Text Mining

By
YatishaBhoir
Under the guidance of Prof.MadhuriGedam
Department of Computer Engineering,
Shree L R Tiwari College of Engineering,
Kanakia Park, Mira Road(E), Thane - 401107

Abstract: Text mining has become an exciting research field as it tries to discover valuable information from unstructured texts. The unstructured texts which contain vast amount of information cannot simply be used for further processing by computers. Text mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics.Therefore, exact processing methods, algorithms and techniques are important to extract this valuable information which is completed by using text mining. In this paper, we have discussed about the general idea of text mining, it's techniques and applications.

***Keywords**: Retrieval, Extraction, Categorization, Clustering, Summarization, Visualization.*

*****

1. Introduction

Text mining has become important research vicinity. A very large number of information stored in different places in unstructured structure. Approximately 80% of the world's data is in unstructured text [1]. As the most natural form of storing information is text, text mining is believed to have a commercial potential higher than that of data mining. In fact, a recent study indicated that 80% of a company's information is contained in text documents. Text mining, however, is also a much more complex task (than data mining) as it involves dealing with text data that are inherently unstructured and fuzzy. Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining [3]. The unstructured text cannot be easily used by computer for more processing. So, there is a need for some technique that is useful to extract some precious information from unstructured text. This information is then stored in text database format which contains structured and few unstructured fields. Text can be sited in mails, charts,

SMS, newspaper articles, journals, product reviews, and organization records [2].

There are five basic text mining steps as under:

Text mining steps:

a) Collecting information from unstructured data.

b) Convert this information received into structured data

c) Identify the pattern from structured data

d) Analyse the pattern

e) Extract the valuable information and store in the database [4].

The overall goal is, essentially, to turn text into data for analysis, via application of naturallanguageprocessing (NLP) and analytical methods.

2. A Framework Of Text Mining

Text mining consists of two phases that is text refining and knowledge distillation. Text refining transforms free formtext documents into intermediate form and knowledge distillation deduces patterns from the intermediate form. Intermediate form(IF) can be structured or semi-structured. Structured form contains relational data representation and semi-structured form contains conceptual graph representation. Intermediate form can be of two types, document-based wherein each entity represents a document, or concept based wherein each entity represents an object or concept of interests in a specific domain. Document based includes mining a document-based intermediate form which deduces patterns and relationship across documents. Clustering, categorization and visualization are the examples of document-based IF. Mining a concept-based IF derives pattern and relationship across objects or concepts. Predictive modelling, associative discoveryand visualization are the examples of concept-based intermediate form. A document-based IF can be transformed into a concept-based IF by realigning or extracting the relevant information per 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 [3].

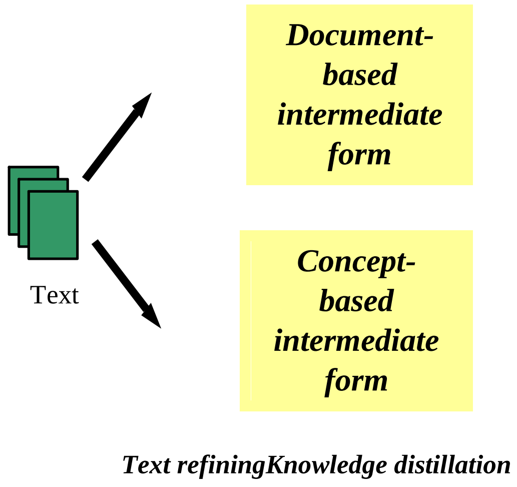
Document-based intermediate form
Text
Concept-based intermediate form
Text refiningKnowledge distillation

Figure 1: A text mining framework

Figure 1 describes that the text refining converts unstructured text documents into an intermediate form (IF). IF can be document-based or concept-based. Knowledge distillation from a document-based IF deduces patterns or knowledge across documents. A document-based IF can be projected onto a concept-based IF by extracting object information relevant to a domain. Knowledge distillation from a concept-based IF deduces patterns or knowledge across objects or concepts.

3. Data Mining vs. Text Mining

Data Mining	Text Mining
process directly	Linguistic processing or natural language processing (NLP)
Identify causal relationship	Discover heretofore unknown information
Structured numeric transaction data residing in rational data warehouse	Applications deal with much more diverse and eclectic collections of systems and formats

4. Text Mining Techniques

Text mining techniques are produced by natural language processing to analyse, understand and generate text. This technique is information extraction, information retrieval, summarization, categorization, clustering and information visualization. This are all techniques used in the data mining process per the situation that occurs. Eachof this technique plays important role in different situations of text mining. In the following sections, we discussed each of these technologies and the role that they play in text mining.

