A Survey of Text Mining Concepts

Dr. G. Rasitha Banu
MCA., M.Phil., Ph.D.,
Professor of Department of MCA, Vels University,
Chennai, Tamil Nadu, India

VK Chitra
MCA., Mphil Scholar,
Department of Computer Science,
Mother Teresa Women's University,
Chennai, India

Abstract - The explosive growth of databases in almost every area of human activity has created a great demand for new, powerful tools for turning data into useful knowledge. Text Mining has become an important research area. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. This survey article is intended to provide easy accessibility to the main ideas for non-experts about Text Mining Techniques, applications and Difference between text mining and data mining have been presented.

Key Words —Text mining, information extraction, Text classification (or Categorization), summarization, clustering

I. INTRODUCTION

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

Text classification is commonly used to handle spam emails, classify large text collections into topical categories, used to manage knowledge and also to help Internet search engines. A major characteristic of text categorization is high dimensionality of the feature space .the native feature space consists of hundreds of thousands of terms for even a moderate sized text collection. Various feature selection methods are discussed in this paper to overcome the problem of the high dimensionality. This survey also focuses on the various approaches and also the applications of text categorization.

Although the differences in human and computer languages are expansive; there have been technological advances which have begun to close the gap. The field of natural language processing has produced technologies that teach computers natural language so that they may analyze, understand, and even generate text. Some of the technologies [3] that have been developed and can be used in the text mining process are information extraction, summarization, categorization, clustering, and concept linkage and information visualization. In the following sections we will discuss each of these technologies and the role that they play in text mining.

II. CONCEPTS OF TECHNOLOGY

A. *TEXT MINING*

Text mining is a variation on a field called data mining, which 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.

Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding.

Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Figure 1 on next page, depicts a generic process model for a text mining application.

Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

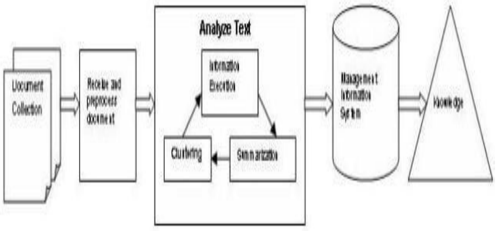

Figure 1. An example of Text Mining

Text mining is procedure of synthesizing the information by analyzing relations, the patterns and rules from the textual data. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. The functions of the text mining are text summarization, text categorization and text clustering.

B. INFORMATION EXTRACTION

A starting point for computers to analyze unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process called pattern matching. The software infers the relationships between all the identified people, places, and time to provide the user with meaningful information. This technology can be very useful when dealing with large volumes of text. Traditional data mining assumes that the information to be “mined” is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural language documents rather than structured databases. Since IE 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 the KDD module for further mining of knowledge as illustrated in Figure 2.

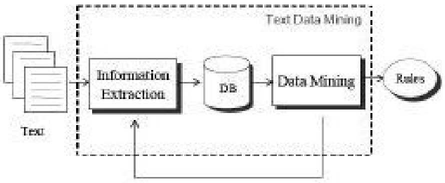

Figure 2. Overview of IE-based text mining framework

C. TEXT CATEGORIZATION

Text categorization (or text classification) is the assignment of natural language documents to predefined categories according to their content. The set of categories is often called a controlled vocabulary. Text classification is the act of dividing a set of input documents into two or more classes where each document can be said to belong to one or multiple classes. Huge growth of information flows and especially the explosive growth of Internet promoted growth of automated text classification. Development of computer hardware provided enough computing power to allow automated text classification to be used in practical applications. The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research Community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre classified documents, the characteristics of the categories.

CATEGORIZATION METHODS

Decision Trees

Decision tree methods rebuild the manual categorization of the training documents by constructing well-defined true/false queries in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. After having created the tree, a new document can easily be categorized by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf. The main advantage of decision trees is the fact that the output tree is easy to interpret even for persons who are not familiar with the details of the model. The tree structure generated by the model provides the user with a consolidated view of the categorization logic and is therefore useful information.

