2014 International Conference on Advances in ICT for Emerging Regions (ICTer): 228 - 234

Smart Web Content Bookmarking with ANN based Key Phrase Extraction Algorithm

B.M. Thosini Kumarika[#1], N.G.J. Dias[#2]

[#]*Department of Statistics and Computer Science,*
University of Kelaniya, Sri Lanka
[1]thosinikumarika@gmail.com
[2]ngjdias@kln.ac.lk

Abstract **— This paper introduces a smart web content bookmarking tool which gives the ability of bookmarking only a selected text other than bookmarking whole pages and keeping a set of links relevant to those pages for later reference. This novel concept helps to collect most important text and paragraphs into one place for the ease of use. The major component of this research is the use of an artificial neural network (ANN) with author identified parameters especially applicable in the domain of small textual contents to extract the key idea of the selected text and to organize bookmarked contents under this system suggested meaningful names. Key phrases are an important mean of document summarization and the manual assignment of key phrases to documents is very laborious. The purpose of this neural network approach is to automate this extraction process and bookmark a selected paragraph with extracted key idea using the developed key phrase extraction method with artificial intelligence. The front end of Smart Web Content Bookmark is developed as a browser extension so that the web users can extend their browsers with it to experience the efficient and meaningful bookmarking method in web.**

Keywords **— Bookmarks, Key Phrase Extraction, Artificial Neural Network.**

I. INTRODUCTION

With the development of technology, it has become essential to use the World Wide Web (WWW) for many day to day life activities. Whenever we navigate in the web, we are used to bookmark the most important and useful web pages so that those pages can be accessed later easily. Due to this facility traditional bookmark list is also referred as the Internet shortcuts.

But it is commonly noted that depending on the interests and relevant subject fields of a particular web user, the most useful focus can be a specific part of a web page rather than the whole page. So, it is clear that it is inefficient to access and view this whole page, when user needs to refer only a specific part of it. But all traditional bookmarking methods are to bookmark the whole page instead of the important small textual contents.

Therefore, the intention of this research is to introduce an efficient and novel method to bookmark all the specific contents or small part of texts in each web page stored in one location corresponding to the interests of every individual web user. Thus, web users can refer those bookmarked small textual contents instead of viewing whole page repeatedly, and since all contents are collected into one place the user can quickly refer this collection to view all bookmarked contents at once. The most outstanding feature in this research is that each bookmarked content will be well organized by the system using the suggested meaningful titles. To accomplish this task, a newer key phrase extraction method which incorporates computational intelligence and web technologies were employed.

Key phrases are the phrases consisting of one or more significant words. They can be incorporated in the search results as subject metadata to facilitate information searches on the web [1]. A set of key phrases related to a document may be viewed as a concise summary of the document. In general key phrases are meant to serve various goals such as summarization, indexing and searching. Key phrase extraction is the process of identifying key phrases from a textual content based on many different algorithms.

Manually assigning key phrases to documents is a very laborious task. Therefore, ways of automating this process, using artificial intelligence—more specifically, machine learning techniques—are of interest. Here, this web content bookmarking process uses a trained artificial neural network to automate key phrase extraction and bookmark small textual contents under meaningful names and organize them.

To implement this new model, a chrome browser extension is developed using a trained neural network, and since it bookmarks small text and paragraphs organizing under meaningful names, it can be introduced as smart web content bookmarking tool especially trained for small texts rather than whole documents. Further, web users can install and experience the features of this smart bookmarking tool easily as it comes as a browser extension.

The remaining of the paper is organized as follows; Section II that contains the currently available approaches for the key phrase extraction algorithms and web bookmarking tools, and Section III introduces the details of the developed smart web content bookmarking tool which uses a trained ANN. The experiments and results are discussed in sections IV and section V encloses the conclusion.

978-1-4799-7732-1/14/$31.00 ©2014 IEEE

II. Related Work

A. Current Approaches to Web Bookmarking

All the browsers by default have the facility of bookmarking the whole web page by keeping a reference to specific URL of that particular page. In addition to that there are web bookmarking for the bookmarks or bookmarked contents. Developed smart web content bookmarking tool is a rich novel method to provide all above features to make bookmarking even more efficient for web users.

