations or stopwords. This technique uses stopwords and word delimiters to partition the documents into candidate keywords which will be given scores for each, based on some calculations. Lastly, extracted keywords will be produced based on their scores.

The second technique is Text Rank. This technique was established by Mihalcea R.and Tarau P., 2004. Text Rank is based on the co-occurrence links between words. It makes use of voting or recommendation between words to extract keywords. The technique is first implemented by constructing a graph which reflects relationship between different vertices (words), which will be extracted from the given texts, and then using ranking algorithm to calculate score of words. Finally, highest score of words in the document will be chosen as keywords after an iterative algorithm is used to compute the ranking value of each vertex of the graph.

The third technique Hulth-2003 N-grams approach which was introduced by Anette Hulth 2003. All unigrams, bigrams, and trigrams were extracted. Thereafter a stop list was used, where all terms beginning or ending with a stopword were removed. Finally all remaining tokens were stemmed using Porter’s stemmer. (Porter, 1980) [3]

As mentioned above, this research will introduce the fusion-based technique. The techniques above produce their results as lists of extracted keywords. This new technique integrates and calculates the results of those three techniques to produce its’ own results of more accurate keywords than the keywords that are produced individually.

## 4- أهداف الدراسة ومبرراتها:

Automatic keywords extraction has a main task of identifying a small set of words, key phrases, keywords, or key segments from a document. Those keywords describe the meaning of the document. The process of identifying keywords should be done systematically and with minimal or no human intervention, depending on the model. The goal of automatic keyword extraction is to apply the power and speed of computation to the problems of access and discoverability, adding value to information organization and retrieval without the significant costs and drawbacks associated with human indexers. [3].

Keyword search is a usable and powerful tool which enables efficient scanning of large document collections. Keywords of any document provide important information about the content of the document which makes up the meaning of the document. Keywords can be used by users to search through information more efficiently or to decide whether to read a document. Keywords can also be used for a variety of language processing tasks such as text categorization, and information retrieval.

Keywords can also be used to enrich the presentation of search results. They are also applied to improve the functionality of IR systems. A system that lists documents related to a primary document’s keywords, and that supports the use of keyword anchors as hyperlinks between documents, enabling a user to quickly access related material. [Jones and Paynter, 2002].

**5**- الدراسات السابقة:

( يعـرض الطـالب عينــة مـن الدراسـات السابقـة النظرية والتطبيقية ذات العلاقـة بمـوضوع البحـث مبــاشرة وكيف تختلف هذه الدراسة عن السابقة)

Extracting keywords manually is too slow, expensive, and full of mistakes. Therefore, most algorithms and systems that help users to perform automatic keyword extraction have been proposed.

Previous work on document-oriented methods of keyword extraction combines natural language processing approaches to identify part-of-speech (POS) tags that are incorporated with either supervised learning, machine-learning algorithms, or statistical methods. [1]

Existing methods can be divided into four categories: simple statistics, linguistics, machine learning, and mixed approaches [4]. These methods are illustrated below:

**Simple Statistics Approaches**

Simple Statistics methods are simple, have limited requirements, and need no training data. Those methods aim to focus on non-linguistic features of the text such as term frequency, inverse document frequency, and position of a keyword. The statistic information of words can be used to identify the keywords in the document. Statistics methods include word frequency, word co-occurrences [5], etc. The purely statistical methods have some advantages. For instance their ease of use and the fact that they generally produce good results.

**Linguistics Approaches**

The Linguistics approaches use the linguistic features of the words, sentences, and documents. Methods that focus on linguistic features such as part-of-speech, syntactic structure and semantic qualities tend to add value, functioning sometimes as filters for bad keywords. During automatic keyword extraction, the advantages of using the lexical resources are compared to a pure statistical method and a relative frequency ratio.

Hulth examines few different methods of incorporating linguistics into keyword extraction. Terms are tested as keywords based on three features: document frequency, collection frequency, relative position of its’ initial occurrence in a document and the term's part of speech tag. The results show that the use of the linguistic features represent the remarkable improvement of the automatic keyword extraction. [3] Indeed, some of the linguistic methods are mixed methods, which combine some linguistic methods with common statistical measures such as term frequency and inverse document frequency.

**Machine Learning Approaches**

The machine learning mechanism works as follows. First a set of training documents is provided to the system, each of the documents has a range of human-chosen keywords. Soon after, the gained knowledge is applied in order to identify keywords from new documents. The Key phrase Extraction Algorithm uses the machine learning techniques and naive Bayes formula for domain-based extraction of technical key phrases. [6][10][11]

**Mixed Approaches**

Other approaches for keyword extraction essentially combine the methods that were mentioned above, or use some heuristic knowledge in the task of keyword extraction like the position, length, layout feature of the words, html tags around of the words, etc.

