Text Mining: Concepts, Applications, Tools and Issues – An Overview

K.L.Sumathy
Research and Development Center
Bharathiyar University Coimbatore

M.Chidambaram, Ph.d
Assistant Professor/CS
Rajah Serfoji Govt. College Thanjavur

ABSTRACT

Nowadays there is an increasing trend in the usage of computers for storing documents. As a result of it substantial volume of data is stored in the computers in the form of documents. The documents can be of any form such as structured documents, semi-structured documents and unstructured documents. Retrieving useful information from huge volume of documents is very tedious task. Text mining is an inspiring research area as it tries to discover knowledge from unstructured text. This paper gives an overview of concepts, applications, issues and tools used for text mining.

General Terms

Document classification, Retrieval

Keywords

Text Mining, Information Retrieval

1. INTRODUCTION

In the contemporary world the text is the most common means for exchanging information. The data stored in the computer can be in any one of the form (i) structured (ii) semi structured and (iii) unstructured. The data stored in databases is an example for structured datasets. The examples for semi structured and unstructured data sets include emails, full text documents and HTML files etc. Huge amount of data today are stored in text databases and not in structured databases. Text Mining is defined as the process of discovering hidden, useful and interesting pattern from unstructured text documents. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining. Approximately 80% percent of the corporate data is in unstructured format. The information retrieval from unstructured text is very complex as it contains massive information which requires specific processing methods and algorithms to extract useful patterns. As the most likely form of storing information is text, text mining is considered to have a high value than that of data mining. Text mining is an interdisciplinary field which incorporates data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing. The remainder of this paper is organized as follows: Section 2 describes the general frame work for text mining. Section 3 explains the steps involved in text mining, section 4 discusses areas of text mining, section 5 explains the Applications & tools used for text mining and section 6 enlightens the challenges in text mining. Section 7 concludes the paper.

2. GENERAL FRAMEWORK FOR TEXT MINING

Text Mining can be visualized as consisting of two phases: (i) Text refining and (ii) Knowledge distillation. Text refining phase transforms the free form text documents into a chosen intermediate form. Knowledge distillation infers patterns or knowledge from intermediate form. The Intermediate Form (IF) can be semi structured such as the conceptual graph representation or structured such as relational data representation. Intermediate form can be document based wherein each entity represents a document or concept based wherein each entity represents an object or concept of interests in specific domain. Mining a document based IF derives patterns and relationships across documents. Document clustering/visualization and categorization are examples of mining from a document based IF.

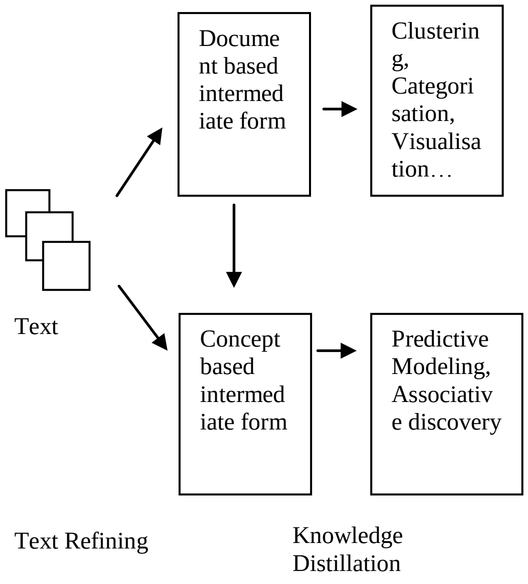
Document based intermediate form
Clustering, Categorisation, Visualisation…
Text
Concept based intermediate form
Predictive Modeling, Associative discovery
Text Refining
Knowledge Distillation

Figure 2.1 General framework for Text Mining

Mining a concept based IF derives pattern and relationship across objects or concepts. Data mining operations such as predictive modeling and associative discovery belong to this category. A document based IF can be converted into a concept based IF by extracting the relevant information according to the objects of interests in a specific domain. It follows that document based IF is usually domain-independent and concept based IF is domain dependent. For example given a set of news articles, text refining first converts each document into a document based IF. One can then perform knowledge refinement on the document based IF for the purpose of organizing the articles, according to their content for visualization and navigation purposes For Knowledge discovery in a specific domain the document based IF of the news articles can be projected onto a concept based IF depending on the task requirement. For example one can extract information related to "product" from the document based IF and form a product database to derive product based knowledge. The general framework for Text Mining is shown in Figure 2.1

