Keyword and Keyphrase Extraction Techniques: A Literature Review

Article *in* International Journal of Computer Applications · January 2015

DOI: 10.5120/19161-0607

CITATIONS
157

READS
8,067

2 authors:

Sifatullah Siddiqi
Jawaharlal Nehru University
11 PUBLICATIONS **193** CITATIONS
SEE PROFILE

Aditi Sharan
Jawaharlal Nehru University
67 PUBLICATIONS **722** CITATIONS
SEE PROFILE

Keyword and Keyphrase Extraction Techniques: A Literature Review

Sifatullah Siddiqi
School of Computer and Systems Sciences

Aditi Sharan
School of Computer and Systems Sciences

ABSTRACT
In this paper we present a survey of various techniques available in text mining for keyword and keyphrase extraction. Keywords and keyphrases are very useful in analyzing large amount of textual material quickly and efficiently search over the internet besides being useful for many other purposes. Keywords and keyphrases are set of representative words of a document that give high-level specification of the content for interested readers. They are used highly in the field of Computer Science especially in Information Retrieval and Natural Language Processing and can be used for index generation, query refinement, text summarization, author assistance, etc. We have also discussed some important feature selection metrics generally employed by researchers to rank candidate keywords and keyphrases according to their importance.

Keywords
Keyword extraction, keyphrase extraction, survey, feature selection, weighting measures

1. INTRODUCTION
Now a days we need to quickly go through large amounts of textual information to find out documents related to our interests and this document space is growing on a daily basis at an overwhelming rate. Now days it is common to store several million web-pages and hundreds of thousands of text files. Analyzing such huge quantities of data can be made easier if we can have a subset of words (Keywords) which can provide us with the main features, concept, theme etc of the document. Appropriate keywords can serve as a highly concise summary of a document and help us in easily organize documents and retrieve them based on their content. Keywords are used in academic articles to give an idea to the reader about the content of the article. In a textbook they are useful for the readers to identify and retain the main points in their mind about a particular section. As keywords represent the main theme of a text, they can be used as a measure of similarity for text clustering.

1.1 What is a Keyword?
International Encyclopedia of Information and Library Science [1] defines "keyword" as "A word that succinctly and accurately describes the subject, or an aspect of the subject, discussed in a document." Both single words (keywords) and phrases (key phrases) may be referred to as "key terms". Manning and Schutze have following to say about phrases in their book Foundations of Statistical Natural Language Processing: "Words do not occur in just any old order. Languages have constraints on word order. But it is also the case that the words in a sentence are not just strung together as a sequence of parts of speech, like beads on a necklace. Instead, words are organized into phrases, groupings of words that are clumped as a unit. One fundamental idea is that certain groupings of words behave as constituents."

1.2 Difference between Keyphrase and Keyword
A keyphrase connotes a multi-word lexeme (e.g. *computer science engineering, hard disk*), whereas a keyword is a single word term (e.g. *computer, disk*). Using single words, as index terms, can sometimes lead to misunderstanding. For example, in phrases like "hot dog", the constituent single words does not have their regular meanings and are thus quite misleading if used as individual indexing terms. Also they may be too general, e.g. words "junior" and "college" are not specific enough to distinguish "junior college" from "college junior". Also, when selected from a controlled vocabulary, keyphrases reduce the problems associated with synonymy and polysemy in natural language.

Humans tend to prefer keyphrases to keywords. The size of the keyphrase depends upon its intented application.

Keywords can be assigned either manually or automatically but the former approach is very time-consuming and expensive. Thus there is a need for automated process that extracts keywords from documents.

Keyword extraction is an important task in the field of text mining. There are many approaches by which keyword extraction can be carried out, such as supervised and unsupervised machine learning, statistical methods and linguistic ones. These approaches are discussed in the next section.

2. APPROACHES FOR KEYWORD EXTRACTION
Broadly speaking there can be different approaches for automatic keyword/keyphrase extraction, each having its own pros and cons, but there are four major methods.

