Adaptive Keywords Extraction using Back Propagation Neural Networks- A Review

Neeraj Sharma

Manish Mann, Ph.D

ABSTRACT
Keyword extraction is important for Knowledge Management System, Information Retrieval System, and Digital Libraries and also for general browsing of the web. Keywords are generally the basis of document processing methods such as clustering and retrieval because processing all the words in the document can be slow. In the existing work, it is observed that the keywords extracted do not include the bold, italic and underlined or words that are of different font size in the document. However, enhanced fonts are the major source of keywords in the document. Further it is also observed that the synonyms of the keywords are not included in the keywords search space and this may be a one of the most important source of keyword search space as many words are used in document by their synonyms as well. In the proposed work, the keyword extraction is not based on merely the predefined keyword dictionary, but the key words are extracted from the particular document based on some features like repetitive frequency of a particular word or form using neural network approach. Also, in the presented system, the extracted keywords are specific to the document and not the common for each document.

The back propagation neural network results are more reliable if an exhaustive training samples are provided to the network. More is the training of the network, more precise keyword extraction is possible. A large no. of feature set may slow down the network operation. Therefore, an optimum no. of features set is likely to be designed that completely describe the document under study.

Keywords
Neural networks, back propagation, clustering, data mining

1. INTRODUCTION
As the amount of information in the modern world grows rapidly, it is becoming more and more difficult to maintain and process document archives for Knowledge Management Systems, Information Retrieval Systems, and Digital Libraries. This is true especially for very large archives with millions or more articles. Processing all the words in the documents, as if they are of equal importance, as basis for finding relevant articles would be slow and not practical. That is why it is important to have a set of good keywords that represent the actual contents of the document. However, it is not possible to have all documents labeled by experts. It is therefore useful to be able to automatically identify keywords in the documents that are just as good as assigned keywords.

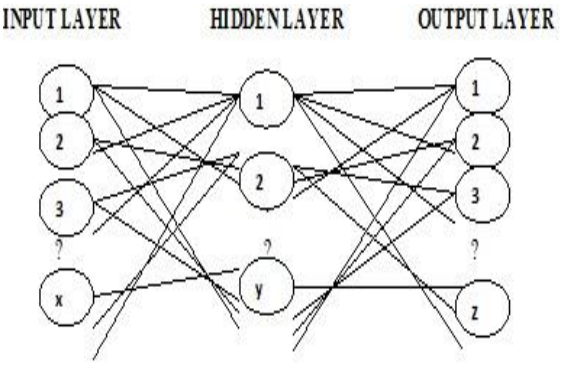

Fig 1: Model of Input and Output Relation in Neural Network

Keywords are often the basis of document processing methods such as clustering and retrieval since processing all the words in the document can be slow. Common models for automating the process of keyword extraction are usually done by using several statistics-based methods such as Bayesian, K-Nearest Neighbor, and Expectation-Maximization. These models are limited by word-related features that can be used since adding more features will make the models more complex and difficult to comprehend.

Back propagation serve as the network in generalizing the relationship of the title and the content of articles in the archive by following certain features such as the position of word in the article, position of word in the sentence, frequency of the word, format applied, and other attributes defined beforehand. Neural networks have been long known for being able to solve problems not solvable by other traditional problems. Also, back propagation networks are considered universal approximations and are capable of solving non-linear problems. The advantage of using neural network is that it is flexible throughout different types of datasets. This data-driven approach is also relatively faster as compared to hand-crafted systems such as expert systems. In other words, back propagation network can be trained specifically towards a category of articles since articles of different category may have different format, content, or definition of useful keyword.

2. BACKGROUND
In the existing work, it is observed that the keywords extracted do not include the bold, italic and underlined or words that are of different font size in the document. However, enhanced fonts are the major source of keywords in the document. Further it is also observed that the synonyms of the keywords are not included in the keywords search space and this may be a one of the most important source of keyword search space as many words are used in document by their synonyms as well.

3. OBJECTIVE
In the proposed work, the keyword extraction is not based on merely the predefined keyword dictionary, but the key words

are extracted from the particular document based on some features like repetitive frequency of a particular word or form using neural network approach. Also, in the presented system, the extracted keywords are specific to the document and not the common for each document.

4. LITERATURE SURVEY

Keyword extraction is important for Knowledge Management System, Information Retrieval System, and Digital Libraries and also for general browsing of the web. Keywords are generally the basis of document processing methods such as clustering and retrieval because processing all the words in the document can be slow. Automating the process of keyword extraction is usually done by using several statistics-based methods such as Bayesian, K-Nearest Neighbor, and Expectation- Maximization. These models are limited by word-related functionalities that can be used since adding more features will make the models more complex and difficult to comprehend. A Neural Network, specifically a back propagation network, can be used in generalizing the relationship of the title and the content of documents in the archive by following word features other than TF-IDF, such as word position in the sentence, paragraph, or in the entire document, and formats such as heading, and other features defined beforehand. In order to explain how the back propagation network works, a rule extraction method will be used to extract symbolic data from the resulting back-propagation network. The rules fetched can then be transformed into decision trees performing almost as accurate as the network plus the advantage of being in an easily comprehensible format [1].

