

Techniques for Mining Text Documents

Ranveer Kaur

M.Tech, Computer Science and Engineering
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Shruti Aggarwal

Assistant Professor, Computer Science &
Engineering
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

ABSTRACT

Data mining, being the active research area deals with analysis of data so that useful, valid and novel patterns can be extracted. Data mining is the core process of KDD (Knowledge Discovery in Databases). Today, a large amount of data is available in textual format and text mining is the emerging technique in the field data mining. This paper briefly discusses the how the text mining works and various techniques for text mining.

General Terms

Data Mining, Text Mining

Keywords

Text Mining, Text Mining Techniques, Information Extraction, Topic Tracking, Association Rule Mining

1. INTRODUCTION

Data Mining is one of the important research areas. Data Mining is formed by the combination of words Data+Mining, which means, it mines the data to get knowledge from it because data is of no use unless knowledge can be fetched out of it. This seems to be important by taking into account the consideration that “we are data rich but information poor” [8]. Data mining is also called as Knowledge Discovery from Databases (KDD).

Data Mining being quite similar to KDD but it always varies from the KDD in a manner that KDD is complete process by which we get knowledge regarding any specific domain or application so that, this knowledge can be used for decision making purposes in business, industries etc. While Data Mining is a step in KDD process aimed at discovering patterns and relationships from pre-processed and transformed data [1].

Hence Knowledge Discovery=Data Preparation + Data Mining+ Evaluation/Interpretation of patterns. Complete process can be shown in Fig1.

We always mine the data by taking an assumption that knowledge will be discovered from it. The data available in the real world can be of any format like text data, online data, audio/video data, spatial data (like maps), time series data, relational and transactional data etc.

We can extract knowledge from any kind of data by using certain tools. So types of data mining systems depending on type of data can be divided as:

- **Text Data Mining:** The mining process applied on free text to get useful relationships from it.
- **Multimedia Data Mining:** the process of finding important patterns form audio/video data.

- **Web Mining:** when data mining process is applied on web (online) data, then it is called as Web Mining.
- **Spatial Data Mining:** When type of data used to extract important features is in the form of maps, geographical information then it is called as spatial data mining.

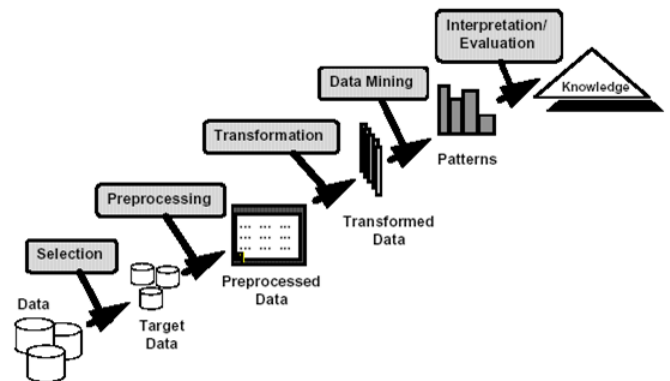


Fig 1. Steps in KDD process [1]

So the basic goals of Knowledge Discovery (Data Mining) process are Verification and Discovery [1]. In Verification, User's hypothesis is tested only for the verification; it should be either accepted or rejected. In discovery, new patterns are generated autonomously by the system itself.

A. Data Mining Task

Discovery goal of Data Mining process can be divided further into two different goals/tasks performed by Data Mining System [1][7]:-

1. Descriptive Modelling
2. Predictive Modelling

In Descriptive Modelling, data is described completely so that it can be presented in human understandable form [2][5].

In Predictive Modelling, value of one variable is predicted from unknown values i.e. future data is predicted [1][7].

B. Data Mining Methods

Description and Prediction goals are achieved via following Data Mining methods [1][7]:-

Table1. Text Mining Methods

Data Mining Task	Goal	Importance
Classification	Predictive	Finding a model that will classify the data into predefined classes.
Clustering	Descriptive	Identify the finite set of clusters to describe the data.
Association Rule Discovery	Descriptive	Finding relationships between the data.
Regression	Predictive	Learning a function which maps a data item to a real-valued prediction variable and finding relationships between variables.
Deviation Detection	Predictive	Find all the significant changes form previously measured values.

2. TEXT MINING

As there is huge growth in web, digital libraries, Medical data hence online textual documents are becoming more effective. So extracting knowledge out of these text documents is one of major research area. An effective reader will always generate relationships between texts so that a hypothesis can made. The extraction of useful patterns out of the textual resources is known as Text Mining. Text Mining is also known as Intelligent Text Analysis, Text Data Mining, and Knowledge Discovery in Text (KDT).