1. **Rule Based Linguistic approaches:** These approaches are generally rule based and are derived from the Linguistic knowledge/features. These approaches may be more accurate but are computationally intensive and require domain knowledge in addition to language expertise. These approaches use the linguistic features of the words mainly sentences and documents. The linguistic approach includes the lexical analysis, syntactic analysis discourse analysis and so on.
2. **Statistical approaches:** These approaches are generally based on linguistic corpus and statistical feature derived from the corpus. Most important advantage of them is that they are independent of the language on which they are applied and hence the same technique can be used on multiple languages. These methods may not give as accurate results compared to linguistic ones, but the availability of large amount of datasets has made it

possible to perform statistical analysis and achieve good results.

3. **Machine Learning approaches:** Machine Learning approaches generally employ supervised learning methods. In these methods keyword are extracted from training documents to learn a model, the model is tested through a testing module. After a satisfactory model is built it is used to find keywords from new documents. This approach includes Naïve Bayes, Support Vector Machine, etc. However, supervised learning methods require a tagged document corpus which is difficult to build. In absence of such a corpus unsupervised and semi-supervised learning methods are used.

4. **Domain specific approaches:** Various approaches can be applied to a specific domain corpus, which exploit the backend knowledge related to the domain (such as ontology) and inherent structure of that particular corpus to identify and extract keywords.

3. MODES OF KEYWORD AND KEYPHRASE GENERATION

There are two fundamental approaches for automatic keyphrase generation:

1. **Keyphrase/keyword assignment**: In this approach, the set of possible keyphrases is bounded by a predefined vocabulary of words. The objective is to find a small set of terms that describes an individual document, independently of the domain to which it belongs.

The advantages are simplicity and consistency. Similar documents can be represented by the same keyphrases and the use of a controlled vocabulary ensures the required scope of document coverage.

Drawbacks of this approach are: (1) It is expensive to create and maintain controlled vocabularies and thus they are not always available. (2) Potential keyphrases occurring in the document are ignored if they are not in the vocabulary.

2. K**eyword/keyphrase extraction**: This approach selects the most significant words present in the document and the selection does not depend on any vocabulary and extracted words are present in the document itself.

The advantages are: (1) No need to create and maintain vocabularies, and (2) important keywords and keyphrases that occur in the document can be selected.

The drawbacks of this approach are: (1) lack of consistency; because similar documents can be represented by different keyphraes and (2) it is tough to select the most relevant keyphrases; i.e., the required scope of the document coverage is not ensured.

4. SOME LINGUISTIC PROPERTIES OF KEYPHRASES

After identifying phrases we can use linguistic approaches to filter out the candidate keyphrases. Keyphrases generally have some linguistic patterns which can narrow down our keyphrase search space. These properties are dependent on the parts of speech (POS) of the phrase constituents [2]. Few of the major linguistic patterns for a phrase in English are:

A N (Adjective Noun), e.g. *Linear function*

N N (Noun Noun), e.g. *Regression coefficients*

A A N (Adjective Adjective Noun), e.g. *Gaussian random variable*

A N N (Adjective Adjective Noun), e.g. C*umulative distribution function*

N A N (Noun Adjective Noun), e.g. *Mean squared error*

N N N (Noun Noun Noun), e.g. *Class probability function*

N P N (Noun Preposition Noun), e.g. *Degrees of freedom*

5. RELATED WORK

5.1 Statistical Approach

G. Salton et al. in 1975 proposed a method [3], discrimination value analysis which ranks the words in the text according to how well they are able to discriminate the documents of a collection from each other. Value of a term in this approach is dependent on variation in average separation between individual documents which results when the given term is assigned for content identification. Words achieving the greatest separation are expected to be the best words.

In 1995, J.D. Cohen proposed an approach to draw index terms from text [4]. It doesn't use any stop list, stemmer, or any language and domain-specific component, allowing for easy application in any language or domain with slight modification. The method utilizes n-gram counts, which results in a function similar and more general than a stemmer.

In 2002, M. Ortuño et al. demonstrated that important words of a text have a tendency to attract each other and form clusters [5]. He argues that the standard deviation of the distance between successive occurrences of a word is such a parameter to quantify this self-attraction.

