A SURVEY ON TEXT MINING PROCESS AND TECHNIQUES

Sathees Kumar B [2], Karthika R [1]
Asst. Professor[2], M.Phil. Scholar[1],
Department of Computer Science,
Bishop Heber College (Autonomous),
Trichirappalli-620 017

ABSTRACT

Text mining has become an important research area. It deals with machine supported analysis of text. The unstructured texts which contains massive amount of information cannot simply be used for further processing by the computer and knowledge from unstructured text completed by using text mining. It uses the techniques from information retrieval, information extraction as well as natural language processing and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics. In this paper we have discussed briefly about the text mining process and the techniques used in the text mining.

KEYWORDS

Text Mining, Data Mining, Text Mining Process, Clustering.

1. INTRODUCTION

Text mining is defined as, "The extraction of information from technical literature". It has three components such as,

- Information Retrieval
- Information Processing
- Information Integration

Text mining deals with the machine supported analysis of text. It assumes that text mining is essentially corresponds to information extraction and the extraction of facts from texts. Text mining helps to extract information from unstructured data and find a pattern which is novel and unknown earlier. The steps involved in the overall process of the text mining can be given as follows

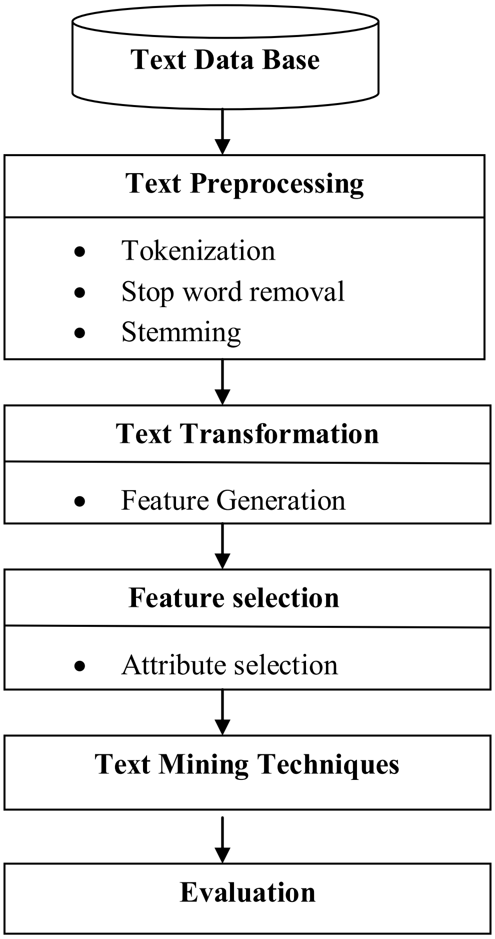

Fig 1: Flow Diagram of Text Categorization

Step 1: TEXT PREPROCESSING

Text preprocessing is the first step in the textmining, it follows three sub steps such as

1.1 Tokenization

Text document has a collection of sentences, this step divide the whole statement into words by removing spaces, commas etc.

1.2 Stop word removal

This step involves removing of HTML, XML tags from web pages and the process of removal of stop words like "a", "of" etc are performed.

1.3 Stemming

These techniques are used to find out the root or stem of a word. Stemming is the process of converting the word to their stem.

Step 2: TEXT TRANSFORMATION

Text transformation means to convert text document into the bag of words or vector space document model notation, which can be used for further effective analysis.

Step 3: FEATURE SELECTION

This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure give advantage of smaller dataset size, less computations and minimum search space required.

Step 4: TEXT MINING METHODS

There are different text mining methods as in data mining had been proposed such as clustering, classification, information retrieval, topic discovery, summarization, topic extraction.

Step 5: EVALUATION

This phase includes evaluation and interpretation of results in terms of calculating precision and recall, accuracy etc.

Data mining and Text Clustering

Data mining is the process of extracting the hidden patterns from data. It is often used to apply to the two separate processes such as,

-
- Knowledge Discovery
- Prediction

Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Predictive modeling provides predictions of future events. Text clustering is an unsupervised technique in which there is no pre-defined input and output. It is based on the concept of dividing the similar text into the same cluster. Each cluster consists of number of texts. Clustering is a technique used to group similar text documents but it differs from categorization.

2. RELATED WORKS

Navathe [1] proposed, Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text.

