Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 6, June-2016

Documents Subject Identification and Clustering based on Subject

Shruthi S N[1], Kavitha G[2]

[1]*PG Student, Department of Computer Science and Engineering, U B D T College of engineering.*

[2]*Assistent Professor, Department of Computer Science and Engineering, U B D T College of engineering.*

Abstract --- *With the dramatic growth of textual information over the Internet or databases, there is an increasing need for the system that can automatically discover useful knowledge from the text. Text Mining is the process of applying automatic methods to analyze and structure textual data in order to create useable knowledge from previously unstructured information. Standard text mining techniques of text document usually rely on word matching. This paper describes how to recognize the subject of each document in the directory and categorizes into related subject directory. mPDF and PDF parser are the powerful PHP libraries utilized in this work for recognizing the subject. Document clustering is a technique used to group similar documents. This work proposes a tool for maintaining the large set of PDF documents and having many applications.*

Keywords-- Data mining; Text mining; Subject identification; clustering; PDF parser;

I. INTRODUCTION

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc. Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data.

Data mining contains different subfields includes, web mining, graph mining, sequence mining, temporal data mining, mining of multimedia, distributed data mining, text mining, and spatial data mining. Here system considers text mining for its main procedure. Text databases consist of huge collection of documents. They collect this information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc. Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is semi-structured. For example, a document may contain a few structured fields, such as title, author, publishing date, etc. But along with the structure data, the document also contains unstructured text components, such as abstract and contents. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users require tools to compare the documents and rank their importance and relevance. Therefore, text mining has become popular and an essential theme in data mining.

Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users. The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared to offline applications.

II. RELATED WORK

There were many researches gone under text mining related to subject identification and clustering. Some of them are considered here for this system references.

- In paper 1[6], with title “Text Mining: A Burgeoning technology for knowledge extraction” authors described overview of text mining. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from a usually a large amount of different unstructured textual resources. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can work with unstructured or semi structured data sets. At the beginning there is the raw text input denoted as text corpus representing a collection of text documents, like memos, reports, or publications. Grammatical parsing and preprocessing steps transform the unstructured text corpus into a semi-structured format denoted as a text database. Subsequently a structured representation is created by computing a document-term matrix from either the text corpus or the text database. The document-term matrix is a bag-of-

words mechanism containing term frequencies for all documents in the corpus. This common data structure forms the basis for further text mining analysis, like text classification, syntax analysis, relationship identification, information extraction & retrieval, and document summarization. The only Drawback is they didn't reveal which technique is better.

- In paper 2[4], with title "Document Categorization using Data Mining Techniques" authors described data mining techniques. In this paper authors designed an innovative solution which provides single as well as multiple research articles aggregation reducing redundancies. This system gives short condensed and accurate summarized categorical contents presenting innovative authentic information from multiple relevant technical articles. Authors have selected only 'Abstract', and 'Introduction' sections instead of whole document for optimized summarization. The research articles to be summarized are first preprocessed using sections segmentation and converted into plain text. It removes all stop words and tokenizes the article. Drawback of the approach is authors have not suggested one data mining technique to execute.

- In paper 3[5], authors described a Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering. Author described Text mining is interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Standard text mining and information retrieval techniques of text document usually rely on word matching. An alternative way of information retrieval is clustering. In which document pre-processing is an important critical step in the clustering process and it has a huge impact on the success extract knowledge. Document clustering is a technique used to group similar documents .proposed procedure contains steps include text data pre-processing, tokenization, stop ward elimination, steming, vector space model, dimensionality reduction and clustering the documents. The main Drawback is they described only centric clustering.

- In paper 4[7], authors described text mining techniques and its applications. They defined system that is starting with a collection of documents; a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system. Mentioned technologies are information extraction, topic tracking, summarization, categorization, clustering, Concept Linkage, Information Visualization. Mentioned applications are most often used in the following sectors: Publishing and media, Telecommunications, energy and other services industries, information technology sector and Internet, Banks, insurance and financial markets, Political institutions, political analysts, public administration and legal documents, Pharmaceutical and research companies and healthcare. Drawback of this work is authors have not reported which technique is efficient.

- In paper 5[8], authors presented design of Intelligent Interface for Document Understanding. The paper deals with the design of Intelligent Interface for Document Understanding. Document Understanding is an important aspect of Web-based Information Retrieval. It is a process of extracting useful information from the document such as text file, PDF file and web page. Intelligent interface is an AI program which acts as an interface between human and computer. Intelligent Interface helps user to understand web documents semantically and delivering useful information to the users, which results in enhancing utilization of e-resources. The main drawback is here authors consider only text for understanding documents.

III. PROPOSED SYSTEM

The huge amount of technical information is published worldwide in the form of research articles every year. Reading these multiple domain specific articles one by one to get desired information is just time-consuming, sometimes unnecessary, irrelevant and impossible.

To solve above problems an innovative system is developed which predict the paper theme and put it in the correlated directory. In this system, PDF parser and mPDF are utilized for conversion of read only document to write format, resolves whole document and make collection of words. Using PHP, collected words are contrast with dataset. Finally displays result as total words in the document, number of words recognized, number of words expected and whether the paper related to particular topic or not.

