2nd International Conference on Recent Innovations in Science, Engineering and Management
JNU Convention Center, Jawaharlal Nehru University, New Delhi (ICRISEM-15)
22 November 2015 www.conferenceworld.in ISBN: 978-81-931039-9-9

AN EMPIRICAL STUDY OF FEATURE EXTRACTION APPROACHES

V. Jayaraj[1], P.Rajadurai[2], V.Mahalakshmi[3]

[1]Associate Professor, [2,3]Research Scholar,Department of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli, Tamil Nadu, (India)

ABSTRACT

Prevailing information systems allow companies to apprehend huge amounts of data. Much of this data captured is structured and can be analyzed using old-fashioned database software. Increasingly, however, plethora of textual data is unstructured, and confronts simple attempts to make sense of it. Manual analysis of this unstructured textual data is impractical, and as a result, numerous text mining methods are being developed to automate the process of analyzing this unstructured data.

Keywords: Business Intelligence, Data Mining, Data Warehouse, Resume mining, Text Mining.

I. INTRODUCTION

Perched on a new century, most of the business houses principally depend on the abundant digital data available across the globe. The advent of internet has changed the business scenario and plethora of data is available for decision making and to improve business transactions. Globalization has changed the very face of the business nature and many innovative techniques have emerged to cater to the need of the evolving business needs. With the aid of these new technologies, the raw information is mined to extract useful meaningful data to help the business grow. Plethora of information are textual based document and, the information in most cases are stored in an unordered manner, from this unordered collection of text document, the process of extracting information or deriving knowledge is termed as text mining.

The last few decades has witnessed a colossal growth in the volume of information available across the globe. Most of this information is unused since it requires a stringent methodology to mine and extract. In order to obtain the knowledge from this vast array of dataset, many powerful tools and techniques have been developed. Since the ability to process this information by the humans is limited, the amount of available information increases exponentially, which leads to mammoth information saturation. It is very imperative to detect useful patterns from the available information to retrieve data

Hovering on anurbane technological world, the business house's strategies and promotions have changed dramatically in the recent years since the importance of data in their business plays a pivotal role in almost all of their business activities. Today computer has become a part and parcel of human life and the colossal volume of data available across the globe provides a helping hand to discover useful meaningful information to enhance the business in various activities related to decision making and decision support. Hence the need for an efficient data mining tool or model is imperative for every business organization. Text mining, also known as text data mining [5] or knowledge discovery from textual databases [4]. Afolabi (2008) refers generally to the

process of extracting remarkable and non-trivial patterns or information from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases .

The phrase "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information [7].Data mining is all about looking for patterns in data, whereas text mining is about looking for patterns in text and documents. However, the artificial resemblance between the two conceals real differences. Data mining is described as the extraction of implicit, previously unknown, and potentially useful/meaningful information from data [10]. However in text mining, the information to be extracted is clearly and explicitly stated in the text or document. It's not concealed at all and from a human point of view, the only sense in which it is "previously unknown" is that human resource restrictions make it infeasible for people to read the text themselves. The problem, of course, is that the information is not couched in a manner that is agreeable to automatic processing and extraction. Text mining is a process to extract the text in a form that is suitable for consumption by computers directly, without any human interaction and help.

The last few decades has witnessed a colossal growth in the volume of information available across the globe. Most of this information is unused since it requires a stringent methodology to mine and extract. In order to obtain the knowledge from this vast array of dataset, many powerful tools and techniques have been developed [9]. Since the ability to process this information by the humans is limited, the amount of available information increases exponentially, which leads to mammoth information saturation. It is very imperative to detect useful patterns from the available information to retrieve data.

Mining is the common term used to denote the extraction process from the available huge volume. One such mining technique commonly used in today's world, is the Data mining. Data mining also known as Knowledge Discovery is the process of extracting or mining essential information from the vast amount of data. It is most commonly used in extracting information from ordered pattern. In major business, data are complex in nature and exist in deferent formats and often organized in a poorly manner (i.e.,) in an unordered manner. From these sources, the data mining alone cannot be efficient to extract the useful information. In order to extract the essential text from these textual documents, a new powerful tool has been used called text mining.