3. STEPS INVOLVED IN TEXT MINING

The steps involved in the process of text mining is shown below

3.1 Text preprocessing

The text preprocessing step is further divided into

a.tokenisation b.stopword removal c.stemming.

a.tokenisation

Text documents contain a collection of statements. This step segments the whole text into words by removing blank spaces, commas etc.

b.stopword removal

This step involves removing of HTML, XML tags from web pages. Then the process of removal of stop words such as 'a', 'is', 'of' etc is performed.

c.stemming

Stemming refers to the process of identifying the root of a certain word. There are basically two types of stemming (i) inflectional and (ii) derivational. The most commonly used algorithm is porter's algorithm for stemming.

3.2 Text transformation

Text document is represented by the words it contains and their occurrences. Two approaches used for document representation are a.bag of words b. vector spaces.

3.3 Feature selection

It is also known as variable selection. It is the process of selecting a subset of important features for use in model creation. This phase mainly performs removing features which are redundant or irrelevant. feature selection is the subset of more general field of feature extraction.

3.4 Text mining methods

At this point Text mining becomes data mining. Data miming methods such as clustering, classification information retrieval etc., can be used for text mining.

3.5 Interpretation/Evaluation

Analyzing the results.

4. AREAS OF TEXT MINING

Text mining is an interdisciplinary field which incorporates areas such as information retrieval, information extraction, data mining, computational linguistics and natural language processing. The areas of text mining is shown in figure 4.1

4.1 Information Extraction (IE)

It is the process of automatically extracting structured information from unstructured and/or semi structured text documents. An IE system involves identifying entities such as names of people, companies and location, attributes and relationship between entities. It does this by pattern recognition. It is the process of searching for predefined sequences of text in text documents. Since information extraction addresses the problem of transforming a corpus of Textual documents into a more structured database, the database constructed by an IE module can be provided to KDD module for further mining of knowledge as shown below.

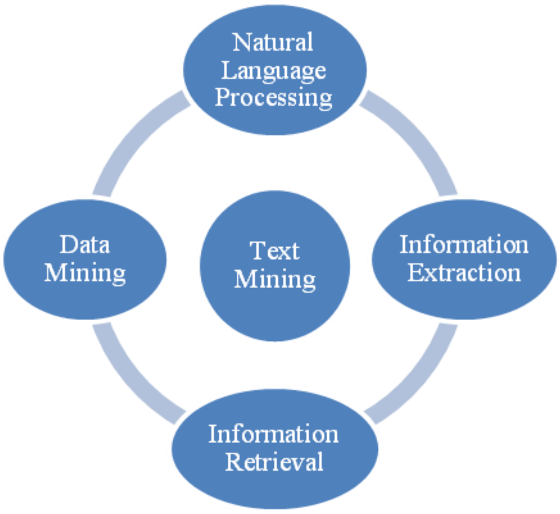

Figure 4.1 Areas of Text Mining

4.2 Information Retrieval (IR)

It is defined as the methods used for representation, storage and accessing of information items where the information handled is mostly in the form of textual documents, newspapers ad books which are retrieved from databases according to the user request or queries. Information Retrieval is considered as an extension to document retrieval where the documents are processed to condense or extract the particular information requested by the user. An IR system allows us to narrow down the set of documents that are relevant to a specific problem. The most well known IR systems are search engines such as Google. IR systems can speed up the analysis significantly by reducing the number of documents for analysis.

4.3 Natural Language Processing (NLP)

Natural Language Processing is the most challenging problem in the field of artificial intelligence. It is the study of human language so that computers can understand natural languages similar to that of humans. Natural Language Processing is concerned with Natural Language Generation (NLG) and Natural Language Understanding (NLU) . NLG makes sure that generated text is grammatically correct and fluent. Most NLG systems include a syntactic realizer to ensure that grammatical rules such as subject verb agreement are obeyed and text planner to decide how to arrange sentences, paragraph and other parts coherently. The best Known NLG application is machine translation.NLU consists of at least one of the following components: a.tokenizer, lexical analyzer, syntax analyzer and semantic analyzer.

4.4 Data Mining

It refers loosely to finding relevant information or discovering knowledge from large volumes of data. Data mining attempts to discover statistical rules and patterns automatically from data. Data mining tools can predict behaviors' and future trends allowing businesses to make positive knowledge based decisions. The overall goal of data mining process is to extract information from a data set and transform it into an understandable structure for further analysis.