In 2008, J.P. Herrera et al. tackled the problem of finding and ranking the relevant words of a document by using statistical information referring to the *spatial* use of the words [6]. Shannon's entropy of information was used for automatic keyword extraction. The randomly shuffled text was used as a standard and the various measures used in the original document text were normalized by corresponding measures of random text.

P. Carpena et al. proposed to automatically extract keywords from literary texts through a generalization of the level statistics analysis of quantum disordered systems [7]. They consider frequencies of the words along with their spatial distribution along the text, and is based on the observation that important words are significantly clustered whereas irrelevant words are distributed randomly in the text. No reference corpus is needed in this approach and it is especially suitable for single documents for which no priori information is available.

5.2 Supervised Approach

Turney (developer of *Extractor*) first formulated keyphrase extraction as a supervised learning problem [8]. He argues that, all phrases of a document are potential keyphrases, but only those phrases which match with human assigned keyphrases are the correct keyphrases. Turney uses genetic algorithm and a set of parametric heuristic rules for keyphrase extraction.

KEA (Keyphrase extraction algorithm) was developed by Frank et al. [9]; in this system a classifier is build based on the

Bayes' theorem from training documents, and then it is used to extract keyphrases from new documents. KEA analyzes the input document on orthographic boundaries e.g. punctuation marks, newlines etc. to find candidate phrases. Two features are utilized: tf-idf and first occurrence of the term.

Song et al. (2003) proposed a system called KPSpotter which combined Information Gain with several Natural Language Processing techniques, such as First Occurrence of Term and Part of Speech [10]. WordNet was incorporated into KPSpotter to improve extraction accuracy.

Hulth (2003) used linguistic knowledge (i.e., *part-of-speech tags*) to determine candidate sets: potential pos-patterns were used to identify candidate phrases from the text [11]. It was shown that, using a *pos tag* as a feature in candidate selection results in a considerable improvement in the keyphrase extraction.

Turney used statistical associations between keyphrases to improve the coherence of the extracted keyphrases [12].

Tang et al. (2004) applied Bayesian decision theory for keyword extraction [13]. They utilized word linkage information and defined two 'local context' features.

Yasin Uzun (2005) used naïve Bayesian classifier, utilizing the features such as TFxIDF score, distance of the word from the beginning of the text, paragraph and the sentence to identify keywords in the text [14]. It has been assumed that the keyword features are normally distributed and independent.

K. Zhang et al. considered keyword extraction as a classification problem [15], in which the words/phrases in a document were to be classified into three groups: 'good keyword', 'indifferent keyword', and 'bad keyword'. Keyword extraction was then achieved by a SVM classification model that was trained in advance.

Medelyan and Witten (2006) proposed KEA++ which enhances automatic keyphrase extraction by using semantic information on terms and phrases extracted from a domain-specific thesaurus [16].

Nguyen and Kan (2007) performed keyphrase extraction from scientific articles by using features which capture salient morphological phenomena found in scientific keyphrases [17].

C. Zhang et al. (2008) used CRF (Conditional Random Field) model to extract keywords [18]. CRF is a new probabilistic model for segmenting and labelling sequence data. It is an undirected graphical model that encodes a conditional probability distribution with a given set of features.

Jiajia Feng et al. (2011) proposed an algorithm based on sequential patterns applicable on a document which is represented as sequences of words. Important sequential patterns are extracted which reflect the semantic relatedness between words [19]. Statistical as well as pattern features within words were used to build the keyword extraction model. The algorithm is language independent and does not require a semantic dictionary to get the semantic features.

Bao Hong et al. (2012) proposed an improved keyword extraction method (Extended TF). They used linguistic features of keywords like word frequency, part of speech, syntactical function of words, location appeared & word's morphology [20]. On the base of the characteristics of each feature, weights were ascribed to different features and the SVM model was used for further optimization.

5.3 Unsupervised Approach

Steier and Belew (1993) utilized the mutual information statistics to discover two-word keyphrases [21]. The mutual information statistic was used to measure the information content of phrases.