Text Mining is an art of acquiring potentially useful knowledge from the textual document. Data Mining cannot derive its impact on extracting useful details from large unstructured materials based on natural language. But text mining will be the solution. The process of text mining is also termed as Information Extraction or Information Retrieval or Document Classification. The process of extracting information from huge volumes of data is necessary in order to obtain proper knowledge from the useful information by snubbing unwanted information.This useful information enables the administrator to arrive to the decision correctly and facilitates to improve their business to a greater extent. Most of the business records are maintained in the form of documents and hence the documents are in unstructured format. In order to obtain knowledge from this unstructured document requires more manual process and time. Hence to elude this snag, text mining plays a pivotal role in extracting the essential much needed information by categorizing the text. Thus, text mining can be also termed as "Categorization of Texts".

The globalization has led to cut throat competition in business and the decision making is very dominant factor for the success of a business. Thus business intelligence is the process of developing the strategies in order to

gain competitive advantage for business. To accomplish this, countless techniques have been emerged and utilized. One such technique that helps business intelligence is data mining. But data mining has certain limitations as it can be efficient in mining the information from the structured internal data and predicting the trends based on these data or knowledge to make wise decision. But some of the business handles the textual information which is not in a structured manner such as project report, employee resume, and competitor's profile.These documents are presented in unstructured form, since the details are described in the form of text using natural language and each of the documents followed the own developers style. Unlike the data in the database, the information in the textual document are scattered through the text. From such kind of documents, the information can be mined with proper care. The afore said details can be implemented by the proposed algorithm after developing a configuration file to identify the useful information from the document, and various patterns can be easily extracted and organized in order to provide the meaningful information. Through the extracted information, the administrator can able to make the decision to enhance the business.

II. RELATED WORKS

Self-Organizing feature Map proposed by Kohonen. Kohonen believed that, a nervous network's outer input receiving model was to divide the nervous network into different regions, these regions have different corresponding features to the input model, and such an input process is finished automatically. The connecting weights of various neurons have a certain distribution, the nearest neurons excite each other, while the distant neurons inhibit each other, and the more distant neurons have a relatively weak inter-inhibition effect. In a word, Self-Organizing feature Map method is a teacher-free clustering method, and compared with the traditional model clustering methods, its former cluster centers could be mapped on a contour or plane, with the topological structure maintained original. A toolkit named as "Learning Pinocchio (LP)", was applied on resumes to learn Information Extraction rules from resumes written in English. The information identified in their task includes a flat structure of Name, Street, City, Province, Email, Telephone, Fax and Zip code. Learning Pinocchio is an adaptive system for IE, based on a kind of transformation based like rule learning. Rules are learned by generalizing over a set of examples marked via XML tags in a training corpus.

Sumitha mageshwari et al., 2010 proposed a two layer model approach by applying a notion of special features on skills section directly the comparison between the features would not be effective. The features are extracted from similar resumes and each feature consists of a skill type and its skill values. For example, 'programming languages' is askill type and 'c, c++, java' are skill values. The skill information itself forms a hierarchy where skill types form one layer and skill values form another layer. This approach considers only the skills section of the résumés to determine the "*Degree of Specialness*." Authors have made an attempt to categorize the skills in a résumé by clustering them on the basis of skill types or skill values which helps to select and provide relevant resumes.

The K-Means Nearest Neighbor algorithm is a machine-learning instance-based technique. This method does not construct models but stores the training instances. For each new instance, the algorithm compares the distance feature-vectors to the training set. The nearest neighbors are selected based on the distance of the features of the new instance i.e., the similarity between the new instance and the training set vectors. "K" in the

algorithm defines the number of nearest neighbors. The classification of the new object is based on the distance between K-clusters and the object. The object is assigned to the cluster with the minimum distance[11].