Information Retrieval

Information retrieval (IR) includes tracing and recovery of information from stored data. It is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full text or other content-based indexing.The most well-known information retrieval systems are Google search engines which recognize those documents on the World Wide Web that are associated to a set of given words. It is measured as an extension to document retrieval where the documents that are returned are processed to extract the useful information crucial for the user [5][4]. Thus, document retrieval is followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage. IR in the broader sense deals with the whole range of information processing, from information retrieval to knowledge retrieval [8][4].

Information Extraction

Information extraction identifies key phrases and relationships within text by analysing unstructured text. Process of pattern matching is used tolook for predefined sequences in text. The main goal is to extract useful information from text. Information extraction includes tokenization, sentence segmentation, identification of named entities. It identifies the useful patterns of data from semi-structures or unstructured text. Most useful information such as name of the person, location and organization are extracted without proper understanding of the text [7].

Categorization

Text categorization is a kind of "supervised" learning where the categories are known in advance and firm in progress for each training document. Then, its key projected utilize was for indexing scientific literature by means of controlled words. It was only in the 1990s that the field fully developed with the availability of continuous increasing numbers of text documents in digital form and the requirement to organize them for easier use. Categorization automatically assigns one or more category to free text document. Categorization is supervised learning method because it is based on input output examples to classify new documents. Predefined classes are assigned to the text documents based on their content. The typical text categorization process consists of pre-processing, indexing, dimensionally reduction, and classification. The main goal is to train classifier based on known examples and then unknown examples are categorized automatically [8]. There are different statistical classification techniques like Naïve Bayesian classifier, Nearest Neighbour classifier, Decision Tree can be used to categorize text.

Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. The Difference between clustering and classification is, Clustering is unsupervised learning whereas classification is supervised learning. Clustering finds a natural grouping of instances given an un-labelled data. Classification includes pre-labelled data. Clustering is basically the process of grouping physical or abstract objects into classes of similar objects. In data mining K-means is frequently used clustering algorithm, in text mining field also it obtains good results. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. The organization of management information systems makes use of clustering technology as organizational database contain thousands of documents [8].

Summarization

Summarization is the process which reduces the text document to retain most important points from the original document. It reduces the text document with a computer program. It creates summary of most important points from text documents rather than reading full document every time. So, the time would be preserved by this technique. Technologies which make summary considers variables such as length, writing style and syntax. There are online tools available for text summarization. There are professional text summarization applications available online for summarization of text. Text summarization software or application processes and summarizes the large text document. It saves the user’s time.

Summarization process include following steps:

(1) Pre-processing obtain a structured representation of the original text.

(2) To transform summary structure from text structure algorithm is applied in next processing step.

(3) In the invention step the final summary is obtained from the summary structure [8].

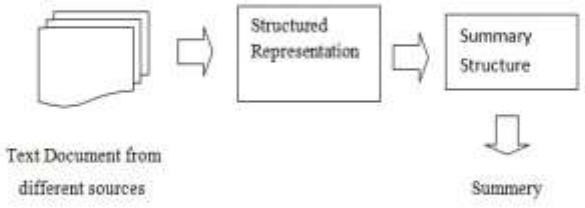
Structured Representation
Summary Structure
Text Document from different sources
Summery

Figure 2: Text Summarization

Information Visualization

Information visualization as it name describes it visualizes the text document that is it improves and simplifies the document. It discovers the relevant information into simplified form.To represent individual documents or groups of documents text flags are used to show document category and to show density colours are used. Visual text mining puts large textual sources in a visual hierarchy. The user can interact with the document by zooming and scaling. Information visualization is applicable to government to identify terrorist networks or to find information about crimes [8].

5. Applications of Text Mining

Text mining is the new area of computer science and it is useful in different area of organizations. Text mining is widely use nowadays in different areas of business. There are many applications of text mining. In the following section, we discussed the different applications of text mining.These text mining applications can give you an idea of how this technology is helping organizations today.

Risk management

Risk management is very important in every organization.Insufficient risk analysis is often a leading cause of failure. This is true in many financial industries where risk management software is based on text mining technology which can dramatically increase the ability to mitigate risk. It provides the ability to link together information and be able to access the right information at the right time.