This paper basically focuses on reviewing the various fields that are active in Text Mining and how the text mining works.

2.1 Text Mining Working

Text Mining applies various techniques from different areas to extract semantically useful information from the text [5]. These techniques have been taken from the areas like Information Retrieval, Natural Language Processing (NLP), Information Extraction and Data Mining.

2.1.1 Information Retrieval:

IR is term used for finding all those documents that satisfy the user's query. Hence IR system reduces our search to the limited documents only. If IR and Text Mining are used together then Text Mining system can produce efficient results and IR can speed up this process.

2.1.2 Natural Language Processing

NLP is analysis of natural languages so that computers can understand them. NLP can perform Part-of-speech tagging, Parsing, Word Sense Disambiguation.

2.1.3 Information Extraction

IE is the process of gathering structured information from unstructured text. NLP extracts the part of speech tags, Parsing results and these results are passed through IE phase. IE can perform Term Analysis, Name Entity Recognition, and Fact analysis.

2.1.4 Data Mining

Data Mining is process of extracting useful and novel information from the database. When DM is used with Text Mining then Data Mining is applied on the facts generated from the IE phase to give the useful patterns.

A text mining process can be shown as:

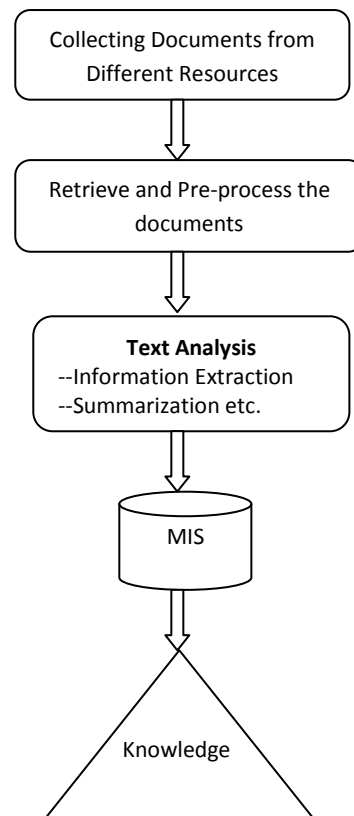


Fig2. Process of Text Mining [3][6]

Text mining process starts with collection of documents from various sources. With rapid development of the internet, textual data has been easily available from the online resources. Data can also be collected from the books, magazines, journals, newspapers etc. Any of the text mining tools is applied to fetch a particular document which further undergoes the pre-processing phase so as to improve the efficiency of the system. Then text analysis phase is executed to extract the information which would completely define the document content. This phase includes the process like information extraction, summarization, and feature selection etc and depending on the goals of organization, these techniques can be used in the hybrid way. Then Management Information System will handle the resulting data leading to the vast amount of the knowledge. The knowledge so produced after the text mining process can be used for decision making purposes [6].

3. TEXT MINING TECHNIQUES

The knowledge extracted from the text mining process can help the businesses in decision making process. Various Techniques have been suggested for text mining like Information Extraction, Topic Tracking, Summarization, Categorization, Clustering, Information Visualization, Concept Linkage, Question Answering and Association Rule Mining [3][6]. Some of these are described here:

3.1 Information Extraction [3][6]:

Being the initial point, Information Extraction software finds key phrases and relationships from text. Pattern Matching is the base process that is used for Information Extraction which includes extracting the predefined sequences from the text.

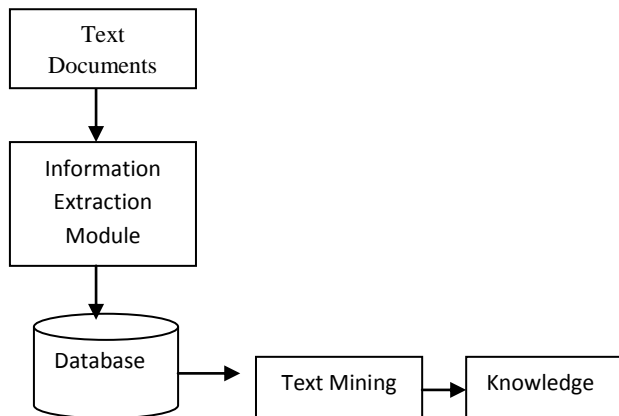


Fig3. Information Extraction Process [6]

Fig 1 depicts that a large corpus consisting of textual documents, when given as input to the Information Extraction module imposes a problem of transforming it to structured database and database so formed can be given as input to the knowledge discovery tool to generate important information out of it .e.g. Ratan Naval Tata, KBE is an Indian businessman who became chairman of the Tata Group. Information Extraction software should identify people, places and other meaningful information out of the text like from the given example:

Table2: Information Extracted

Ratan Tata	Indian Businessman
Tata Group	Chairman

Information Extraction is highly valuable when dealing with large volume of text because data in today's world is mainly available in form of electronic documents which has large amount of text.

