Keyword Extraction: A Neural Network Approach

Neeraj Sharma[1], Dr. Manish Mann
[1]*Research Scholar, LR Institute of Engineering and Technology, HP India*

***Abstract* - T**his research focused on using neural networks, back-propagation to be precise, in extracting keywords from script archives. News articles and journal articles of different groups are used as the dataset in training and testing the network. These digital articles would be of consistent format, for example HTML and PDF with consistent document pattern. It is assumed that the articles will be in supported format for that the initial parsing to proceed since an external library will be used to extract information from the documents and noisy documents will lead to wrong extraction of keywords.

Keywords *– neural networks, back propagation, documents, archieves, histogram*

I. INTRODUCTION

As the content of information in the modern world accrues fast, it is becoming more and more ambitious to maintain and process document extracts for Knowledge Management Systems, Information Retrieval Systems, and Digital Libraries. This is true especially for huge extracts with thousands or more articles. Processing all the words in the articles, as if they are of same importance, as basis for finding acceptable articles would be very slow and not practical. That is why it is of utmost importance to have a set of good keywords that describe the actual abstracts of the document. However, it is not practical to have all articles labeled by experts. It is therefore useful to be able to automatically acknowledge keywords in the articles that are just as good as awarded keywords.

Automatic keyword extraction (AKE) is the job to discover a small set of words, key phrases, keywords, or key segments from an article that can explain the meaning of the document. Since keyword is the smallest entity which can convey the meaning of article, many text mining applications can take benefit of it, e.g. automatic indexing, automatic summarization, automatic categorization, automatic clustering, automatic filtering, topic recognition and tracking, information revelation, etc. Therefore, keywords extraction can be considered as the core technology of all automatic processing for documents.

Neural network topologies, or architectures, are formed by organizing neuron-like cells into fields (also called layers) and linking them with weighted interconnections.

Characteristics of these topologies include connection types, connection schemes, and field configurations. There are two primary connection types, excitatory and inhibitory. The three primary cell interconnection schemes are in tri-field, inter-field and recurrent connections. Field (layer) configurations combine fields of cells, information flow and connection schemes into a coherent architecture. Field configurations include lateral feedback, field feed-forward, and field feedback. Neural Networks may be viewed as a kind of application of kernel methods to density estimation in multi-category problem. The neural network has 3 layers: input, hidden and output.

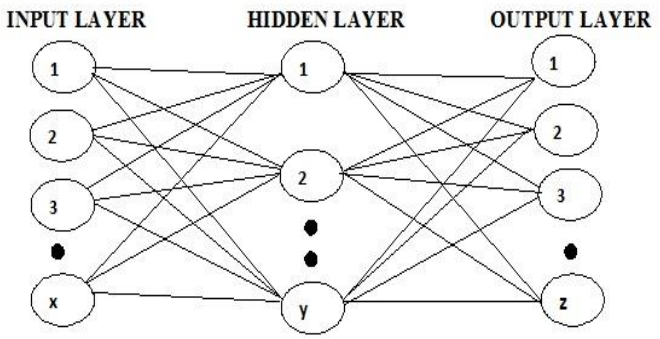

Fig.1.1; Model of Input and Output Relation in Neural Network

II. 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 [3]. The major difference between a classical 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 [7]. 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 [1]. 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 articles are to be searched or processed regularly [8]. The previous clustering process focused on hierarchical clustering of Multi-view point articles, 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 [6]. In classical method only one view point is used as a reference that is k-means algorithm for similarity between the articles. In one another method cosine with multi view point based similarity measure is used between the articles. The multi view will provide more information

assessment than classical method and reduce the not required documentation [2]. Articles are unstructured data comprising of natural language. Document surrogate means the structured data reorganized from original articles to map them in computer systems. Document surrogate is traditionally represented into a queue of words [10]. Also, back propagation networks are considered universal approximations and are capable of solving non-linear queries [9]. This is true especially for very large extracts with millions or more articles. Processing all the words in the articles, as if they are of equal importance, as basis for finding relevant articles would be slow and not practical [5]. As demanding as the upbringing of the neural network, knowledge representation is also vital [4].

III. PROBLEM FROMULATION

Articles are unstructured data consisting of natural language. Document surrogate means the structured data reorganized from original articles to process them in computer systems. Document surrogate is usually represented into a list of words. Because not all words in a document reflect what it contains, it is necessary to select important words related with its content among them. Such important words are called keywords and they are selected with a particular equation based on TF (Term Frequency) and IDF (inverted Document Frequency). Actually, not only TF and IDF but also the location of each word in the document and the inclusion of the word in the title should be considered to select keywords among words contained in the text. The equation based on these factors gets too complicated to be applied to the selection of keywords.

IV. FUTURE SCOPE

The presented work is implemented on notepad format textual document, particularly, only the textual content. However, the textual content can be in the form of table and same can be enhanced to tabular textual content. Also, the work does not include the numerical figures if are made part of the keywords. Texts from the scene may also be incorporated in the future work. That will be a complete package for keywords extraction from the given document.

Since the fashion words or expressions are used when writing procedural documents stick to some plain writing styles and guidelines, we deduce that keywords can be mined not just by making an allowance for the frequency of appearance of the words, but also by their location in the document, paragraphs, or individual sentence as well as by the way and the places where they are used. For instance, significant words may be likely at the front as well as final parts of the document. Significant words might also seem frequently at the start of paragraphs.

ACKNOWLEDGMENT

The successful realization of this thesis work was only possible due to collaboration of many people in many ways. To all of them I would like to pay my gratitude. First of all, I would like to thank my worthy supervisors duo Dr. Manish Mann, associate professor and Mr. Ravinder Thakur, assistant professor in the department of computer science and engineering at LR Institute of Engineering and Technology, Solan. They provided me all possible guidance and support. They suggested me many ideas and solved my puzzles when I was in need. I would like to thank many students of B.Tech. and M.Tech. who cooperated me in collection of samples, which I have used in the thesis work. I would like to thank *Mr. Virender Sood*, lab in charge of our M.Tech. computer labs for well maintenance of lab and computers.

V. REFERENCES

[1] 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.

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

[3] 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.

[4] Azcarraga, A.P. and Setiono, R. (2002). Generating Concise Sets of Linear Regression Rules from Artificial Neural Networks, International Journal on Artificial Intelligence Tools , 11(2), 189-202.

[5] Azcarraga, A.P., Yap Jr., T.N.: Extracting meaningful labels for websom text extracts. In CIKM'01, Proc. of the 10th International Conference on Information and Knowledge Management (2001) 41– 48.

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

[7] 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

[8] 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.

[9] Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feed- forward networks are universal approximators, Neural Networks , 2(5), 359-366

[10] Jo, T. 2003. "Neural based approach to keyword extraction from articles". In ICCSA'03 Proc. of International Conference on Computational Science and Its Applications, LNCS 2667, 456–461.