Knowledge management

In very large documents, we can’t able to find out important information quickly. For example, in healthcare industries. Here organizations are challenged with large amount information.for example, volumes of clinical patient data. So knowledge management software based on text mining offer a clear and reliable solution.

Cybercrime prevention

Cyber-crimes are the main issues in today’s life. With the increasing use of internet, the risk cybercrimes are also increased. To deal with this, text mining is used. Today, text mining intelligence and anti-crime applications aremaking internet crime prevention easier for any enterprise and law enforcement or intelligence agencies.

Customer care service

Text mining and natural language processing are frequent applications for customer care.Today, text analytics software is frequently adopted to improve customer experience using different sources of valuable information such as surveys, trouble tickets, and customer call notes to improve the quality, effectiveness and speed in resolving problems. Text analysis is used to provide a rapid, automated response to the customer, dramatically reducing their reliance on call centre operators to solve problems.

Fraud detection through claims investigation

In text analytics, most information is collected as text. It is very effective technique. Insurance companies are taking advantage of text mining technologies. They combine the results of text analysis with structured data to prevent frauds and swiftly process claims.

Contextual Advertising

Digital advertising is growing field of application for text analytics. For contextual retargeting, different companies have made text mining the core engine.Contextual advertising provides better accuracyas compared to the traditional cookie-based approach. It completely preserves the user's privacy.

Business intelligence

Business intelligence is used to support decision making in very large companies. Text mining plays very important role here. It enables the analyst to quickly jump at the answer. Applications such as the Cogito Intelligence Platform (link to CIP) can monitor thousands of sources and analyse large data volumes to extract from them only the relevant content.

Content enrichment

While it's true that working with text content still requires a bit of human effort, text analytics techniques make a significant difference when it comes to being able to more effectively manage large volumes of information. Text mining techniques enrich content, providing a scalable layer to tag, organize and summarize the available content that makes it suitable for a variety of purposes.

Spam filtering

E-mail is effective to way to communicate with people, but it comes with a dark side: spam. Today, spam is a major issue for internet service providers**,** increasing their costs for service management and hardware\software updating; for users, spam is an entry point for viruses and impacts productivity. **T**ext mining techniques can be implemented to improve the effectiveness of statistical-based filtering methods**.**

Social media data analysis

Today, social media is one of the most common source of unstructured data; organizations have taken notice. Social media is increasingly being recognized as a valuable source of market and customer intelligence, and companies are using it to analyse or predict customer needs and understand the perception of their brand. In both needs Text analytics can address both by analysing large volumes of unstructured data, extracting opinions, emotions and sentiment and their relations with brands and products**.**

6. Merits and demerits of text mining

Merits of text mining:

i) The names of different entities and relationship between them can easily be found from the corpus of documents set using the technique such as information extraction.

ii) The challenging problem of managing great amount of unstructured information for extracting patterns e is solved by text mining[8].

Demerits of Text mining:

i) The information which is initially needed is nowhere written.

ii) To mine the text for information or knowledge no programs can be made to analyse the unstructured text directly[8].

7. Conclusion

In this paper,we have discussed about the framework, techniques, how text mining is different from data mining, applications of text mining and merits and demerits of text mining. The main goal of text mining is to extract useful patterns from large volume of data. By using different techniques of text mining, we can achieve useful patterns from large volume of text documents. There are different techniques like information extraction, information retrieval, summarization, categorization, clustering and information Visualization are used in different situations of text mining to extract useful patterns from text document. As we discussed in the applications of text mining, it is widely used in different areas of organizations. So, in today's life, text mining plays very important role in different areas of business.

References

[1] VallikannuRamanathan, T. Meyyappan "Survey of Text Mining", International Conference on Technology and Business and Management, March 2013, pp. 508-514.

[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, No 2, pp.613-622.

[3] Ah-Hwee Tan,"Text Mining:The state of the art and the challenges", Kent Ridge Digital Labs 21 Heng Mui Keng Terrace Singapore 119613.

[4] Shilpa Dang, Peerzada, Hamid Ahmad, "Text Mining: Techniques and its Application",International Journal of Engineering & Technology Innovations, Vol. 1 Issue 4, November 2014

[5] R.Sagayam, S.Srinivasan, S.Roshini, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques".

Internaltional Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5.

[6] Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:23197242, Vol 3 Issue 5, May 2014, pp.5621-5625.

[7] Vishal Gupta and GurupritLehal, “A Survey of Text Mining Techniques and Applications”, Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[8] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil ,“Text Mining Methods and Techniques”,International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014.