ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 10, October 2016

A Comparative Review on Data Mining With Text Mining

V.Sudheer Goud[1], Prof. P. Premchand[2]

Research Scholar, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, A.P, India[1]

Professor, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, T.S, India[2]

ABSTRACT: Data mining include tools and technique for the Extraction of knowledge from enormous warehouse of data. In Data Mining different techniques that used are Association Rule Mining, Sequential Pattern Mining, Clustering, and Classification. Varieties of algorithms are developed for each of these techniques. Data mining may be defined as the discipline of extract helpful information from databases. It also called knowledge discovery. Using a mixture of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future.
Text mining is also called text data mining and it is defined as finding previously unknown and potentially useful from textual data, textual data may be either semi structured or unstructured. Text mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. It converts the words and phrases in unstructured information into numerical values which may be linked with structured information in database and analyzed with ancient data mining techniques. There are many techniques used in text mining such as information extraction, information retrieval, natural language processing (NLP), query processing, and categorization and clustering.

KEYWORDS: Data Mining, Text Mining, Techniques, Issues, Tasks

I.INTRODUCTION

The amount of data stored and presented by machine has been growing at continually growing rate for the last decade. In the business society, companies collect all sorts of information about the business process such as Financial, payroll, and customer data. The data is often among the most valuable assets of a business. In the scientific community, a single experiment can produce terabytes of data. Subsequently, there is growing demand for methods and tools that analyze large volumes of data. However, even storing, let alone analyzing, such huge amounts of data presents many new obstacles and challenges. An often used metaphor that we are drowning in data, and yet starving for knowledge sums up the situation perfectly. The field of data mining has emerged out of this necessity.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 10, October 2016