A risk of the application of tree methods is known as "over fitting": A tree over fits the training data if there is exist an alternative tree that categorizes the training data worse but would categorize the documents to be categorized later better. This circumstance is the result of the algorithm's intention to construct a tree that categorizes every training document correctly; however, this tree may not be necessarily well suited for other documents. This problem is typically moderated by using a validation data set for which the tree has to perform in a similar way as on the set of training data. Other techniques to prevent the algorithm from building huge trees (that anyway only map the training data correctly) are to set parameters like the maximum depth of the tree or the minimum number of observations in a leaf. If this is done, Decision Trees show very good performance even for categorization problems with a very large number of entries in the dictionary.

1. *k-Nearest Neighbor*

The categorization itself is usually performed by comparing the category frequencies of the k nearest documents (neighbors). The evaluation of the closeness of documents is done by measuring the angle between the two feature vectors or calculating the Euclidean distance between the vectors. In the latter case the feature vectors have to be normalized to length 1to take into account that the size of the documents (and, thus, the length of the feature vectors) may differ. A doubtless advantage of the k-nearest neighbor method is its simplicity. It has reasonable similarity measures and does not need any resources for training. K nearest neighbor performs well even if the category-specific documents from more than one cluster because the category contains, e.g., more than one topic. This situation is badly suited for most categorization algorithms. A disadvantage is the above-average categorization time because no preliminary investment (in the sense of a learning phase) has been done. Furthermore, with different numbers of training documents per category the risk increases that too many documents from a comparatively large category appear under the k nearest neighbors and thus lead to an inadequate categorization.

2. *Bayesian Approaches*

There are two groups of Bayesian approaches in document categorization: Naïve and non-naïve Bayesian approaches. The naïve part of the former is the assumption of word (i.e. feature) independence, meaning that the word order is irrelevant and consequently that the presence of one word does not affect the presence or absence of another one. A disadvantage of Bayesian approaches [8] in general is that they can only process binary feature vectors and, thus, have to abandon possibly relevant information.

3. *Neural Networks*

Neural networks consist of many individual processing units called as neurons connected by links which have weights that allow neurons to activate other neurons. Different neural network approaches have been applied to document categorization problems. While some of them use the simplest forms of neural networks, known as perceptions, which consist only of an input and an output layer, others build more sophisticated neural networks with a hidden layer between the two others. In general, these feed-forward -nets consist of at least three layers (one input, one output, and at least one hidden layer) and use back propagation as learning mechanism. The advantage of neural networks is that they can handle noisy or contradictory data very well. The advantage of the high flexibility of neural networks entails the disadvantage of very high computing costs. Another disadvantage is that neural networks are extremely difficult to understand for an average user; this may negatively influence the acceptance of these methods.

4. *Regression-based Methods*

For this method the training data are represented as a pair of input/output matrices where the input matrix is identical to our feature matrix A and the output matrix B consists of flags indicating the category membership of the corresponding document in matrix A. Thus B has the same number of rows like
A (namely m) and c columns where c represents the total number of categories defined. The goal of the method is to find a matrix F that transforms A into B' (by simply computing B'=A*F) so that B' matches B as well as possible. The matrix F is determined by applying multivariate regression techniques.

An advantage of this method is that morphological preprocessing (e.g., word stemming) of the documents can be avoided without losing categorization quality. Thus, regression based approaches become truly language-independent. Another advantage is that these methods can easily be used for both single category and multiple-category problems.

5. *Vector-based Methods*

We discuss two types of vector-based methods: The centroid algorithm and support vector machines. One of the simplest categorization methods is the centroid algorithm. During the learning stage only the average feature vector for each category is calculated and set as centroid-vector for the category. A new document is easily categorized by finding the centroid-vector closest to its feature vector. The method is also inappropriate if the number of categories is very large. Support vector machines (SVM) need in addition to positive training documents also a certain number of negative training documents which are untypical for the category considered. SVM is then looking for the decision surface that best separates the positive from the negative examples in the n-dimensional space. The document representatives closest to the decision surface are called support vectors. The result of the algorithm remains unchanged if documents that do not belong to the support vectors are removed from the set of training data. An advantage of SVM is its superior runtime-behavior during the categorization of new documents because only one dot product per new document has to be computed. A disadvantage is the fact that a document could be assigned to several categories because the similarity is typically calculated individually for each category.

D. SUMMARIZATION

Text summarization is immensely helpful for trying to figure out whether or not a lengthy document meets the user’s needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning. Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered. One of the strategies most widely used by text summarization tools, sentence extraction, extracts important sentences from an article by statistically weighting the sentences. Further heuristics such as position information are also used for summarization.