B. Current Approaches to Key phrase Extraction

In a number of previous works, the authors have suggested various document key phrases extraction methods and they are given in detail below. Key phrase extraction is widely used technique and can be useful in a variety of applications such as retrieval engines [1, 3], browsing interfaces [16], thesaurus construction [17], and document classification and clustering [18].

An algorithm to choose noun phrases from a document as key phrase candidates has been proposed in [5]. Noun phrases are extracted from a text using a base noun phrase skimmer and an off-the-shelf online dictionary.

Chien [6] developed a key phrase extraction system for Chinese and other Asian languages. HaCohen-Kerner et al. [7, 8] proposed a model for key phrase extraction based on supervised machine learning and a combination of the baseline methods. They applied a *decision tree* (a machine learning algorithm) for feature combination.

Hulth et al. [9] proposed a key phrase extraction algorithm in which the hierarchically organized thesaurus, and the frequency analysis were integrated. Inductive logic programming has been used to combine evidences from frequency analysis and the thesaurus.

A graph based model for key phrase extraction has been presented in [10]. A document is represented as a graph in which the nodes represent terms, and the edges represent the co-occurrence of terms. Whether a term is a keyword is determined by measuring its contribution to the graph.

A neural network based approach to key phrase extraction has been presented in [11]. It exploits traditional features such as term frequency, inverted document frequency, and position

tools and social bookmarking sites like Xmarks, Delicious, Find My Bookmarks etc [2]. But none of the above taken into consideration provides the features of content bookmarking. Also they do not have any automated process like providing system suggested titles

(binary) features. The neural network has been trained to classify a candidate phrase as either a key phrase or not. A threshold value of 0.5 has been applied to the output predictions for the crisp classification of the candidate phrase in the categories of: key phrase or non-key phrase.

Turney [12] has treated the problem of key phrase extraction as a supervised learning task. Some of the features are the positional information of the phrase in the document and whether or not the phrase is a proper noun. Key phrases are extracted from candidate phrases based on an examination of their features. Turney's program is called the Extractor.

A key phrase extraction program called Kea, developed by Frank et al. [13, 14], uses Bayesian learning for key phrase extraction tasks. A model is developed from the training documents with exemplar key phrases and corresponds to a specific corpus containing the training documents. Each model consists of a Naive Bayes classifier and two supporting files containing phrase frequencies and stop words (a stop word is an insignificant word, such as "the"), which are commonly used (occur frequently) in a language. The list of such words is called a "stop list" or a "stop word list." The developed model is used to identify the key phrases from a document. In both Kea and Extractor, the candidate key phrases are identified by splitting up the input text according to phrase boundaries (numbers, punctuation marks, dashes, and brackets etc.). Finally, a phrase is defined as a sequence of one, two, or three words that appear consecutively in the text. The phrases beginning or ending with a stop word are not taken into consideration. Kea and Extractor both used supervised machine learning based approaches. Two important features, such as the distance of the phrase's first appearance and the product TF*IDF (used in the information retrieval setting), are considered during the development of Kea. Here TF corresponds to the frequency of a phrase in a document and IDF is estimated by counting the number of documents in the training corpus that contain the phrase at least once. Frank et al. [13, 14] has compared the

performance of Kea to Turney's work and showed that the performance of Kea is comparable to the system proposed by Turney.

An n-gram based technique for filtering key phrases has been presented in [15]. This approach initially computes n-grams such as unigram, bigram, etc., for extracting candidate key phrases. After filtering the candidate key phrases, the phrases are ranked based on the features such as term frequency and the position of the phrase in a document and also in a sentence.

Smart web content bookmarking tool has different key phrase extraction method differs from all the above described approaches since it is used here for a different purpose, which is to identify key concepts for small text and paragraphs. That is a specially developed ANN based key phrase extraction algorithm with author identified parameters that is particularly suits well in the domain of small textual contents.