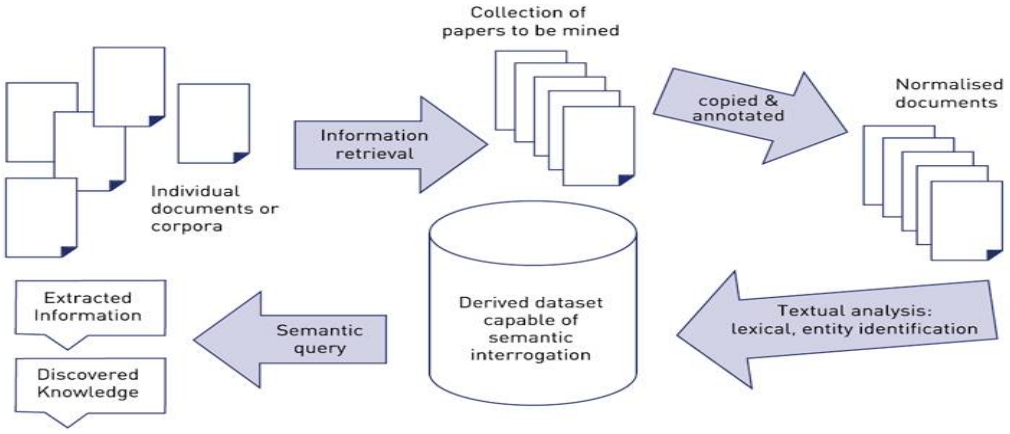


Fig 1: Text mining process

Text mining also referred as Knowledge discovery from textual data is an interdisciplinary field of Data Mining, Statistics and Machine Learning. Basically the data mining techniques are applied on three types of data for knowledge extraction, those are – Structured data (SQL database with rows and columns), Semi-structured data (CSV files, XML files, NoSQL documents), Unstructured data (Text documents, Scientific data, Photographs and Videos).
Technically Text Mining is a process of deriving high quality information from text. In this the hidden patterns and trends are extracted from the textual data. There are multiple utilities of text mining like- Text Classification, Opinion Mining, Text Clustering, and Document Summarization etc. Text mining is that the method of seeking or extracting the helpful info from the matter information. It is an exciting analysis space because it tries to find knowledge from unstructured texts. it's additionally legendary as Text data processing (TDM) and information Discovery in matter Databases (KDT). KDT plays an more and more important role in rising applications, like Text Understanding. Text mining method is same as data processing, except, the data mining tools ar designed to handle structured data whereas text mining will able to handle unstructured or semi-structured information sets like emails hypertext mark-up language files and full text documents etc. [1]. Text Mining is employed for locating the new, previously unidentified info from completely different written resources. Structured information is information that resides in an exceedingly mounted field within a record or file. This information is contained in relational database and spreadsheets. The unstructured information typically refers to info that does not reside in an exceedingly ancient row-column info and it's the other of structured information. Semi Structured data is that the typed information in an exceedingly standard info system. Text mining may be a new space of applied science analysis that tries to resolve the problems that occur within the space of data mining, machine learning, info extraction, linguistic communication process, info retrieval, information management and classification.

II.CHALLENGES FOR DATA MINING

What kinds of database to work on generally data miner can be classified according to its different kids databases relational database, transactional database, object oriented database, deductive database, spatial databases , temporal database, multimedia databases and etc.
What kind of knowledge to be mined, generally in real-time different applications requires different mining techniques to process their data.
What kind of techniques to be utilized, for effective processing appropriate technique should be used.

III.CHALLENGES FOR TEXT MINING

Text mining is still not that famous technique of data mining as others since it is used to extract knowledge from textual data i.e. unstructured data which is still a tough task because of some limitations those are

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 10, October 2016

The uppermost problem in text mining is the ambiguity of the language i.e. the capability of being understood in two or more possible sense. Because one word or phrase may have multiple meanings those can lead to ambiguity problem.
In fields like Bioinformatics there are multiple names for a single gene or protein that may also lead to ambiguity problem.
Since some filtration techniques are applied on the original data but, still sometimes some words remain in the text because they have higher frequency and affect the result.
One more problem with test mining is when we use the social media data i.e. status updates, tweets, comments, reviews etc. most people use slang words like- "btw" for by the way, "ppl" for people etc. these words do not exist in the dictionary that's why they affects the mining results.
Another problem with text mining is cleaning the data, if we extract online texts then we also get the reference addresses of the images linked with the text and those references are hard to remove.

IV.DATA MINING TASKS

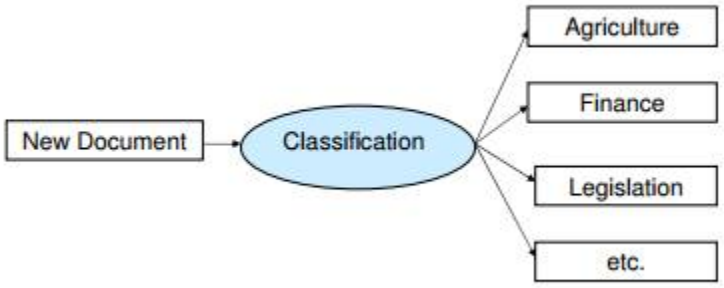
The task of data mining discovers hidden patterns in a large database, so for different patterns different kinds of methods required. So tasks in data mining can be categorized into

- Summarization
- Classification
- Clustering
- Prediction
- Association Rule
- Trend analysis.

V.TEXT MINING TASKS

Technically Text Mining is a process of deriving high quality information from text. In this the hidden patterns and trends are extracted from the textual data.

Text categorization: assigning the documents with pre-defined categories (e.g. decision trees induction).



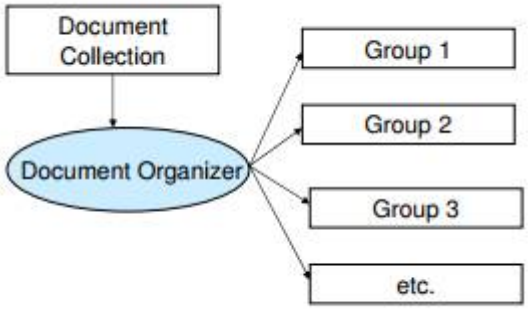
Text clustering: descriptive activity, which groups similar documents together (e.g. self-organizing maps).

