ISSN (Online): 2230-7893
www.IJCEM.org

Information Retrieval Configuration File Text Categorization Algorithm for Improving Business Intelligence

V. Jayaraj[1] and V. Mahalakshmi[2]
[1,2]School of Computer Science, Engineering & Applications,
Bharathidasan University, Tamilnadu, India.

Abstract

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. This paper proposes an methodology to implement the text mining technology to enhance the scope of Business Intelligence. A novel algorithm named information retrieval configuration file is proposed to extract useful information from a set of documents and employs a new technique in extraction where the information gained is relevant and efficient for business needs. This paper mainly focuses on the development of a new algorithm after analyzing and comparing it with the existing algorithm like support vector machine and hidden markov model.

Keywords: Business Intelligence, Data Mining, Data Warehouse, Ordered manner, Text Mining, Unordered Manner.

1. Introduction

The last few decades has witnessed a stupendous growth of information across the internet. The giant of information are unused across the globe and it requires rigid methodology to mine and extract the text. The growth of information is increasing exponentially and it becomes more important to detect useful pattern from the data. While retrieving the needful data from unstructured source it requires more manual process and time. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Mining is the common term used for extracting patterns in the data. Hence to cunning this snag, text mining plays an decisive role in extracting the needed information by categorizing the text. The globalization has lead 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. A technique that helps business intelligence is data mining. But data mining has some restrictions 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 profile. These documents are presented in unstructured form, since the details are prescribed 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 disperse through the text. From such kind of documents, the information can be mined as per the user wish and it 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.

In this paper, Section 2 presents the various reviews and related work which are to understand the paper. Section 3 describes the proposed information retrieval configuration file algorithm and Section 4 describes the experimental results finally section 5 concludes the paper.

2. Related Work

Information retrieval is the activity of obtaining relevant information from a collection of information resources, Diana Inkpen, explains that information retrieval is an task, which is about finding precise data in databases and also presents a more detailed view of an data[1]. R.Baeza-Yates et al., they worked out the strong points is that from the unordered collection of text document, the process of extracting information is termed as text mining. Many term based methods such as Rocchio model, Probabilistic models, Support vector machine models are used in the beginning[2]. Ning Zhong et al., stated that pattern based approach should perform better than the term-based approach and also many pattern discovery technique helps to improve the effectiveness of an data[3].

Raymond J.Mooney states that by integrating natural language processing, machine learning, data mining and

information retrieval can useful for discovering knowledge from large text corpora can be developed[4].Sonia et al., describes that extracting text information based on Hidden Markov Model(HMM) is easy to establish and it has an strong therotical foundation and can process data robustly. HMM is an popular statistical tool for modeling a wide range of time series data. But it is not sufficient in cluster documents using keywords[5].Mladenic and Grobelnic states that support vector machine as classification technique that was first applied to text categorization, it uses all the words in the document collection to identify important keywords. Yang et al., used Naïve based model[6] . Kun Yu et al., presented an effective approach for resume information extraction to support automatic resume management and routing[7]. A cascaded information extraction (IE) framework is designed. In the first pass, a resume is segmented into a consecutive blocks attached with labels indicating the information types. Then in the second pass, the detailed information, such as *Name* and *Address*, are identified in certain blocks (e.g. blocks labeled with *Personal Information*), instead of searching globally in the entire resume. The most appropriate model is selected through experiments for each IE task in different passes. The experimental results show that this cascaded hybrid model achieves better F-score than flat models that do not apply the hierarchical structure of resumes. It also shows that applying different IE models in different passes according to the contextual structure is effective[8].

3. Proposed Method

The core objective of the paper is to develop a methodology to mine the useful information from the unstructured textual content in order to improve the business intelligence. The mining process can be achieved by new emerging technology, text mining. With the help of text mining, the user can able to discover previously unknown knowledge in text, by automatically extracting information from different written resources developed in natural languages. Thus, text mining is the extension of data mining and obtains the goal of extracting meaningful data from different sources of textual documents. In data mining, the collection of data is stored in the repository known as Data Warehouse. Likely, in text mining, the collection of documents is stored in the repository known as Document Warehouse. From this Document Warehouse, the text has to be extracted using text mining. The summary of the proposed methodology to extract the text from different sources of documents and make the extracted text by the decision makers to support for business intelligence is as described below: The user may wants to extract the text from different sources of documents stored in a word file or an excel file or in any text or pdf file. To perform this, a configuration file is developed to ensure that it support all kinds of documents. In this configuration file, a set of configurations with suitable conditions in order to train the data are provided. This can be discern using the regular expression such as [0-9], [a-z], [A-Z]. Based on this regular expression, the text can be identified and extracted. The document warehouse consists of set of documents from which we have to extract the useful text by using our proposed methodology and some of the rules has been applied.