III. Smart Web Content Bookmarking

A. Introduction

Automated neural network based key phrase extraction is the major component which is introduced through this research which is an alternative to existing key phrase extraction methods and the use of this technique helps for efficient and well organized web content bookmarking method called smart web content bookmarking.

Although, it is a Java based trained neural network approach in its back end, the front end of this solution is a browser extension where the users can extend their browsers by installing the extension to experience the efficient and meaningful web content bookmarking.

A brief description about how the system functions is described in this paragraph. With installed browser extension, whenever the user selects a text on a web page and sends it to the interface of the browser extension, selected text is copied to the OS clipboard. Execution of the Java program then begins and the content of the clipboard is the input for the Java program which runs in the background. As the first step, it will identify all the candidate key phrases of that piece of text and the second step is to calculate the feature vectors for each candidate key phrase. Those feature vectors are the inputs for the trained neural network and the output will be either in class positive or negative (means that the corresponding candidate key phrase is an actual key phrase or not). Finally it produces a list of key phrases for the given text, based on the ranked class probabilities by ordering them in the descending order of probability.

B. Framework for NN Based Key phrase Extraction

This section discusses a framework for key phrase extraction using an artificial neural network. Fig. 1 shows a basic framework for key phrase extraction using machine learning techniques. The source text content should be reduced to a set of candidate phrases, which are analyzed in terms of the features of interest. The candidate phrases are represented in terms of feature vectors. The labeled feature vectors corresponding to the candidate phrases are then fed to the neural network, which learns a model that can be used to classify each candidate phrase as being or not being a key phrase. Then the feature vectors are labeled as "positive" (key phrase) or "negative" (not a key phrase).

As shown in Fig. 1, a machine learning algorithm is fed with the labeled feature vectors, which are prepared after performing two major steps candidate phrase extraction and feature extraction on the input text.

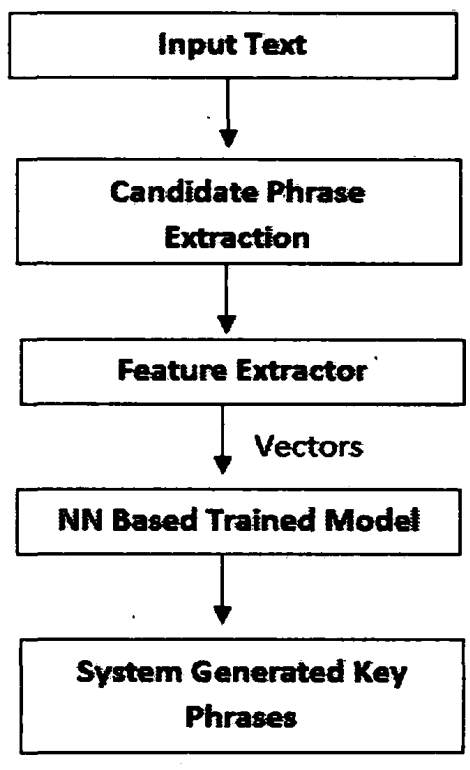


Fig. 1 Framework for neural network based key phrase extraction

C. Candidate Key Phrase Identification

Here the main focus is to identify the noun phrases in the input text. The noun phrases in the document are treated as the candidate key phrases [1]. To identify the noun phrases, documents

should be tagged. The input text segments are passed to a Part-Of-Speech (POS) tagger called OpenNLP [19] to extract the lexical information about the terms. Fig. 2 shows a sample output of the OpenNLP for the following text segment:

"Financial Management means the efficient and effective management of money (funds) in such a manner as to accomplish the objectives of the organization. It is the specialized function directly associated with the top management. The significance of this function is not only seen in the 'Line' but also in the capacity of 'Staff' in overall administration of a company. It has been defined differently by different experts in the field."