3.2 Topic Tracking [3][6]

Topic Tracking is text mining technique in which documents of user's interest are presented to user, based on tracking all those documents that user views. This approach can be used in the companies for alert purposes.

Various approaches have been used for Topic tracking like Vector Space Model, Hierarchical Clustering, Named Entity Recognition Model, Hidden Markov Model, and Based on Keyword Extraction System etc[10].

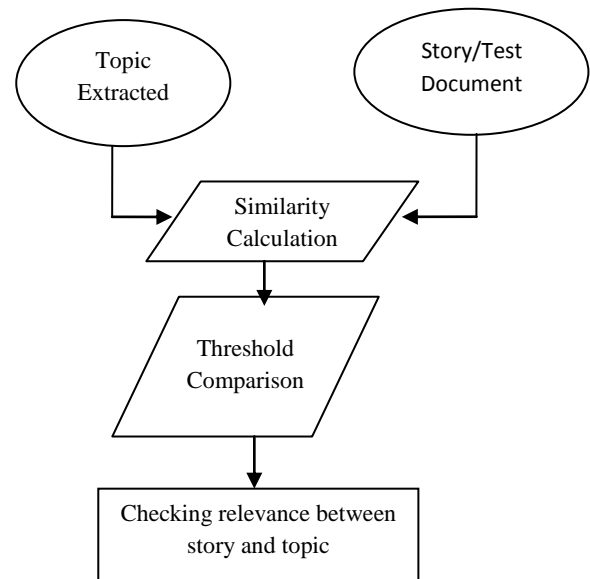


Fig4. Topic Tracking System Architecture [10]

While doing Topic Tracking, Test Document needs to be represented and Similarity is calculated with the help of similarity function. This similarity is between topic and story and then threshold comparison is performed. If similarity is higher than threshold value then story is found to be related to topic otherwise not related to topic.

Topic tracking system can be implemented on any textual database for tracking the events. It helps one to keep updated with all the products in the market.

It can be used in the medical industry to track the complete situation of patient and what procedure has been followed and what are new treatments. It can also be used for education purposes. In news industry, it is highly valuable technique to find which news articles tracks same events and helps to collect distributed information together.

3.3 Summarization [3][6]

Text Summarization is process of expressing large textual documents into reduced length documents while overall meaning remain same. Various techniques can be used for text summarization like sentence extraction i.e. extracting important sentences from an textual document by statistical calculation for sentences like weighting scheme, TF-ISF (Term Frequency- Inverse Sentence Frequency) Further heuristics such as position weighting scheme can also be used for summarization. For example, those phrases which are followed by key phrases like "in conclusion", "at last", "finally", "in the end" etc. depicts the main points of document.

Various methods like statistical, linguistically, heuristic methods are used for text summarization where this system finds how often certain keywords are. Frequency of the keywords is calculated, in which sentence they are present, check for bold text tag etc. This information is helpful to generate summarized view of the original text.

An automatic summarization [11] process can be divided into three steps: (1) Preprocessing step where a structured representation of the original text is obtained; (2) Processing

step where an algorithm only transforms original text into the summary structure of original document; and (3) Generation step the final summary is generated.

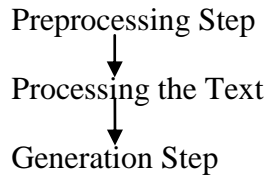


Fig5. Automatic Text Summarization [11]

3.4 Clustering [3][6]

The process in which similar documents are grouped together and dissimilar documents are placed in different cluster.

3.5 Categorization [3][6]:

The Process of placing a document into predefined topic set is called as categorization. It can also be called as classification process.

3.6 Association Rule Mining [3][6]:

Technique by which important association rules i.e. relationships are extracted from large databases is called as Association Rule Mining (ARM). It has been widely used in decision making process in business. E.g. these relationships are currently implemented in various supermarkets where items are placed on the basis of purchasing habits of customers i.e. those items are placed at minimum distance which are purchased frequently [8].