Krulwich and Burkey (1996) used heuristics for extracting keyphrases from a document [22]. The heuristics are syntactic ones, such as italicization, the presence of phrases in section headers, and the use of acronyms.

Muñoz (1996) proposed an algorithm based on Adaptive Resonance Theory (ART) to discover two-word keyphrases [23].

Barker and Cornacchia (2000) proposed a simple system of choosing noun phrases from a document as keyphrases [24].

Tomokiyo et al. (2003) utilized pointwise KL-divergence among multiple language models for measuring both phraseness and informativeness of phrases which can then be unified into a single score to rank extracted phrases [25].

Mihalcea et al. (2004) proposed the TextRank, a graph based ranking model for graphs extracted from texts to rank keywords based on the co-occurrence links between words [26]. It makes use of the concept of "voting" between words to extract keyphrases.

Bracewell et al. (2005) extract noun phrases from a document, and then cluster the terms which have the same noun term [27]. The clusters are then ranked depending on term and noun phrase frequencies. Top ranked clusters are selected as keyphrases for the document.

Liu et al. (2009) proposed to extract keyphrases by utilizing clustering techniques which ensure that the document is semantically covered by these keyphrases [28].

Stuart Rose et al. (2010) described Rapid Automatic Keyword Extraction (RAKE), a domain and language-independent method for extracting keywords from individual documents [29]. RAKE is based on the observation that keywords frequently contain multiple words but they rarely contain punctuation or stop words, such as the function words and , the, and of , or any other words having minimal content.

Luit Gazendam et al. (2010) describe the extraction and ranking of keywords with a restricted vocabulary with the help of a thesaurus for the purpose of ranking [30]. For ranking words it uses a weighting scheme called tf-rr which uses both the term frequency and the number of thesaurus relations realized between the thesaurus terms found in the specific document. This approach doesn't need any kind of training from a reference corpus.

Marina Litvake et al. (2011) proposed DegExt, a graph-based, cross-lingual keyphrase extractor [31]. DegExt uses graph representation based on the simple graph-based syntactic representation of document and enhances the traditional vector-space model by taking into account some structural document features.

Ali Mehri et al. (2011) described a method for ranking the words in texts by use of non-extensive statistical mechanics [32]. The non-extensivity measure can be used to classify the correlation range between word-type occurrences in a text.

5.4 Semi-Supervised Approach

Decong Li et al. (2010) proposed a semi-supervised approach which utilizes a generally accepted notion that the title of a

document is always elaborated to reflect the content of the document and thus key phrases naturally have close semantics to the title [33]. Keyphrase extraction is performed by calculating the phrase importance in the semantic network, through which the effect of the title phrases reaches to the other phrases iteratively. They have modeled a semantic network as a hyper-graph, whose vertices represent phrases and weighted hyper-edges measure the semantic relatedness of binary relations as well as *n*-ary relations among phrases and the knowledge base "Wikipedia" is used to estimate the semantic relatedness.[34] Decong Li et al. (2011) proposed a transductive method which represented the phrases in the document as a hypergraph and the hypergraph was expanded to include the implicit phrases, which were then ranked by an inductive learning approach. The highest ranked phrases were termed as implicit key phrases which are those keyphrases which contribute greatly in understanding the document though they do not appear in the text. These phrases are also appropriate to be called as key phrases, and are a beneficial complement to the extracted ones.

6. FEATURE SELECTION FOR KEYWORD/KEYPHRASE EXTRACTION

Intuitively we can classify features for the extraction of keywords and keyphrases in two broad categories.

6.1 Features based on Phraseness

These give us an idea about how strongly different parts of a phrase attach to each other or in other words how much the phrase formed is qualified to be called as a phrase.

Mutual information: Fano in 1961 originally defined mutual information (I), between particular events x and y, in our case the occurrence of particular words, as follows:

$$I(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

This type of mutual information is roughly a measure of how much one word tells us about the other.