Financial_NNP Management_NNP means_VBZ the_DT efficient_JJ and_CC effective_JJ management_NN of_IN money_NN (funds)_NNS in_IN such_JJ a_DT manner_NN as_IN to_TO accomplish_VB the_DT objectives_NNS of_IN the_DT organization._NN It_PRP is_VBZ the_DT specialized_JJ function_NN directly_RB associated_VBN with_IN the_DT top_JJ management._NN The_DT significance_NN of_IN this_DT function_NN is_VBZ not_RB only_RB seen_VBN in_IN the_DT 'Line'_JJ but_CC also_RB in_IN the_DT capacity_NN of_IN 'Staff'_JJ in_IN overall_JJ administration_NN of_IN a_DT company._NN It_PRP has_VBZ been_VBN defined_VBN differently_RB by_IN different_JJ experts_NNS in_IN the_DT field._NN

Fig. 2 Example output of POS tagger

In the Fig. 2, NN, NNS, NNP, JJ, DT, VB, IN, PRP, WDT, MD, etc. are lexical tags assigned by the tagger. In this figure, the tagger's output is shown in a box and the meanings of those tags are given below. This is not the complete tag set. These are some examples of tags used by the OpenNLP.

NN: Noun (Singular), NNS: Noun (Plural), NNP: Proper Noun, JJ: Adjective, DT: Determiner, VB: Verb, IN: Preposition, PRP: Pronoun, MD: Modal, CC: Coordinating conjunction, etc.

To detect noun phrases the output of the POS tagging segment should be further processed programmatically. A fixed pattern of part of speech tags are searched for, and word/tag pairs matching the pattern, are pulled out from the text as noun phrases. For noun phrases, this pattern or regular expression is the following [4]:
(Adjective | Noun)*(Noun
Preposition)?(Adjective | Noun)* Noun

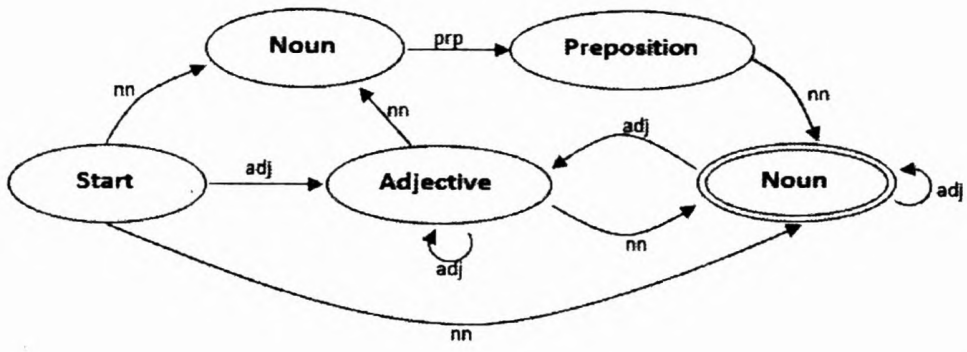


Fig. 3 The finite state machine for noun phrase identification

The base noun phrase finder is implemented on the basis of a finite state machine (FSM), as shown in Fig. 3. The input to the noun phrase finder is a tagged sentence. The sentences in a document are tagged by the POS tagger. Each state in the FSM represents a part-of-speech and an arc is labeled with the tag of a word in the input sentence.
Table I shows a list of noun phrases identified by our noun phrase extractor.

TABLE I
SAMPLE OUTPUT OF NOUN PHRASE EXTRACTOR

Noun Phrase Number	Noun Phrases
1	Financial Management
2	effective management
3	money (funds)
4	manner
5	objectives
6	organization
7	specialized function
8	top management
9	significance
10	function
11	overall administration
12	company
13	different experts
14	field

D. Features Extraction

After identifying the candidate key phrases, we should design and compute a set of features based on which we can determine whether a candidate phrase is a key phrase or not. This is the time where the process of neural network begins and those calculated features are fed into the Multilayer

Perceptron (MLP) Network and all candidate phrases will be ranked based on the output probabilities of the neural network. Figure 4 shows an example for such multilayer feed-forward neural network.

Fig. 4 A multilayer feed-forward neural network: (x_1, x_2, x_3) are the input vectors fed into the network model and weighted connections existing between each layer are established during the training phase.

