Keywords Recommender for Scientific Papers Using Semantic Relatedness and Associative Neural Network

F Nugroho¹, I D Sumitra²
¹ Department of Postgraduate of Information System, Universitas Komputer Indonesia, Indonesia
² Department of Information Management, Universitas Komputer Indonesia, Indonesia

*fajarnugroho.id@gmail.com

Abstract. The purpose of this paper is to extract keywords from scientific papers. Keyword extraction is used to assist in suggesting keywords for indexers or to produce a summary in a form of keywords for scientific writings. Scientific writings are also categorized as structured document because the scientific papers have parts such as title, abstract, content, and conclusion, so we are going to find important information on those parts. So the method that suitable was Associative Neural Network Algorithm and using trained Associative Neural network with chosen keywords data sets we are able to find keywords features in scientific documents. So the automated keywords extraction and pre-included keywords from the writer will be compared to choose either the most relevant or complementary and providing recommendation or perhaps the extracted automated keywords and the keywords from the writer are exactly the same.

1. Introduction

Keyword extraction is tasked with the automatic identification of terms that best describe the subject of a document[1]. Automatic keyphrase extraction as the automatic selection of important, topical phrases from within the body of a document [2] Keywords extraction is a collection of short phrases which captures main topics discussed in a document[3].

Scientific papers are also categorized as structured document because the scientific papers have parts such as title, abstract, content, and conclusion, so we are going to find important information on those parts. The frequency by which a word appears in the different parts of the document, the use of the word as part of the abstract, and figure/table captions, and various word formatting options such as bold-faced, italicized and large fonts[4] is an important thing on the keyword extraction. Generally the method used for Keyword Extraction are statistics, linguistics, machine learning[5]. Statistics approach used for find the statistical information of the words are used to search for the keywords in the document. Linguistics Approach used for finding the noun phrase is considered with the context of the documents and machine learning approach used for find keywords in a document, but before that, the machine must be trained with some dataset the suitable of main topics.

The method that suitable is Associative Neural Network Algorithm and using trained Associative Neural network with chosen keywords data sets we are able to find keywords features in scientific documents.
2. Methods
A Statistical Approach of Keyword Extraction for Efficient Retrieval (Shruti Luthra, Dinkar Arora, Kanika Mittal, and Anusha Chhabra). Mentioning that technique is words have high frequency in the all terms and page are chosen candidate keywords and have graph that is built in word-net[5]. Keyword Extraction using Semantic Analysis (Mohamed H. Haggag) This model for extracting keywords based on their relatedness weight among the entire text terms. Relationship between semantic is important[6]. Keyword Extraction using Auto-associative Neural Networks (German Aquino, Waldo Hasperue, Laura Lanzarini). Mentioning the techniques used to extract keywords text-documents quickly from Spanish language without a large training set[7]. It covers any lexical or functional association, so it is a more general concept than semantic similarity. Dissimilar entities may still be related by many possible relationships, such as meronymy (or “part of” relation, as in finger, hand), antonymy (opposite meanings, as hot, cold), or any kind of functional relationship or frequent association (for example, penguin, Antarctica, that are not linked by any lexical relation)[8].

An associative neural network (ANN) is a combination of an ensemble of the feed-forward neural networks and the K-nearest neighbor (KNN) technique. The introduced network uses the correlation between ensemble responses as a measure of distance among the analyzed cases for the nearest neighbor technique and provides an improved prediction by the bias correction of the neural network ensemble both for function approximation and classification[9].

KNN method used as:

$$z(x)=\frac{1}{k}\sum_{j\in N_k(x)} y(X_j)$$

Where $z(x)$ is a predicted value for case $x$, $N_k(x)$ is the collection of the $k$ nearest neighbors of $x$ among the input vectors in the training set. Other approaches besides used combining statistical approach, semantic approach and machine learning or can use some heuristic knowledge such as the position of text, length, layout feature of words and especially when keyword extraction for website is can use Html tags and meta tags [10].

3. Results and Discussion
Automatic keywords extraction (AKE) from scientific papers is used as an assistant to suggest keywords for indexers or to exert summary feature in the form of keywords for inaccessible scientific papers. Thus automatic keywords extraction and keywords input by the writer will be compared either by the most relevant or complementary and giving recommendations, or automatic keywords extraction and writer keywords are exactly the same. The methodology is depicted in Figures 1, 2, 3 and 4 below:

![Figure 1. AKE dan Author Keywords is different](image)

From Figure 1, AKE keywords are different from keywords given by the writer or indexer.
From Figure 2, both AKE and author keywords intersect therefore from the same words the system could give keywords recommendation for indexers.