All clustering methods have to assume some cluster relationship between the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. The major difference between a traditional dissimilarity/similarity measure and latter is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved [2]. The term clustering is used in many research communities to define methods for association of unlabeled data. Clustering is useful in numerous exploratory pattern-analysis, assemblage, decision-making, and machine-learning circumstances, along with data mining, document recovery, image segmentation, and pattern organization. The basic purpose of clustering is to organize objects of data into many separate clusters basically as the intra cluster in which resemblance is maximum and other type in which the inter cluster difference between those is maximum. There have been various clustering algorithms available every year and the efficiency of algorithms relies on the aptness of the similarity measure to the data at hand. The objects which are to be determined should be in the same kind of cluster at the same time the location of the points from someplace to begin this dimension should be outer surface of the cluster and this application is known as Multi viewpoint-based Similarity. Multiple viewpoints clustering provide an ability to determine important structures within the rule base by providing a way to structure both hierarchically and orthogonally. Recently, multi-view clustering methods have been proposed to expand over conventional single-view clustering. It is possible to make use of more than single point of indication for creating new concept of identity [3]. Some cluster relationships have to be considered for all clustering methods surrounded by the data objects which will be applied on. There may be a similarity between two objects which can be defined as a choice of explicitly or implicitly. The main distinctness of latter concept with a traditional dissimilarity/similarity measure is that the aforesaid dissimilarity/similarity exercises only a single view point for which it is the base and where as the mentioned Clustering with Multi-viewpoint Based Similarity Measure uses many distinguished viewpoints that are objects and are assumed to not be in the same cluster with two objects being measured. By making use of multiple viewpoints, countless descriptive evaluation could be accomplished. [4]. Clustering methods have to consider some cluster relationships among the data objects that they are carried upon. Similarity between a pair of objects can be characterized either explicitly or implicitly. Using multiple viewpoints, more informative assessment of affinity could be achieved [5]. Clustering is one of the data mining and text mining methodologies which is used to analyze datasets by dividing it into meaningful groups. The objects in the dataset may have certain relationships among them. All clustering algorithms consider this before they are applied to datasets. The existing algorithms for text mining make use of a one viewpoint for measuring similarity between objects. Their drawback is that the clusters can't express the complete set of relationships among objects. To overcome this drawback, a new similarity measure known as multi-viewpoint based similarity measure is there to ensure the clusters show all relationships among objects. The empirical study revealed that the hypothesis "multi-viewpoint similarity can bring about more informative relationships among objects and thus more meaningful clusters are formed" is proved to be true and it can be used in the real time applications where text documents are to be searched or processed frequently [6]. Clustering problems are assumed in which the available attributes can be split into two independent subsets, such that either subset suffices for learning. Example applications of this multi-view setting include clustering of web pages which have an intrinsic view (the pages themselves) and an extrinsic view (e.g., anchor texts of inbound hyperlinks); multi-view learning until now has been studied in the context of classification. Multi-view versions of k-Means and EM greatly improve on their single-view counterparts [7]. As we know a cluster is a collection of similar objects situated together and are divergent to other cluster objects. With multiple viewpoints, more beneficial measurement of similarity could be accomplished. Two criterion functions for document clustering are inter-cluster and intra-cluster relation between objects. The previous clustering process focused on hierarchical clustering of Multi-view point documents, which are not focused on less and high dimensional data. Especially, the bisecting divisive clustering approach is considered here. This advance consists in iteratively splitting a cluster into two sub-clusters, starting from the main dataset [8].

The database object that describes multiple attributes is referred to as high dimensional data space. In high dimensional data, the common distance measures may be influenced by noise. Existing clustering algorithms are evaluated based on partitioning, hierarchical, density based and grid based. These methods assume many kinds of cluster relationships among the clustered objects. Similarity among two objects may be defined as implicitly or explicitly. Our main objective is to cluster web documents [9]. Clustering is the one of the very important task in data mining .The aim of clustering is to find the fundamental structures in data and divide them into meaningful subgroups for supplementary study and examination. Disadvantage of existing K-Means clustering with MVS measure is that it doesn't best position to

cluster the data points. This problem will give rise to gain less optimal solution for clustering method. There is a solution to the Multi-view point based similarity measure with NMF clustering to predict k value. Latter gives a detailed study on the multi-view point clustering approach with the NMF clustering method [10]. The clustering will have some clustering relationships between the documents or objects that we are applied on. In traditional method only one view point is used as a reference that is k-means algorithm for similarity between the documents. In one another method cosine with multi view point based similarity measure is used between the documents. The multi view will provide more information assessment than traditional method and reduce the not required documentation [11].