Discovering good features of a classification task is an important task. In our work, we have used a set of features such as phrase frequency, phrase occurrence and phrase length. These features are newly proposed as it is especially suitable for small textual contents.

The feature weighting methods and the feature value normalization methods for different features used in our key phrase extraction task are discussed below.

1) *Phrase Frequency (PF)*: A noun phrase occurring frequently in a text is assumed to be more important in that text. The number of times a phrase (noun phrase) independently occurs in its entirety in a corpus is considered to be the Phrase Frequency. The value of this feature is normalized by dividing it with the maximum value of the feature.

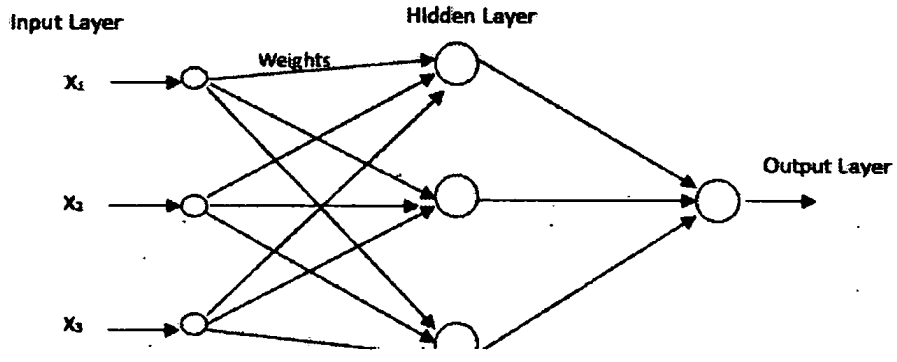
$$F_{freq} = PF / Maximum (PF)$$

2) *Phrase Occurrence:* Key phrases usually occur early in the text paragraph. So, the candidate key phrases that occur early in a text should be given higher score. We consider the position of the first occurrence of a phrase in a document as a feature. Unlike the previous approaches [11, 13], which assume the position of a phrase as a binary feature, in our work, the score of a phrase that occurs first in the i-th sentence in a document is computed using the following formula assuming that each sentence has a fixed number of words:

$$F_{occ} = 1/square\text{-}root\ (i)$$

3) *Phrase Length:* This feature can be considered as the structural feature of a phrase. Phrase length becomes an important feature in the key phrase extraction task. We find that key phrases consisting of 4 or more words are relatively rare in our corpus. The value of this feature is normalized by giving a value in the range of 0-1 based on the phrase length of each candidate key phrase.

Employing above features, the multilayer perceptron feed-forward neural network is trained and in this supervised learning process the sigmoid function is used as activation function. The aim of this learning process here is to classify candidate key phrases as it is a key phrase (positive) or not a key phrase (negative). The Backpropagation learning algorithm is used to train the network by adjusting the weights of the neural network with minimizing the error rate.



The output of the neural network model provides the most suitable key phrases for a given textual content based on the ranked probabilities. This list of key phrases will be the system suggested titles for the given input textual content which is to be bookmarked.

The neural network framework called "encog" is used to train and implement the neural network with the help of Eclipse Java IDE. All the other steps such as candidate phrase identification and feature extraction is also done programmatically using Java, since it is platform independent and easy to implement using Eclipse Java framework.

IV. Evaluation

There are two usual practices for evaluating the effectiveness of a key phrase extraction system. One method was to use human judgment. We asked human experts to give scores to the key phrases generated by the system. Another method, which is less costly, is to measure how well the system-generated key phrases match the author-assigned key phrases. We followed the second approach to evaluate the proposed key phrase extraction system by computing its precision and recall using author-provided key phrases for the documents. For our experiment, precision and recall have been defined as follows:

precision = the proportion of the extracted key phrases that match the key phrases assigned by a document's author(s). So, the precision is defined as follows:

$$Precision = N/K,$$

where N is the number of key phrases matched and K is the number of system generated key phrases.