From Figure 3, both AKE and author keywords are exactly the same for every extracted word. The proposed method for this research is depicted in Figure 4 below:

From proposed method in Figure 4, all scientific papers are preprocessed initially, then identify the pattern or position of each word in each part, next is to divide into several words, and finally identify per word interconnection based on training data, so we expect the method to provide relevant keywords in scientific papers. The process within information retrieval or text mining requires several stages which basically to prepare so the text can be arranged in a structured manner. One of the ways of text mining is text preprocessing. Text preprocessing is a stage where the application selects data to be
processed in each document. It covers (1) case folding, (2) tokenizing, (3) filtering, and (4) stemming in Figure 5.

![Diagram of preprocessing steps]  

**Figure 5.** Proses Preprocessing

In this section is the result of research keywords for scientific papers. The first is we show the example of the dataset and the end of result will show combining machine learning method and semantic relatedness.

For now, dataset obtained from kaggle that is a set of keywords from many journals. Below is example in Table 1 of dataset.

| **Table 1. Dataset From Kaggle** |
|---|---|---|
| **Id** | **Year** | **Abstract** |
| 1 | 1987 | An efficient method of self-organizing |
| 2 | 1988 | In modeling studies or memory based on neural networks |
| 3 | 1994 | Deformable models are an attractive approach to recognizing |

Below is Table 2 of the total dataset after processing data.

| **Table 2. Total keywords** |
|---|---|
| **Number of documents** | 7241 |
| **Total Number of Keywords** | 29793 |

Based on the research conducted by combining machine learning technique and semantic relatedness, the results were obtained. Machine learning method which is ANN algorithm have used training of
dataset, that is dataset from kaggle have many keywords and for seeing the terms of the sentence used the strengthened by semantic relatedness.

4. Conclusion
In this research is keywords extraction intended for the scientific document which is combining of two methods that are machine learning and semantic relatedness, we consider interesting that an algorithm can learn from examples of dataset and can learn the term of keywords whether is important or not. For the further when used a large of dataset the process training still takes a long time.

Acknowledgements
We would like to express sincere gratitude to Mr. Assoc. Prof. Dr. Ir. Eddy Soeryanto Soegoto, MT as Rector of Universitas Komputer Indonesia (UNIKOM), and Mr. Dr. Yeffry Handoko Putra, ST., M.T as Head of Master of Information Systems. Special thanks are given to Irfan Dwiguna Sumitra, M.Kom., Ph.D. who has given full support and guidance so that this paper can be realized.

References
[1] Luthra, S., Arora, D., Mittal, K., & Chhabra, A. 2017. A Statistical Approach of Keyword Extraction for Efficient Retrieval. *International Journal of Computer Applications*, 168(7), pp. 5-8
[2] Turney, P. D. 1999. Learning to extract keyphrases from text. National Research Council. *Institute for Information Technology, technical report ERB-1057*.
[3] Turney, P. D. 2000. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4), pp. 303-336.
[4] Azcarraga, A., Liu, M. D., & Setiono, R. 2012. Keyword extraction using backpropagation neural networks and rule extraction. In *The 2012 international joint conference on neural networks (IJCNN)* (pp. 1-7), IEEE.
[5] Luthra, S., Arora, D., Mittal, K., & Chhabra, A. 2017. A Statistical Approach of Keyword Extraction for Efficient Retrieval. *International Journal of Computer Applications*, 168(7), pp. 1-10
[6] Haggag, M. H. 2013. Keyword extraction using semantic analysis. *International Journal of Computer Applications*, 61(1), pp. 1-6.
[7] Lanzarini, L. C., Hasperué, W., Estrebou, C. A., Ronchetti, F., Villa Monte, A., Aquino, G. O., ... & Luna, C. 2017. Minería de datos y big data: aplicaciones en señales y textos. In *XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires)*.
[8] Gracia, J., & Mena, E. 2008. Web-based measure of semantic relatedness. In *International Conference on Web Information Systems Engineering* (pp. 136-150). Springer, Berlin, Heidelberg.
[9] Tetko, I. V. 2002. Associative neural network. *Neural Processing Letters*, 16(2), 187-199.
[10] Kaur, J., & Gupta, V. 2010. Effective approaches for extraction of keywords. *International Journal of Computer Science Issues (IJCSI)*, 7(6), pp. 144.