Rule1: Identify the originates of the documents from the datawarehouse

$S = \sigma (D / N)$

Where

S – Source of document

σ (D / N) – Selection of a document from N document in Warehouse

Rule 2:Identify the type of an document.It may be any one of the following type word or html file or pdf file.

Rule 3.The next step is to develop a configuration file with set of conditions to extract the required information based on the user requirement.

Config, ζ = {x1 $\rightarrow$ x2 / x1 is the set of conditions to validate the text in the document,

x2 is the set of values for the condition x1, where x1,x2 $\in$ X,

X is the selected Document from the warehouse}

Following the creation of configuration file, the next process involved in this mining process is to choose the document reader to read the document. The document reader can be chosen based on the document type.

Based on these 3 rules, the next step is to extract the text from the document by reading the document using document reader and to identifying the text using the configuration file developed in step2.

$\lambda = \Pi\ d1 \in D$

$\alpha = \lambda / \zeta$

where,

ζ – configuration file

λ– text, d1 read by the document reader from selected document, D

α– extracted text by verifying the condition given in configuration file, ζ with λ.

Finally the resultant obtained in α is summarized by storing it a 2-dimensional array. The resultant obtained may be look like as follows: Finally, by reading the values from this 2-dimensional array, the resultant value has to be concatenated to form the textual data or it may stored the value in an ordered manner.

Thus, the text mining process has to be sustain successfully by extracting the useful information from the document

stored in a document warehouse. From the mined result, the business executives can make the decision to improve the business intelligence.

$$\begin{pmatrix} \alpha 1 & Field1 & Value1 \\ \alpha 2 & Field2 & Value2 \\ . & & \\ . & & \\ . & & \\ \alpha n & Fieldn & Valuen \end{pmatrix}$$

3.1 Algorithm : Text Mining Algorithm

Input : Information from the Document Warehouse, (D / N)

Supported Input: Configuration File, ζ developed for D (as in section Configuration File described above)

Output : indispensable mined text, α

```
Begin
Persual the Inputted textual document, D
Scanning the Configuration File, ζ
Choose the Document Reader for D, DR
 x= 1
 z = 1
 Repeat
 DR = DLx / D, where DLx is the read line by DR
 For y = 1 to ζ.End
 If DR.equals(ζ.condy) then
      Field = ζ.valuey
      Value = ζ.condy
 End loop
 End if
 Next
Arr[x][z] = {DR, Field, Value}
z++
x++
Until DR = EOF
For a=1 to n
For b=a+1 to n
Navigate (Arr[a][b] → table[a][b]);  //store the array value
          to table
 Next
 Next
End
```

The process of algorithm *IRCF Text Mining Algorithm* is explained in this section. (IRCF – Information Retrival based on Configuration File). The initial step is to get the input document from the user. The document can be chosen from the document warehouse. Also, the configuration file is inputted for the algorithm to extract the text from the document. From these two inputs, the mining process can be performed as follows:

First, with the help of the chosen Document Reader, the Document is read line by line. Upon fetching each line, it is passing through the configuration file to check for the condition. If the condition with the configuration file matches with the read line, then the value is fetched from the configuration file and stored in the two dimensional array. This process is iterated until the document reader reaches the end of the file. Upon completing this process, the final step is to read the data from the two dimensional array and store the value in the corresponding field in the table. Thus the extraction process is completed successfully and the text from unstructured document is converted into a structured table. From the value stored in the table, the decision maker can make the decision to improve their process.

4. Experimental Results

The proposed methodology has been experimentally verified and the result proves that the proposed methodology satisfies the aim of the paper. To do the experiment, we have to undertaken 100 candidate's resume, which is of different format. From those 100 candidates, a particular group of members has to be selected. First, the selection process can be carried out manually by examining each resume and identifying the qualifying candidate. The time period can be noted.