Recall = the proportion of the key phrases assigned by a document's author(s) that are extracted by the key phrase extraction system. So, the recall is defined as follows:

$$Recall = N/M,$$

where N is the number of key phrases matched and M is the number of author assigned key phrases. Previous studies have used those measures and have found that it is an appropriate method to measure the performance of a key phrase extraction system [5, 8, 18, 20].

To build up the corpus which is used for the experiments on the key phrase extraction task, articles were downloaded from the websites relevant in domains Economics and Medical. The proposed key phrase extraction algorithms are

compared with an existing system called Kea [13, 14], which is now a publicly available key phrase extraction system. Kea uses Bayesian learning for key phrase extraction task. The author assigned key phrases for the small textual content which is used as an example in the section III C of the paper is given below:

Author assigned key phrases: Financial Management, Management of money, Controlling funds, Specialized function

Table II produces top 5 key phrases extracted by the proposed ANN based system and Kea respectively.

TABLE III
TOP 5 KEY PHRASES EXTRACTED BY THE PROPOSED ANN BASED SYSTEM AND KEA

ANN Key	Kea Key
Financial Management	Financial
effective management	Management
Money(funds)	objectives
specialized function	organization
top management	function

Table II shows that two key phrases extracted by ANN based system exactly match with the author assigned key phrases and a close match with author assigned key phrases can be observed from key phrases extracted by ANN based system compared to that of Kea.

Table III shows the overall performance comparison between the Kea and our neural network based key phrase extractor.

TABLE IIIII
PERFORMANCE COMPARISON BETWEEN THE PROPOSED ANN BASED KEY PHRASE EXTRACTION METHOD AND KEA

Number of Key Phrases	Average Precision		Average Recall	
	ANN	Kea	ANN	Kea
3	0.30	0.10	0.50	0.20
5	0.40	0.40	0.50	0.25
10	0.40	0.20	0.66	0.50

By changing the number of key phrases to be extracted for each test case both precision and recall measures are taken for developed ANN and Kea. Table III data shows that most of the time (about 82%) the number of key phrases extracted by ANN that match the author assigned key phrases is greater than that of Kea. For an example, when considering the test case of 10 key phrases, 4 key phrases extracted by the proposed ANN match with the author assigned key phrases while only 2 key phrases extracted by Kea match the author assigned key phrases.

From Table III, we can clearly conclude that proposed ANN based key phrase extraction algorithm outperforms Kea for all three cases shown in the table.

V. CONCLUSIONS

The final outcome of this research is a smart web content bookmarking tool which bookmarks only a specific textual content of a web page relevant to each web user's interest and it was a need for a long time among web users who often happen to refer bookmarked web page many times for a small important textual content. This tool comes in a form of a browser extension to web users and the major component developed through this research is a neural network based efficient key phrase extraction method. This intelligent key phrase extraction outperformed existing famous key-phrase extraction method called Kea. This new approach uses features like phrase frequency, phrase occurrence and phrase length as they are more suitable to extract key concepts from a small text corpus. Using this novel key phrase extraction method, those content bookmarks are well organized under system suggested meaningful titles.

In the future, the proposed systems can be improved by (1) improving the candidate phrase extraction module of the system by implementing ANN using different, effective learning algorithm than BP algorithm and (2) incorporating new features such as document structural features, lexical features, etc.

ACKNOWLEDGMENT

We would like to thank the authors of the references we have used throughout this research for their valuable effort and everyone who helped us to implement this research work.

REFERENCES

[1] Y. B. Wu, Q. Li, "Document key phrases as subject metadata: incorporating document key concepts in search results". Journal of Information Retrieval, Volume 11, Number 3, 2008, pp.229-249.

[2] Elise Moreau. About.com Web Trends. [Online]. Available: http://webtrends.about.com/od/pro5/tp/bookmarking-tools-bookmarklets.htm

[3] S. Jones, M. Staveley, "Phrasier: A system for interactive document retrieval using key phrases", In: proceedings of SIGIR'99, Berkeley, CA, 1999.