**Rapid Automatic Keyword Extraction (RAKE)**, an unsupervised, domain-independent and language-independent method used for extracting keywords from individual documents. Authors provide details of the algorithm and its configuration parameters, and present results on a benchmark dataset of technical abstracts, showing that RAKE is more computationally efficient in achieving higher precision and comparable recall scores than TextRank. Next, authors describe a novel method for generating stoplists. Stoplists are used to configure RAKE for specific domains and corpora. Finally, they apply RAKE to a corpus of news articles and define metrics for evaluating the exclusivity, essentiality, and generality of extracted keywords, enabling a system to identify keywords that are essential or general to documents in the absence of manual annotations.

**TextRank for Keyword Extraction**

The expected end result for this application is a set of words or phrases that represent a given natural language text. The units to be ranked are sequences of one or more lexical units extracted from text, which represent the vertices that are added to the text graph. Any relation that can be determined between two lexical units is a potentially useful connection (edge) that can be added between two such vertices. Authors used a co-occurrence relation, controlled by the distance between word occurrences; two vertices are connected if their corresponding lexical units co-occur within a window of maximum words, where can be set anywhere from 2 to 10 words. Co-occurrence links express relations between syntactic elements, and similar to the semantic links found useful for the task of word sense disambiguation [2], they represent cohesion indicators for a given text.

The vertices added to the graph can be restricted by syntactic filters, which select only lexical units of a certain part of speech. One, for instance, can consider only nouns and verbs in addition to the graph, and consequently draw potential edges depending on relations that can be established between nouns and verbs. Authors experimented with various syntactic filters, including: all open class words, nouns and verbs only, etc., with best results observed for nouns and adjectives only. [7]

**N-gram approach**

In a first set of runs, the terms were defined in a manner where all unigrams, bigrams, and trigrams were extracted. Thereafter a stoplist was used, where all terms beginning or ending with a stopword were removed. Finally all remaining tokens were stemmed using Porter’s stemmer (Porter, 1980). In this paper, this manner of selecting terms is referred to as the N-gram approach.[9] The overview of the related works reveals that the automatic keyword extraction is faster and less expensive than human intervention.

**6**- منهجية البحث:

(يحدد الطالب المنهج/المناهج التي ستستخدم في دراسة البحث ويعطى الطالب فكرة عن أسلوب البحث الذي سيتم استخدامه ومجتمع الدراسة والعينة وأدوات جمع المعلومات البيانات وطرق تحليلها وإجراءات الدراسة، وذلك حسب نوعية الدراسة بالإضافة لجدول زمني لمراحل الإنجاز).

**The Corpus**

The dataset “Hulth 2003” consists of 2000 Inspec abstracts for journal papers from Computer Science and Information Technology. The abstracts are divided into a training set with 1000 abstracts, a validation set with 500 abstracts, and a testing set with 500 abstracts

The abstracts are from the years 1998 to 2002, from journal papers, and from the disciplines Computers and Control, and Information Technology. Each abstract has two sets of keywords—assigned by a professional indexer—associated to them: a set of controlled terms, i.e., terms restricted to the Inspec thesaurus; and a set of uncontrolled terms that can be any suitable terms.

Both the controlled terms and the uncontrolled terms may or may not be present in the abstracts. However, the indexers had access to the full-length documents when assigning the keywords. For the experiments described here, only the uncontrolled terms were considered, as these to a larger extent are present in the abstracts (76.2% as opposed to 18.1%).[3]

**Rake implementation**

1. Partition the document based on the generated stop words, word delimiters, and phrase delimiters into candidate keywords;
2. Calculate keyword score; a. Word Frequency: the frequency of candidate keywords. b. Word Degree: favor words than occur more often and in longer candidate keywords;
3. Calculate the ratio of degree to frequency. The words with higher ratio are the extracted keywords;
4. Adjoin keywords: when a stopword occurs within two candidate keywords, it is considered adjoining keyword, if and only if it is occurred at least twice in the document;
5. Final step, calculate T, which is the top T scoring candidates are selected as keywords for the document. T is computed as one-third the number of words in the document.

**TextRank implementation**

To enable the application of PageRank algorithms to natural language texts, a graph will be built to represent the words in the text after removing stop words, and interconnects words or other text entities with meaningful relations. Regardless of the type and characteristics of the elements added to the graph, the application of PageRank algorithms to natural language texts consists of the following main steps:

1. Clean the tested text from useless data;
2. Identify the cleaned text as a set of tokens and add them as vertices in the graph;
3. Identify relations that connect text tokens, and use these relations to draw edges between vertices in the graph. Edges can be directed or undirected, weighted or non-weighted;
4. Repeat the PageRank algorithm until convergence;
5. Sort vertices based on their final score. Use the values attached to each vertex for ranking/selection decisions;
6. Finally, evaluate and compare the automatically extracted keywords with the manually suggested dataset of keywords.