III. OBJECTIVE OF THE RESEARCH WORK

The foremost objective of this research work is to develop a novel clustering algorithm to mine information from resumes to enhance and ease the recruitment processes in organizations. Apart from this objective, the following are the secondary objectives,

1. To analyze, survey, compare and summarize various existing algorithms related to text mining and after careful analysis find the snags present in the existing algorithms.
2. Apply stringent norms to extract important features present in the resumes to form.

IV. PERFORMANCE EVALUATION MEASURES

We used the F-measure to evaluate the accuracy of the clustering algorithms. The F-measure is a combination of *precision* and *recall* values used in information retrieval. Each cluster obtained can be considered as a result of query, whereas each pre-classified set of documents can be considered as a desired set of documents for that query. We treat each cluster as if it was the result of a query and each class as if it was the relevant set of documents for a query. The recall, precision, and F-Measure for natural class Ki and cluster Cj are calculated as follows:

$$\text{Recall} = \frac{n_{ij}}{|K_i|}$$

$$\text{Precision} = \frac{n_{ij}}{|C_j|}$$

Where, n_{ij} is the number of members of class K_i in cluster C_j ,The corresponding F-Measure $F(K_i, C_j)$ is defined as:

$$F(K_i, C_j) = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

The overall F-measure is calculated as follows

$$\text{Overall } F(C) = \sum \text{Max}\,(F(K_i, C_j))$$

This Overall F Measure is the weighted sum of the F – measures of the best clusters and a higher F(C) value indicates higher accuracy.

The second evaluation measures we do is calculate the *Rank index* (RI) measures the percentage of decisions that are correct.A true positive (TP) decision assigns two similar documents to the same cluster, a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors we can

commit. A false positive (FP) decision assigns two dissimilar documents to the same cluster. A false negative (FN) decision assigns two similar documents to different clusters.

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

V. CONCLUSION & FUTURE WORK

The process of selecting appropriate resume from a huge volume of resumes is one of the problems often faced by the recruiters in most of the companies. We have extended the notion of special features to extract the special skills, and experience from given set of resumes. In future work with the help of the experimental results we have shown that there is 40-90% reduction in the number of features that the recruiter needs to browse through to select appropriate resumes.The resume selection process is a tedious work as resume contains freeform texts which are difficult to compare and also has a hierarchical structure containing different sections. As each resume contains different sections with each section containing different types of text, an integrated approach has to be developed by considering information in each section. In future, we have to developed an approach to overcome this problem of extracting the important features present in the resume and cluster the similar resumes with accuracy.

REFERENCES

[1] Afolabi I.T., Musa G.A., Ayo C.K. and Sofoluwe A. B (2008). Knowledge discovery in online repositories: a text mining approach. European Journal of Scientific Research, 22 (2): 241-250

[2] Feldman R., Regev Y and Gorodetsky M (2008). A modular information extraction system. Intelligent Data Analysis, 12(1): 51-71.

[3] Feldman Ronen, and Ido Dagan (1995). Knowledge Discovery in Textual Databases (KDT), KDD, 95: 112-117.

[4] Hearst M.A (1999). Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 3-10.

[5] Hearst, Marti A (1997). Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, 112-117.

[6] Kohonen T (1982). Self-Organized Formation of Topologically Correct Feature Maps," Computer Journal of Biological Cybernetics, 43(1): 59-69.

[7] Sebastiani F (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1): 1-47.

[8] Sumit Maheshwari., Sainani A and Reddy P.K (2010). An approach to extract special skills to improve the performance of resume selection. In Databases in Networked Information Systems, Springer Berlin Heidelberg, 256-273.

[9] V.Jayaraj and V.Mahalakshmi, "Augmenting Efficiency of Recruitment Process using IRCF text mining Algorithm" Indian Journal Of Science And Technology(IJST) , ISSN:0974-6846 July 2015.

[10] Witten I.H and Frank E (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[11] Yiu-Ming Cheung (2003). k*-Means: A new generalized k-means clustering algorithm, Pattern Recognition Letters 24: 2883–2893.

.