Mean and variance: One way of discovering the relationship between two words is to compute the ***mean*** and ***variance*** of the offsets (signed distances) between the two words in the corpus. The mean is simply the average offset between two words. The variance measures how much the individual offsets deviate from the mean. If the offset is the same in all cases, then the variance is zero. If the offsets are randomly distributed (which will be the case for two words which occur together by chance, but not in a particular relationship), then the variance will be high.

The mean and deviation characterize the distribution of distances between two words in a corpus. We can use this information to discover keyphrases by looking for pairs with low deviation. A low deviation means that the two words usually occur at about the same distance. Zero deviation means that the two words always occur at exactly the same distance.

6.2 Features based on Informativeness

These give us an idea about how important is the current keyword or keyphrase. They can be further classified in three categories.

1) Based on term weight

There are various measures to determine the importance of a word in a document as well as in a corpus. Term weighting measures can be useful in identifying keywords and stopwords from a text corpus. Following are some of the important term weighting measures:

Term count: It is just the no of times a word occurs in the document and this count is called term count. An important word in the document is expected to be repeated many times and therefore this measure assigns a higher value to a word which occurs a higher no of times.

Term frequency: Total no. of times a word occurs in a corpus is called its term frequency. To calculate term frequency we add up all the occurrences of a word in all the documents in the corpus.

Document frequency: If a word ***w*** occurs ***n*** no of documents in a corpus of ***N*** documents where ***n*** < ***N,*** then ***n*** is called the document frequency of word ***w.***

Inverse document frequency (IDF)

IDF is an informativeness score that embodies the principle that the more rare a word is, the greater the chance it is relevant to those documents in which it appears. Pure frequency based measure is too simple for evaluating the importance of a term and it doesn't use global information. IDF helps us in distinguishing one document from other in a collection.

IDF = log (N/ d_m)

Where N = no of documents in the corpus, d_m = no of documents containing the word m.

A word which occurs in all the documents in the collection (common word), IDF has value 0 and it has highest value for those words which occur only once. The IDF score has long been used to weight words for information retrieval but it is a weak identifier of informative words.

X^I measure : $X^I_w = f_w - d_w$

Where f_w = term frequency of word **w,** d_w = document frequency of word **w**. Informative words tend to exhibit "peaked" distributions with most occurrences coming in a handful of documents. For two words with the same frequency, the one that is present in lesser no of documents will have the higher score. This score has a higher value for frequent words and lower value for low frequency words.

Average frequency: It is the ratio of total occurrence of a word in the corpus with the total no of documents in the corpus.

Relative frequency: It is the ratio of total occurrence of a word in the corpus with the no of documents in which that word occurs. Relative frequency of a word is always greater than or equal to its average frequency. It is equal only when that word is present in all the documents and if this condition is satisfied that word might be a stopword since they have the highest probability of being present in all the documents.

Term length (TL): TL represents number of tokens included in a keyphrase. Concepts expressed by longer phrases are expected to be more specific, and thus more relevant.

Tf-idf of a keyword: The **Tf–idf** weight (term frequency–inverse document frequency) is used to evaluate how important a word is to a document in a collection or corpus. Importance increases in proportion to the frequency of

occurrence of a word in the document but is countered by the total no of documents having that word.

$$\textbf{Tf.idf} = (\log N/d_m) * \sum_{i=1}^{i=n} f_w(i)$$

TF x IDF of a phrase: Term Frequency (TF) of a phrase measures the number of occurrences of a phrase in a document. The document frequency of a phrase is defined to be the number of documents in the collection that contain that phrase. The TF of a phrase p in a document d is

$$TF_{(p,d)} = \frac{freq(p,d)}{size(d)}$$

Where, freq (p,d) is the term frequency of the phrase p in document d and size(d) is the number of phrases in the document d. The IDF of a phrase in a document d is -

$$IDF_p = -\log_2(\frac{N}{DF_p})$$

Where, N is the number of phrases in the total collection, DFp is the number of document in the collection where the term p is present. The tf-idf of a phrase p in a document d is given by

$$TF - IDF = TF_{(p,d)} * IDF_p$$

2) Based on location in the document

First N terms: Only the first N terms from the document are selected. The logic is that the important keyphrases are found in the beginning of the document as generally important information is put at the beginning.