**N-grams implementation**

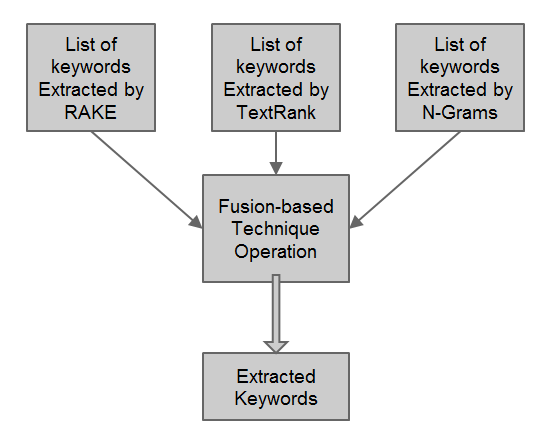
1. Only non-alphanumeric characters that were not present in any keyword in the training set were removed; (keeping e.g., C++)
2. Numbers were removed only if they stood separately (keeping e.g., 4YourSoul.com).
3. Proper nouns were kept;
4. The stemming and the stoplist applied were different;
5. The stems were kept even if they appeared only once (which is true for 80.0% of the keywords present in the training set).

**Fusion-based Technique implementation**

1. After making the three lists of extracted keywords from the techniques discussed above, the fusion-based technique can be implemented;
2. Select potential keywords from the lists. We can select the most or least mutual keywords among the lists;
3. Assign weights to the potential keywords based on their semantic properties and the relationships between them;
4. Use an existing algorithm based on the assigned weights, to get more relevant and accurate list of fusion-based extracted keyphrases.

**Figure 1. Summarizes the system environment of the proposed study.**

Figure 1. Summarizes the system environment of the proposed study.



**Table 1. Summarizes the different phases of the proposed study.**

Table 1: A Time-Table for the Different Phases of the Current Study

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of phases** | **Phases** | **Activities** | **Times in weeks** |
| Phase 1 | Surveying | Review different automatic keywords extraction systems that have been developed | 5 |
| Phase 2 | Implementation of RAKE | Implement and design the system | 4 |
| Phase 3 | Implementation of TextRank | Implement and design the system | 4 |
| Phase 4 | Implementation of N-grams | Implement and design the system | 4 |
| Phase 5 | Implementation of the  Fusion-based technique | Implement and design the system | 6 |
| Phase 6 | Testing | Test and evaluate the automatic keywords extraction systems. | 5 |
| Phase 7 | Comparing | Comparing the new result with the manually extracted results. | 2 |
| Phase 8 | Documenting | Final revision and writing the thesis using Latex. | 6 |
| **Total** |  |  | **36** |

**7**- أية ملاحظات أخرى: (تقتضيها طبيعة البحث)

8- المصادر والمراجع:

( يضع الطالب مجموعة مختارة من المصادر بأشكالها المختلفة والمراجع من الكتب والمقالات والدوريات والمخطوطات والوثائق التي ستتم الاستفادة منها في البحث) ويتم ترتيبها حسب أسلوب التوثيق المعتمد من كلية الدراسات العليا.

**[1]** Stuart, R., Dave, E., Nick, C.and Wendy, C. (2010) “*Automatic keyword extraction*

*from individual documents*”. Text Mining: Applications and Theory edited by Michael W. Berry and Jacob Kogan, pages 3-20.

**[2]** Mihalcea, R., and Tarau, P. (2004). *“TextRank – Bringing Order into Texts”*. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain. Volume 4, Issue 4, pages 404–411.

**[3]** Hulth, A. (2003). *“Improved Automatic Keyword Extraction Given More Linguistic Knowledge”*. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Japan. Pages 216-223.

**[4]** Liu Z., Liu J., Yao W. and Wang C., (2010). *“Keyword Extraction Using PageRank on Synonym Networks “*. E-Product, E-Service and E-Entertainment (ICEEE), International Conference. Pages 1-4.

**[5]** Matsuo Y., Ishizuka M., (2004). *“Keyword extraction from a single document using word co-occurrence statistical information”*. International Journal on Artificial Intelligence Tools. Volume 13, pages 2004.

**[6]** Witten I., Paynte G., Frank E., Gutwin C. and Manning C., (1999). *"KEA:Practical Automatic Keyphrase Extraction".* In Proceedings of the 4th ACM Conference on Digital Library.  Pages (254-255).