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

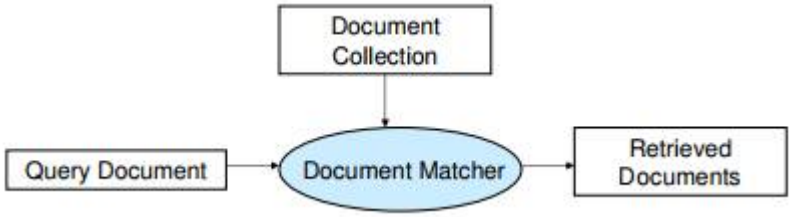
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 10, October 2016



Concept mining: modelling and discovering of concepts, sometimes combines categorization and clustering approaches with concept/logic based ideas in order to find concepts and their relations from text collections (e.g. formal concept analysis approach for building of concept hierarchy).

Information retrieval: retrieving the documents relevant to the user's query.



Information extraction: question answering.

VI.CONCLSUION

In this paper we studied challenges and tasks of data mining and text mining, and both these are little different in the way of discovering patterns and processing of input, data mining applicable for structured data where as text mining can be applicable for unstructured data. Text mining objectives are text analysis, information extraction, document mining and topic modeling.

REFERENCES

[1].Varsha C. Pande , Dr. A.S. Khandelwal ,A Survey Of Different Text Mining Techniques, IBMRD's Journal of Management and Research, Online ISSN: 2348-5922, Volume-3, Issue-1, March 2014
[2].Gobinda G. Chowdhury ,Natural Language Processing
[3]. 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
[4]. DivyaNasa, Text Mining Techniques- A Survey,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012 ISSN: 2277 128X
[5]. Falguni N. Patel, Neha R. Soni,Text mining: A Brief survey, International Journal of Advanced Computer Research, ISSNprint: 2249-7277, ISSN online: 2277-7970 Volume-2 Number-4 Issue-6 December-2012
[6] Techniques, Applications and Challenging Issue in Text Mining , Shaidah Jusoh and Hejab M. Alfawareh , College of Computer Science & Information Systems, Najran University P.O Box 1988, Najran, Saudi Arabia
[7] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 10, October 2016

[8]. ShaidahJusoh , Hejab M. Alfawareh, Techniques,Applications and Challenging Issues in Text Mining, International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November -2012 ISSN (Online): 1694-0814
[9] H. J. kamber and J. pei, Data Mining: Concepts and Techniques, San Francisco: Margan Kaufman, 2011.
[10] Text Classification Using Data Mining, S. M. Kamruzzaman1 Farhana Haider2 Ahmed Ryadh Hasan ICTM-2005.
[11] Web Data Mining, Chapter 9 Opinion Mining and Sentiment Analysis, Authors: Liu, Bing, Department of Computer Science, University of Illinois, Chicago, 851 S. Morgan St., Chicago, IL, 60607-7053, USA.

BIOGRAPHY

V.Sudheer Goud, Post Graduated in Master of Computer Application (MCA) from OU,1994, Post Graduated in Master of Business Administration (MBA) from OU, 2006, Advanced Level Course in Computer Science(ALCCS) that is equivalent to Post Graduated in Master of Technology in Computer Science & Engineering (M.Tech) from IETE , New Delhi in 2013 and Pursuing PhD in Computer Science in ANU. He is currently working as an Associate Professor, Department of Computer Science and Engineering in Holy Mary Institute of Technology and Science (HITS), Bogaram,(V), Keesara(M), Ranga Reddy .District, Telangana State, India. He has 22 years of Teaching Experience. His research interests include, Data Mining, Cloud Computing and Information Security.

Prof. P. Premchand , He is currently working as an Professor, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, Telangana State