Last N terms: Only the last N terms of the document are selected. The logic is that the most important keyphrases are found in the last part of the document since important keyphrases are found in their concluding parts of the document

At the beginning of a paragraph: It weights terms according to their relative position in a paragraph. The logic is that the important keyphrases are likely to be found near to the beginning of paragraphs.

At the end of its paragraph: Weights a term according to its relative position in its paragraph. The logic is that the important keyphrases are likely to be found near to the end of paragraphs.

Resemblance to title: Rates a term according to the similarity of its sentence with the title of the article. Phrases similar to the title will have a higher score.

Maximal section headline importance: Rates a term according to its most important presence in a section or headline of the article. It is known that some parts of papers are more important from the aspect of presence of keyphrases such as abstract, introduction and conclusions.

Accumulative section headline importance: It is very similar to the previous one but it weights a term according to all its presences in important sections or headlines of the article.

3) Miscellaneous

Negative Brackets: Phrases found in brackets are unlikely to be keyphrases. Therefore, they are defined as negative phrases, and will grant negative scores.

Shorter concept subsumption: If a keyphrase is (stringwise) included in a longer keyphrase with a *higher frequency*, the frequency of the shorter keyphrase is transferred to the count of the longer one e.g. "computer science engineering"= 6 and "computer science"=4 are re-ranked as "computer science engineering"=10 and "computer science"=0

Longer concept boosting*:* If a keyphrase is included in a longer one with a *lower frequency*, the average score between the two keyphrase frequency is computed. Such score is assigned to the less frequent keyphrase and subtracted from the frequency score of the higher ranked one. For example, if "computer science engineering"=4 and "computer science"=6, the average frequency is 5, so that "computer science engineering"= 5 and "computer science" = 6–5=1.

Spread: Spread of a phrase is the distance between its first and last occurrences in a document. Both values are computed relative to the length of the document. High values help to determine phrases that are mentioned both in the beginning and at the end of a document and hence are spread throughout the document.

7. CONCLUSION

A survey of various approaches put forward by various researchers in recent years for the purpose of keyword and keyphrase extraction was done along with the brief description of different feature selection metrics generally used to rank the candidate keywords and keyphrases according to their importance in the analyzed text. The survey work was divided into major categories such as supervised, unsupervised, semi-supervised and statistical approaches for keyword extraction and major works done in those areas were listed chronologically.

8. REFERENCES

[1] Feather, J. and S. P., International encyclopedia of information and library science. London & New York: Routledge, 1996

[2] Justeson, J., Katz, S., "Technical terminology: some linguistic properties and an algorithm for identification in text", Natural Language Engineering 1, 9-27, 1995

[3] G. Salton, C. S. Yang, C. T. Yu, "A Theory of Term Importance in Automatic Text Analysis", Journal of the American society for Information Science, 26(1), 33-44, 1975.

[4] J. D. Cohen, "Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting" Journal of the American Society for Information Science, 46(3): 162-174, 1995

[5] M. Ortuño et al., "Keyword detection in natural languages and DNA", Europhys. Lett. 57, 759, 2002

[6] J.P. Herrera, P.A. Pury, "Statistical keyword detection in literary corpora", The European physical journal, 2008

[7] P. Carpena et al., "Level statistics of words-Finding keywords in literary texts and symbolic sequences", Physical Review E, 79, 03512(R), 2009

[8] Turney P. D., "Learning algorithms for keyphrase extraction", Information Retrieval, 2: pp 303-336, 2000

[9] Frank E., Paynter G.W., Witten I.H., Gutwin C., Nevill-Manning C.G., " Domain-specific keyphrase extraction", Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 668-673. San Francisco, CA, USA, 1999

[10] Song M. et al.," KPSpotter: a flexible information gain-based keyphrase extraction system", Proceedings of the 5th ACM international workshop on Web information and data management, Pages 50 – 53, 2003

[11] Hulth A. "Improved automatic keyword extraction given more linguistic knowledge", Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 216-223. Association for Computational Linguistics, Morristown, NJ, USA, 2003