# **[7]** Wei Y., (2012), "*An Iterative Approach to Keywords Extraction".* In Proceedings of ICSI (2). Volume 7332, pages 93-99.

**[8]** Hulth A 2004 “*Combining machine learning and natural language processing for automatic*

*keyword extraction”*. Stockholm University, Faculty of Social Sciences, Department of

Computer and Systems Sciences (together with KTH).

**[9]** C.Y. Lin and E.H. Hovy. 2003. Automatic evaluation of summaries

using n-gram co-occurrence statistics. In Proceedings of Human Language Technology Conference (HLT-NAACL 2003)**),** Edmonton, Canada, May

**[10]** P. Turney. 1999. Learning to extract keyphrases from text. Technical report, National Research Council, Institute for Information Technology.

**[11]** Peter D. Turney. 2000. Learning algorithms for keyphrase extraction. Information Retrieval,**2(4):303–336.**

**اسم الطالب / الطالبة: أمجد كميل توفيق ايوب**

**\* توصية لجنة الدارسات العليا في القسم**

**¨ موافق ¨ غير موافق**

**رقم الجلسة ـــــــــــــــــــــــ تاريخها: / /**

**توقيع رئيس اللجنة: ــــــــــــــــــــــــــــــــــ**

##### **\* تنسيب لجنة الدراسات العليا في الكلية المعنية**

**¨ موافق ¨ غير موافق**

**¨ ملاحظات:**

**رقم الجلسة ـــــــــــــــــــــــ تاريخها: / /**

**توقيع رئيس اللجنة: ــــــــــــــــــــــــــــــــــ**

**\* قــرار مجـلس الدراسـات العليـــــا**

##### ¨ **موافق** ¨ **غير موافق**

**¨ ملاحظات:**

**رقم الجلسة ـــــــــــــــــــــــ تاريخها: / /**

**رقم القرارـــــــــــــــــــــــــــــــــــــــــــــــــــــــ**

**توقيع عميد كلية الدراسات العليا : ــــــــــــــــــــــــــــــــــــــ**

**الشكل العام**

**يكتب مشروع خطة الرسالة باللغة العربية، مع ملخص واضح باللغة الإنجليزية ولا يزيد عن (250) كلمة. أما إذا رغبت الكلية في إستعمال لغة أخرى للكتابة بها، فتكتب بتلك اللغة ويرفق بها ملخص واضح باللغة العربية لا يزيد عن (250) كلمة، ويتضمن الملخص هدف الرسالة وأسئلتها وأدواتها ونتائجها وتوصياتها وأهميتها.**

**إرشادات عامة:**

**مشروع خطة رسالة الماجستير في اللغة العربية**

**· نوع الخط Simplified Arabic عادي**

**· حجم الخط 13**

**· الحدود لكل صفحة علوي وسفلي ويمين وشمال 2.5 سم**

**· عدد الصفحات من 7 إلى 10 صفحة فقط**

**· عدد المصادر والمراجع من الكتب والمقالات والمخطوطات والوثائق المختارة لا يتجاوز 12 مرجع رئيسي مرتبة ترتبا هجائيا.**

**· عدم إضافة مختصر كون مشروع خطة رسالة الماجستير في الأصل أن يكون مختصرا.**

**مشروع خطة رسالة الماجستير في اللغة الإنجليزية**

**· نوع الخط Times New Roman عادي**

**· حجم الخط 12**

**· الحدود لكل صفحة علوي وسفلي ويمين وشمال 2.5 سم**

**· عدد الصفحات من 7 إلى 10 صفحة فقط**

**· عدد المصادر والمراجع من الكتب والمقالات والمخطوطات والوثائق المختارة لا يتجاوز 12 مرجع رئيسي مرتبة ترتبا هجائيا.**

**· عدم إضافة مختصر كون مشروع خطة رسالة الماجستير في الأصل أن يكون مختصرا.**

**صفحة الغلاف لمشروع الرسالة المكتوب باللغة العربية**



**استخراج الجمل المفتاحيه المعتمد على اندماج المقالات العلميه**

**إعداد**

**امجد كميل توفيق ايوب**

**المشرف الرئيسي**

**د. معاذ رفعت الزغول**

**كلية الدراسات العليا**

**جامعة البلقاء التطبيقية**

**السلط- الأردن**

**22,03,2015**

**صفحة الغلاف لمشروع الرسالة المكتوب باللغة الإنجليزية**



**A Fusion-based Extraction of Key Phrases from Abstracts of Scientific Articles**

**By**

**Student (Amjad Kameel Ayoub)**

**Supervisor (Dr. Muath Refat Al-Zghool)**

**Assistant Professor**

**Faulty of Graduate Studies**

**Al-Balqa' Applied University**

**Salt- Jordan**

**22, 03, 2015**