Description of the Proposed System

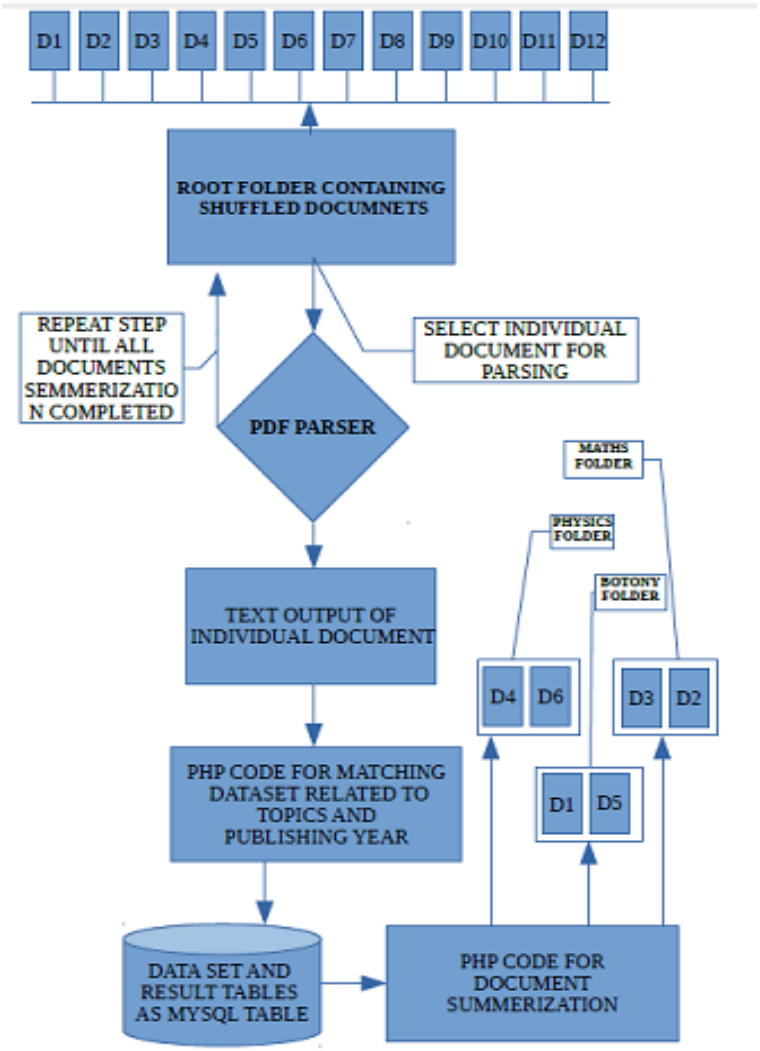

Fig: documents Subject identification and clustering system

System method contains following steps:

- In the first step, take into account a very large number of different subjects PDF papers. For instance take d1,d2,d3,d4,d5,d6,d7,d8,d9,d10,d11,d12 documents. Here these documents are jumbled.
- Firstly system provides an option for user to choose subject from index list. Index having all subject names. It includes economics, physics, biology, chemistry and history.
- User is allowed to upload single document at every instance.
- PDF document is modified to write format by the utilization of PDF parser and mPDF.
- PDF parser processes the document. It draws the content of document and also its metadata.
- Refine the distilled information.
- Database having subjected and verbs connected datasets. Next step is to number the suited keywords in the present document excluding verbs.
- Exhibits result as total number of words in the document excluding verbs, total number of words recognized, finally shows whether the present document correlated to the chosen subject or not. It makes decision by contrasting words of recognized with words of expected.

3.1. Applications

- Pattern Recognition.
- Text summarization.
- Text visualization.
- Analysis of web pages.

IV. EXPERIMENTAL RESULT

This system's input is large collection of different subject documents. They must be in PDF format. First user selects the subject from list. Later uploads the document to the system. It saves in main directory with modified name. Then PDF parser and mPDF process this document. Finally it concludes the subject of uploaded paper and put it in correlated directory. These categorized documents are helpful for further analysis and researches. Output is displayed in the form of pie charts.

Consider if one user desires to recognize the subject of one particular document. So user uploads this document to system. System processes and outputs as the total number of words in this document, number of words expected (i.e. 1% of total words), number of words expected and displays whether this document belongs to that selected subject or not. It also displays matched words and its location in the document.

V. CONCLUSION

The rapid growth of stored information in almost every area of live scenario has created a great demand for new, powerful tools for turning data into useful knowledge. This problem of information overload is further aggravated due to the unstructured, textual data form of the majority of the data. Text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. The new researcher or scholar as readers only searches for the most related published research articles of their interested areas for latest research developments in the same field. This system minimizing readers efforts deciding whether to go ahead with the retrieved articles for further readings helping in his/her own work.

This work can further be modified for future improvements as a wide range of different sub-domains of the specific field can be covered. In coming days, one can implement system that categorizes different types of documents except PDF format.

REFERENCES

[1] Han J. and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2nd Edition.
[2] Luke Welling and Laura Thomson, "PHP and MySQL Web Development", PEARSON Education, 3rd Edition.
[3] N K Nagwani, "Summarizing large text collection using topic modeling and clustering based on MapReduce framework". Journal of Big data, 2015, DOI 10.1186/s40537-015-0020-5.
[4] Ms.Sunita R. Patil, Student, NMIMS, Mumbai, Dr.Mrs.Sunita M. Mahajan, Principal, ICS, MET, Bandra, "Document Categorization using Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 2, Issue 10, October – 2013.
[5] A. Anil Kumar, S. Chandrasekhar, "Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 5, July – 2012.
[6] Anshika Singh, Dr.Udayan Ghosh, "Text Mining: a burgeoning technology for knowledge extraction", International Journal of Scientific research Engineering and Technology (IJSRET), Volume 1, Issue12, pp 022-026, March 2013.
[7] Vishal Gupta, Gurpreet S. Lehal, "A survey of text mining techniques and applications", Journal of emerging technologies in web intelligence, VOL. 1, NO. 1, AUGUST 2009.
[8] Rahul Khokale, Dr.Mohd. Atique, "Design of intelligent interface for document understanding",International Journal of Engineering Research and Technology (IJERT), Vol. 1 Issue 4, June – 2012.