[4] Introduction to Noun Phrase Detection. [Online]. Available: https://files.ifi.uzh.ch/cl/hess/classes/ecl1/termerCIE.html

[5] K. Barker, N. Cornacchia, "Using Noun Phrase Heads to Extract Document Key phrases", In: H. Hamilton, Q. Yang (eds.): Canadian AI 2000. Lecture Notes in Artificial Intelligence, Vol.1822, Springer-Verlag, Berlin Heidelberg, 2000, pp.40-52.

[6] L. F. Chien, "PAT-tree-based Adaptive Key phrase Extraction for Intelligent Chinese Information Retrieval", Information Processing and Management", 35, 1999, pp.501-521.

[7] Y. HaCohen-Kerner, "Automatic Extraction of Keywords from Abstracts", In: V. Palade, R. J. Howlett, L. C. Jain (eds.): KES 2003. Lecture Notes in Artificial Intelligence, Vol.2773, Springer-Verlag, Berlin Heidelberg, 2003, pp.843-849.

[8] Y. HaCohen-Kerner, Z. Gross, A. Masa, "Automatic Extraction and Learning of Key phrases from Scientific Articles. In: A. Gelbukh (ed.): CICLing 2005. Lecture Notes in Computer Science, Vol.3406, Springer-Verlag, Berlin Heidelberg, 2005, pp.657-669.

[9] A. Hulth, J. Karlgren, A. Jonsson, H. Boström, "Automatic Keyword Extraction Using Domain Knowledge" , In: A. Gelbukh (ed.): CICLing 2001. Lecture Notes in Computer Science, Vol.2004, Springer-Verlag, Berlin Heidelberg, 2001, pp.472-482.

[10] Y. Matsuo, Y. Ohsawa, M. Ishizuka, "KeyWorld: Extracting Keywords from a Document as a Small World", In: K. P. Jantke, A. shinohara (eds.): DS 2001. Lecture Notes in Computer Science, Vol.2226, Springer-Verlag, Berlin Heidelberg 2001, pp.271-281.

[11] J. Wang, H. Peng, J-S. Hu, "Automatic Key phrase Extraction from Document Using Neural Network", ICMLC 2005, 2005, pp.633-641.

[12] P. D. Turney, "Learning algorithm for key phrase extraction", Journal of Information Retrieval, 2(4), 2000, pp.303-36.

[13] E. Frank, G. Paynter, I. H. Witten, C. Gutwin and C. Nevill-Manning, "Domain-specific key phrase extraction", In: proceeding of the sixteenth international joint conference on artificial intelligence, San Mateo, CA,1999.

[14] I.H. Witten, G.W. Paynter, E. Frank, et al, "KEA: Practical Automatic Key phrase Extraction", In: E.A. Fox, N. Rowe (eds.): Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries. ACM Press, Berkeley, CA, 1999, pp.254-255.

[15] N. Kumar, K. Srinathan, "Automatic key phrase extraction from scientific documents using N-gram filtration technique", In Proceeding of the eighth ACM symposium on Document engineering, Sao Paulo, Brazil, 2008.

[16] C. Gutwin, G.W. Paynter, I.H. Witten, C.G. Nevill-Manning and E. Frank, "Improving browsing in digital libraries with key phrase indexes." Journal of Decision Support Systems, Vol.27, no 1-2, 1999, pp.81-104.

[17] B. Kosovac, D. J. Vanier, T. M. Froese, "Use of key phrase extraction software for creation of an AEC/FM thesaurus", Journal of Information Technology in Construction, 5, 2000, pp.25-36.

[18] S. Jonse, M. Mahoui, "Hierarchical document clustering using automatically extracted key phrase", In proceedings of the third international Asian conference on digital libraries, Seoul, Korea, 2000, p.113-20.

[19] openNLP on the Apache Software Foundation. [Online]. Available: https://opennlp.apache.org/

[20] Q. Li, Y. B. Wu, "Identifying important concepts from medical documents", Journal of Biomedical Informatics, 2006, pp.668-679.