Next, the selection process can be carried out by using the proposed IRCF text mining algorithm. Based on the algorithm, the qualifying candidate can be chosen and the time period to obtain the result has to be noted. By comparing these two periods, we can conclude that the proposed algorithm performs well than the existing methods. Thus our proposed methodology satisfies the aim of the paper and helps the decision makers to improve the business intelligence. Our proposed methodology is compared with the model in paper [8] and the average is calculated. The comparison proves that the proposed method performs well. Precision (P), recall (R) were used as the basic evaluation metrics and macro averaging strategy was used to calculate the average results.

These models is tested to extract the personal detailed information and the educational detailed information. For both these information, precision and recall can be calculated. The results are tabulated below:

Model	Personal Detailed Info		Educational Detailed Info	
	Avg.P(%)	Avg.R(%)	Avg.P(%)	Avg.R(%)
SVM	86.83	76.89	67.36	66.21
HMM	79.64	60.16	70.78	76.80
IRCF	88.90	80.45	75.66	80.78

Table1.Comparison Data

From this table, IRCF achieves better results compared to SVM and HMM. Thus our proposed method performs well to mine the information from the text. The comparison chart is shown below:

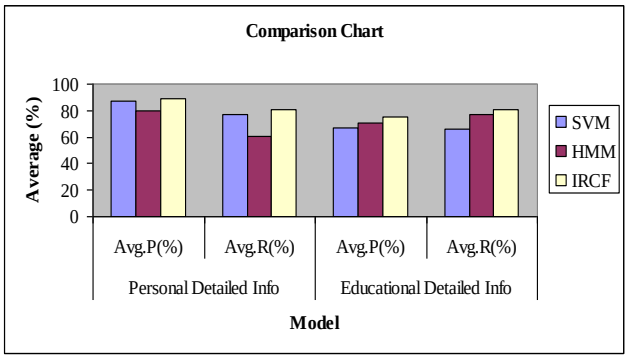

Figure2. Comparison Chart

5. Conclusion

The new proposed algorithm developed was experimentally tested using real time data and from the result of the experimental setup, we can conclude that the proposed algorithm IRCF performs well in extracting meaningful useful text from the unstructured document without any manual intervention. Also, the proposed methodology can be applied in all areas to mine the text from unstructured document. From the experimental results the proposed algorithm ICRF clearly outperformed the HMM and SVM algorithm in terms of speed regarding execution and scanning.

References

1.Diana Inkpen . Information Retrieval on the Internet,University of Toronto.

2.R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

3.Ning Zhong, Yuefeng Li, and Sheng-Tang Wu. Effective Pattern Discovery for Text Mining, IEEE.Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.

4.Raymond J.Mooney and Un Yong Nahm .Multilingualism and Electronic Language Management. Proceedings of the 4th International MIDP Colloquium,September 2005.

5.Sonia Bansal and ReenaGang. A novel probabilistic approach for efficient information retrieval. International Journal of Computer Applications, November 2010.

6.Mladenic and Grobelnic. Feature Selection for Unbalanced Class Distribution and Naive Bayes.In Proceedings of the 16th International Conference on Machine Learning (1999) 258–267.

7.Yang, Y., Pedersen J. O. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14th International Conference on Machine Learning (1997) 412–420.

8.Kun Yu, Gang Guan, Ming Zhou. Resume Information Extraction with Cascaded Hybrid Model, Proceedings of the 43rd Annual Meeting of the ACL, June 2005.

Authors Profile

Dr.V.Jayaraj holds a Ph.D in Computer Science from Bharathidasan University, Tiruchirappalli, Tamilnadu in 2009. He has more than 19 years of teaching experience and has guided more than 50 M.Phil Scholars. He is currently working as an Associate Professor and guiding Ph.D scholars in the Department of Computer Science in Bharathidasan University. Presently he is doing his research on Mobile Computing and Data Mining domains.

V.Mahalakshmi received her Master Degree in M.Tech (Information Technology) from BharathidasanUniversity, Tiruchirappalli, Tamilnadu, India in 2011. she is currently doing his Doctoral Degree in School of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli, Tamilnadu, India. and her research works are under the domains of Data Mining and Information retrieval. She is also member in different societies.