5. APPLICATIONS AND TOOLS OF TEXT MINING

5.1 Applications of Text Mining

Text mining applications are most often used in the

Following sectors

(i) Telecommunications, energy and other service industries.

(ii) Information Technology sector and Internet

(iii) Publishing and media

(iv) Banks, insurance and financial markets.

(v) Political instituions, political analysis, public administration and legal documents

(vi) Pharmaceutical and research companies and health care.

(vii) Bio-Informatics, Business Intelligence and national security

5.2 Tools used in Text Mining

A high level overview of text mining tools are presented to provide a comparison of text mining capabilities perceived strengths, potential limitations, applicable data sources and output results as applied to chemical biological and patent information . Examples of tools are given below include organization name tool function output and website reference

Organization	Tool	Function	Output	website
Megaputer	Text analyst	Grouping document based semantic network	To mine knowledge hidden in text collection to make better business decision	http://www.megaputer.com
SAS Institute	Text finder	Develop High	Quick search of specific	http://wwww.parcel.com
IBM	Intelligent Miner	Information retrieval,summarisation	Document based clustering and categorization	http://www.ibm.com/software/data/iminer
Carrot search	Carrot	Cluster small collection of documents	Search result or document abstract based on themes	Http://carrot2.org
Rapid-I inc	Rapid Miner	Clustering ,classification and sentimental analysis of text documents	News filtering and E-mail spam detection	http://rapid-i.com
sysomos	sysomos	To identify how the sentiment towards the identical subjects developed	Social media analysis and product review analysis	http://www.sysomos.com

6. ISSUES IN TEXT MINING

The major challenging issue in text mining arises from the complexity of a natural language itself. The natural language is not free from ambiguity problem. Ambiguity means the capability of being understood in two or more possible ways. In a text document one word may have multiple meanings and one phrase or sentence can be interpreted in different ways which led to various meanings of statement. Although a number of researches have been conducted in resolving the ambiguity problem the work is still immature. Some of the other issues are

(i) Intermediate form

Different text mining uses different Intermediate forms. To obtain fine grain domain specific knowledge it is essential to perform semantic analysis. But semantic analysis methods are computationally expensive and often operate in the order of few words per second. It is a big challenge how semantic analysis can be made more efficient and scalable for very large corpus of text documents.

(ii) Multilingual text refining

Since text mining involves a significant language component it is essential to develop text refining algorithms that process multilingual text documents and produce language independent intermediate forms.

(iii) Domain Knowledge Integration

Integrating domain knowledge in text mining tool plays an important role in text mining. So far no text mining tool has included the domain knowledge.

(iv) Personalized autonomous mining

Current text mining products and applications are still tools designed for trained knowledge specialists. Future text mining tools as part of the knowledge management systems should be usable by technical users as well as management executives. There have been some efforts in developing systems that interpret natural language queries and automatically perform the appropriate mining operations.

7. CONCLUSION

Due to the rapid growth of digital data made available in recent year’s knowledge discovery and data mining have attracted great attention with a forthcoming need for turning data into useful information and knowledge. Consequently there is growing research interest in the topic of text mining. In general text mining consists of analyzing large amount of text documents by extracting key phrases; concepts etc., and prepare the text processed for further analysis with data mining techniques. In this paper an overview of concepts, applications, tools and issues of text mining is presented to give the researchers to carry it to the next level.

8. REFERENCES

[1] Vishal Gupta,Gupreet S Lehal. " A survey of Text Mining Techniques and Applications". Journal of Emerging Technologies in web inteliignce, No.1, August 2009.

[2] Vidya k A,G Aghila, "Text Mining Process,Techniques and Tools: an overview",International journal of information technology and knowledge management ,july-december 2010,volume 2,no2,pp.613-622.

[3] Ah-hwee Tan, "Text Mining:The state of the art and the challenges",In proceedings of the PAKDD workshop on Knowledge discovery from advanced databases,pp.65-70,1999.

[4] Vishwadeepak singh baghela, Dr.s.p.tripathi," International journal of computer science issues",vol.9,issue3,pp.545-552,may2012.

[5] Vallikannu ramanathan, T.Meyyappan,"International conference on technology and business management" pp.508-514,March 2013.

[6] Lokesh kumar,Parul kalraBhatia,"Text Mining :concepts,process and applications " , Journal of global research in computer science ,pp.36-39,march 2013.

[7] Falguni N.patel,Naeha R.soni,"International journal of Advanced computer research", vol.2, pp.243-248, december 2013.