5. METHODOLOGY

The proposed work is based on extraction of key words based on documents word's data base arranged in an array. A histogram of each word is extracted by incrementing the histogram count against each word array index.

The words histogram method basically removes the unnecessary words from the search like prepositions, articles verbs etc. These are the words that have the maximum count and need to be removed from the search space. Now the document is remained with some words that are specific to the document matter. The document may further be filtered out by the title of the document and the words appearing the title are given prior importance.

Flow Chart

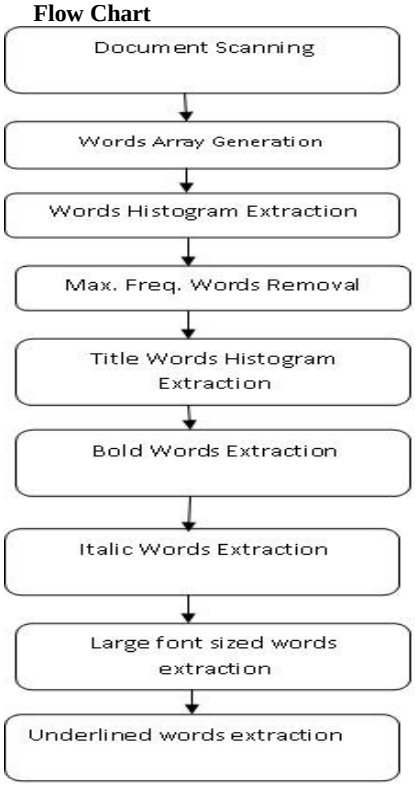

Fig. 2 : Flow Chart for Proposed Scheme

6. CONCLUSION

The presented approach shows fair results on different category of textual documents. The approach adopted here is that no keywords should be left if it has the probability of being a keyword. Rather, if a word have less probability of keyword but is selected as keyword, it is fine. As the purpose is not to leave the authentic keyword. Also, the presented work is on textual document and for testing purposes, only notepad documents are taken.

The back propagation neural network results are more reliable if an exhaustive training samples are provided to the network. More is the training of the network, more precise keyword extraction is possible. A large no. of feature set may slow down the network operation. Therefore, an optimum no. of features set is likely to be designed that completely describe the document under study.

7. REFERENCES

[1] Arnulfo Azcarraga and Michael David Liu, Rudy Setiono, "Keyword Extraction Using Back-propagation Neural Networks and Rule Extraction". WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.

[2] Duc Thang Nguyen, Lihui Chen, Chee Keong Chan, "Clustering with Multi-viewpoint based similarity measure". IEEE transactions on knowledge and data engineering, vol. 24, no. 6, june 2012

[3] Aggadi Gnanesh, M.Sudhir Kumar, "An advance towards standard utilities of document clustering". International Journal of Computer and Electronics Research [Volume 2, Issue 4, August 2013].

[4] K.A.L.V Prasanna, Mr. Vasantha Kumar, "Performance evaluation for multi-viewpoint based similarity measure for data clustering". Journal of Global Research in Computer Science Volume 3, No. 11, November 2012.

[5] S. Sesha Sai Priya, k. Rajini Kumari, "The clustering with multi-viewpoint based similarity measure". IJCST Vol. 3, Issue 1, Spy. 5, Jan. - March 2012.

[6] Gaddam Saidi Reddy, Dr.R.V.Krishnaiah, "Clustering Algorithm with a Novel Similarity Measure". IOSR Journal of Computer Engineering (IOSRJCE) 2278-0661 Volume 4, Issue 6 (Sep-Oct. 2012), PP 37-42.

[7] Steffen Bickel and Tobias Scheffer, Humboldt-Universit¨at zu Berlin, "Multi-View Clustering". Proceedings of IEEE international conference on data mining 2004.

[8] B.Amuthajanaki, K.Jayalakshmi, "A hierarchical divisive clustering based multi-viewpoint similarity measure for document clustering". International Journal of Advances in Computer Science and Technology Volume 2, No.8, August 2013.

[9] S. Chandrasekhar, K. Sasidhar, M. Vajralu, "Study and analysis of multi-viewpoint clustering with similarity measures". International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 10, October 2012.

[10] R.Saranya, P.Krishnakumari, "Clustering with multi-viewpoint based similarity measure using NMF". International Journal of scientific research and management (IJSRM), Volume 1, Issue 6, Pages 316-322, 2013.

[11] Annavazula Mrinalini, A. Rama Mohan Reddy, "Implementation of a multi-viewpoint method for similarity measure for clustering the documents". International Journal of Advanced Research in Computer Science and Management Studies, Vol 2, Issue 1, January 2014.