Liritano [2] proposed that A Cluster is a group of related documents, and clustering is the operation of grouping documents on the basis of some similarity measure, automatically without having to pre-specify

categories. The most common Clustering algorithms that are used are hierarchical, binary relational, and fuzzy. Hierarchical clustering creates a tree with all documents in the root node and a single document in each leaf node. The intervening nodes have several documents and become
more and more specialized as they get closer to the leaf nodes.

Haralampos [3] proposed Text mining is also known as Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics.

Setu Madhavi [4] proposed that Using supervised learning algorithms the objective is to learn classifiers from known examples (labeled documents) and perform the classification automatically on unknown examples (unlabeled documents).

V.Gupta [5] said, The clustering is considered better if the contents of documents of intra cluster are more similar than the contents of inter-cluster documents. Clustering is a technique used to group similar documents but it differs from categorization in than documents are clusters on the fly instead of through the use of pre-defined knowledge.

Q.Guo [6], proposed a supervised technique is one which is based upon the set of input-output examples which are basically used to train the model being used, in order to classify the new documents. In this method, pre-defined classes are assigned to the text documents. The goal is to train the classifier on the basis of known examples and then unknown examples are categorized automatically. Here for reducing the dimensionality of the document set, a method called as Index Term Selection is used.

Uma Mahesh J [7] proposed feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification and also proposed a fuzzy similarity-based self constructing algorithm for feature clustering. The words in the feature vector of a document set are grouped into cluster, based on similarity test. Each cluster is characterized by a membership function with statistical mean and deviation.

3. TEXT MINING TECHNIQUES

There are several text mining techniques are used in the text mining process some of them are given as follows,

- Supervised text categorization technique
- Pattern matching algorithm
- Support vector machine technique

3.1 Supervised text categorization technique

Supervised text categorization and clustering are closely related as both are concerned with "grouping" of objects. However, in the supervised setting, these groupings are given by common membership to a class that is assigned to sample documents before the training process starts. The training process then induces hypotheses of how the document space is shaped according to which new documents are assigned.

The Algorithm Library Component acts as the algorithmic backbone of the text mining frame work. It incorporates a number of text mining methods such as conceptual clustering, terminology extraction, pattern matching as well as machine learning techniques such as association rules and classifiers.

3.2 Pattern matching algorithm

Text mining concerns looking for patterns in unstructured text. Pattern matching is to find a pattern, which is relatively small, in a text, which is supported to be very large. Documents contain vast amounts of data that cannot be easily examined one by a human. Mining patterns from a large data set is an important system management task.

3.3 Support Vector Machine Technique

A classification task usually involves separating data into training and testing sets. Support Vector Machine (SVM) is one of the most actively developed classification technique in data mining and machine learning. The goal of SVM is to produce a model based on the training data which predicts the target values of the test data given only the test data attributes. It has been successfully applied to a wide range of pattern recognition problems.

4. SUMMARY

Text mining provides a valuable tool to deal with large amounts of unstructured text data. A major characteristic of the representation paradigm of text mining is high dimensionality of the feature space, which impose a big challenge to the performance of clustering algorithms. Text clustering is a technique that is used to group the text in similar groups. There are few advantage and disadvantage of the text mining they can be given as follows,

Advantage of text mining:

- It solved the problem of managing a great amount of unstructured information for extracting patterns easily
- Reduces the storage problems in the data base

Disadvantage of text mining:

- Programs can not be in order to analyze the unstructured text directly to mine the text for information or knowledge.
- The initial needed information is not given in the text documents.

5. REFERENCES

[1] Navathe, shamkant B and Elmasri Ramez, (2000), "Data mining and text mining in fundamental database system", pearson education pvt.inc, Singapore,841-872.

[2] Liritano S and Ruffolo M(2011), "Managing the knowledge contained in electronic documents: a clustering method for text mining", IEEE 455-458.

[3] Haralampos Karanikas and Manchester (2005), "Knowledge discovery in text and text mining software", center for research in information management.

[4] Setu madhavi and Krishna R (2008), "Experiments on supervised learning algorithms for text categorization", international conference, IEEE computer society 1-8.

[5] V.Gupta, G.S Lehal, " A survey of text mining technique and applications" in journal of emerging technologies in web intelligence, 2009.

[6] Q.Guo and W.D.S.Yu, "A novel approach to the text mining ",2010.

[7] J.Uma Mahesh and S.Lalitha, "Data mining feature clustering algorithms in text classification" in International conference on computer science and information technology, ISBN : 978-93-81693-5, 2012