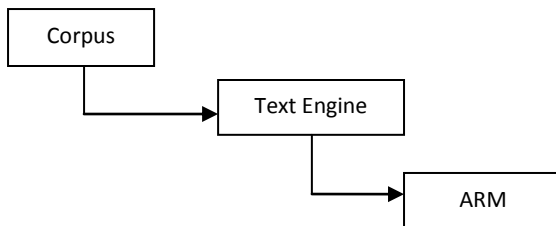


Fig6. Association System

Fig6 shows A corpus is given as input to the text engine for extracting topics out of it. Text mining engine then reads those topics and generates association from those topics. These results are then fed into visualization system for analysis purpose.

3.6.1 ARM application

ARM has been implemented in business for decision making purposes. ARM can also be used in textual databases for generating rules out of the text. This process has been used on many languages like Chinese, Urdu etc. Other languages can also be used to generate association rules which help to generate:

- Statistical thesaurus for any language
- Extracting grammatical rules
- Searching of web textual content in efficient manner

This technique of text mining can be used to generate strong Association Rules for Punjabi text database. ARM will be

combined with Natural language processing techniques to generate a system for generating associations

4. TEXT MINING APPLICATIONS

Text mining, being the part of data mining field seeks to gather knowledge from structured as well as unstructured text. Advancements in text mining technologies have been led to process the natural language text also. There are many applications of text mining in various sectors like banks, insurance and financial markets, healthcare industry, IT sector, telecommunications etc. [2]. These applications are briefly discussed below:

4.1 Market Analysis

With the help of various text mining techniques, market analysis is concerned to analyse the competitors in the market and can also be used to monitor the customers opinions and searing for new potential customers [2][6].

4.2 Customer Relationship Management

It mainly deals with managing the client messages. CRM consists of providing appropriate service to the customer as per their requests and providing quick answers to the questions. CRM need is basically felt in banking and insurance sector [2][6].

4.3 Competitive Intelligence

In this competitive era, each business organization need to gather, analyse the information about themselves, their competitors, customers, products and environment needed to take decisions efficiently [2][6]. So CI aims to find relevant information from the available data. CI is required in medical and pharmaceutical industry can be used to analyse the medical documents for the extraction of articles, scientific abstracts etc [2][6].

4.4 Natural Language Questioning

One of the text mining applications in Natural Language Processing is to develop the websites which can support the questioning in natural language [2][6]

5. CONCLUSION

Data mining being the important as well as active research area helps to extract useful patterns from the data. These patterns generated help in decision making process in industry. Text data mining is also an important field which deals with unstructured and semi-structured text. This paper describes that Text data mining is always used with various fields like Natural Language Processing, Data Mining, Information Extraction, and Information Retrieval. Various techniques for Text mining has been also explained briefly which includes Information Extraction, Topic Tracking, Summarization, Clustering, Categorization and Association Rule Mining.

Association rule mining has various applications in finding relationship from given text. So this technique can also be used to find relationships in natural languages like Punjabi. Not much work has been done on Punjabi language. So this work can be done in future.

6. REFERENCES

- [1] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smith, "Knowledge Discovery and Data Mining: Towards a Unifying Framework", KDD-96 Proceedings, 1996.
- [2] Sergio Bolasco, Alessio Canzonetti, Federico M. Capo, Francesca Della Ratta-Rinaldi, Bhupesh K. Singh, "Understanding Text Mining: a Pragmatic Approach", Roam, Italy, 2002.
- [3] Weiguo Fan, Linda Wallace, Stephanie RichZhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005
- [4] Seth Grimes, "The Developing Text Mining Market", white paper, Text Mining Summit05 Alta Plana Corporatopn, Boston, 1-12, 2005.
- [5] Ananiadou, Sophia and McNaught, John(eds), Text Mining, March 2006.
- [6] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Texhnologies in Web Intelligence, Vol. 1, No. 1, August 2009.
- [7] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.
- [8] Jiawei Han, Michelin Kamber, 2001, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, USA, 70-181.
- [9] [online] www.wikipedia.org/
- [10] Kamaldeep kaur and Vishal Gupta, "Topic Tracking for Punjabi Language", Computer Science and Engineering: An International Journal (CSEIJ), Vol. 1 No. 3, August 2011.
- [11] Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravayan Dehkordy and Asghar Tajoddin (2008), "Optimizing Text Summarization Based on Fuzzy Logic", Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE computer society, 347-352.
- [12] Nazish Asad, M. Younus Javed and Usman Qamar, "Association Rules Mining for Urdu Language", International Journal of Computer and Communication Engineering (IJCEE), ISSN 2010-3743, Vol 1, No. 1, May 2012