[12] Turney P., "Coherent Keyphrase Extraction via Web Mining", Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), pp. 434-439, 2003

[13] Tang J. et al.: Loss Minimization Based Keyword Distillation, Lecture Notes in Computer Science Volume 3007, pp 572-577, 2004

[14] Yasin Uzun, "Keyword Extraction Using Naïve Bayes", Bilkent University, Computer Science Dept., Turkey, 2005

[15] Zhang K. et al. "Keyword Extraction Using Support Vector Machine", Lecture Notes in Computer Science Volume 4016, pp 85-96, 2006

[16] Medelyan O., Witten H. "Thesaurus based automatic keyphrase indexing", Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Pages 296-297, 2006

[17] Nguyen, T.D., Kan, M.Y., "Keyphrase extraction in scientific publications", Goh, D.H.L., Cao, T.H., Sfilvberg, I., Rasmussen, E.M. (eds.) ICADL. LNCS, vol. 4822, pp. 317-326. Springer, 2007

[18] Zhang C. et al., "Automatic Keyword Extraction from Documents Using Conditional Random Fields", Journal of Computational Information Systems 4:3 pp 1169-1180, 2008

[19] Jiajia Feng et al., "Keyword extraction based on sequential pattern mining", Proceedings of the Third International Conference on Internet Multimedia Computing and Service, pages 34-38, 2011

[20] Hong B., Zhen D., "An Extended Keyword Extraction Method", International Conference on Applied Physics and Industrial Engineering, Physics Procedia, Volume 24, Part B, 2012, Pages 1120–1127,2012

[21] Steier A., Belew R., "Exporting phrases: A statistical analysis of topical language", Second Symposium on Document Analysis and Information Retrieval, 1993

[22] Krulwich B., and Burkey C., "Learning user information interests through the extraction of semantically significant phrases", AAAI 1996 Spring Symposium on Machine Learning in Information Access, AAAI Press, 1996

[23] Muñoz,A., "Compound key word generation from document databases using a hierarchical clustering ART model" Intelligent Data Analysis, 1996

[24] Barker, K., and Cornacchia, N., "Using nounphrase heads to extract document keyphrases", Advances in Artificial Intelligence, Lecture Notes in Computer Science, volume 1822/2000, pp 40-52, 2000

[25] Tomikoyo T., Hurst M., "A language model approach to keyphrase extraction", Proceedings of the ACL workshop on Multiword expressions: analysis, acquisition and treatment, Volume 18, Pages 33-40, 2003

[26] Mihalcea, R., and Tarau, P., "TextRank: Bringing order into texts", Proceedings of EMNLP, pp 404-411, 2004

[27] Bracewell et al., "Multilingual single document keyword extraction for information retrieval", Natural Language Processing and Knowledge Engineering, **pp.** 517 – 522, 2005

[28] Liu, Z., Li, P., Zheng, Y., Sun, M., "Clustering to find exemplar terms for keyphrase extraction", Proceedings of Conference on Empirical Methods in Natural Language Processing. pp. 257-266, Singapore 2009

[29] Rose S. et al., "Automatic keyword extractionfrom individual documents", Text Mining: Applications and Theory, John Wiley & Sons Ltd, 2010

[30] Luit Gazendam et al. "Thesaurus Based Term Ranking for Keyword Extraction", Workshops on Database and Expert Systems Applications, pp.49-53, 2010

[31] Litvak M. et al., "DegExt — A Language-Independent Graph-Based Keyphrase Extractor"**,** Advances in Intelligent and Soft Computing, Volume 86, pp 121-130, 2011

[32] Ali Mehri et al., "Keyword extraction by non-extensivity measure", Physical Review E, Volume 83, Issue 5, 2011

[33] Decong Li, Sujian Li, Wenjie Li, Wei Wang, Weiguang Qu, "A semi-supervised key phrase extraction approach: learning from title phrases through a document semantic network", Proceedings of the ACL 2010 Conference Short Papers, pages 296–300, 2010

[34] Decong Li, Sujian Li, "Hypergraph-based inductive learning for generating implicit key phrases", ACM 978-1-4503-0637, 2011