E. CLUSTERING

Document clustering has been investigated in different areas of text mining and information retrieval. Document clustering has been studied intensively because of its wide application in areas such as Web Mining, Search Engine and Information Retrieval. Document clustering is the automatic organization of documents into clusters or groups, so that, documents within a cluster have high similarity in comparison to one another, but are very dissimilar to documents in other clusters.

In other words, the grouping is based on the principle of maximizing intra cluster similarity and minimizing inter-cluster similarity. The major challenge of clustering is to efficiently identify meaningful groups that are concisely annotated.

III. DIFFERENCE BETWEEN TEXT MINING AND DATA MINING

The difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. One application of text mining is in, bioinformatics where details of experimental results can be automatically extracted from a large corpus of text and then processed computationally. Text-mining techniques have been used in information retrieval systems as a tool to help users narrow their queries and to help them explore other contextually related subjects.

Text Mining seems to be an extension of the better known Data Mining. Data Mining is a technique that analyses billions of numbers to extract the statistics and trends emerging from a company's data. This kind of analysis has been successfully applied in business situations as well as for military, social, government needs. But, only about 20% of the data on intranets and on the World Wide Web are numbers - the rest is text. The information contained in the text (about 80% of the data) is invisible to the data mining programs that analyze the information flow in corporations. Text mining tries to apply these same techniques of Data mining to unstructured text databases. To do so, it relies heavily on technology from the sciences of Natural Language Processing (NLP), and Machine Learning to automatically collect statistics and infer structure and meaning in otherwise unstructured text. The usual approach involves identifying and extracting key features from the text that can be used as the data and dimensions for analysis. This process is called feature extraction, is a crucial step in text mining.

Text mining is a comprehensive technique. It relates to data mining, computer language, information searching, natural language comprehension, and knowledge management. Text mining uses data mining techniques in text sets to find out connotative knowledge. Its object type is not only structural data but also semi structural data or non-structural data. The mining results are not only general situation of one text document but also classification and clustering of text sets.

IV. SOME OF TEXT MINING APPLICATIONS

The main Text Mining applications are most often used in the following sectors:

- Publishing and media.
- Telecommunications, energy and other Services industries.
- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Political institutions, political analysts, public administration and legal documents.
- Pharmaceutical and research companies and Healthcare.

The sectors analyzed are characterized by a fair variety in the applications being experimented. However, it is possible to identify some sectorial specifications in the use of TM, linked to the type of production and the objectives of the knowledge management leading them to use TM. The publishing sector, for example, is marked by prevalence of Extraction Transformation Loading applications for the cataloguing, producing and the optimization of the information retrieval.

In the banking and insurance sectors, on the other hand, CRM applications are prevalent and aimed at improving the management of customer communication, by automatic systems of message re-routing and with applications supporting the search engines asking questions in natural language. In the medical and pharmaceutical sectors, applications of Competitive Intelligence and Technology Watch are widespread for the analysis, classification and extraction of information from articles, scientific abstracts and patents. A sector in which several types of applications are widely used is that of the telecommunications and service companies: the most important objectives of these industries are that all applications find an answer, from market analysis to human resources management, from spelling correction to customer opinion survey.

(i) Text Mining Applications in Knowledge and Human Resource management Text Mining is widely used in field of knowledge and Human Resource Management.

(ii) Extraction Transformation Loading: Extraction

Transformations loading are aimed at filing non structured textual material into categories and structured fields. The search engines are usually associated with ETL that guarantee the retrieval of information, generally by systems foreseeing conceptual browsing and questioning in natural language. The applications are found in the editorial sector, the juridical and political document field and medical health care. In the legal documents sector the document filing and information management operations deal with the particular features of language, in which the identification and tagging of relevant elements for juridical purposes is necessary

(iii) Human resource management:

TM techniques are also used to manage human resources strategically, mainly with applications aiming at analyzing staff’s opinions, monitoring the level of employee satisfaction, as well as reading and storing CVs for the selection of new personnel. In the context of human resources management, the TM techniques are often utilized to monitor the state of health of a company by means of the systematic analysis of informal documents.

V. CONCLUSION AND FUTURE WORK

At last we conclude that, Text mining is also known as Text Data Mining or Knowledge-Discovery in Text (KDT), can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge.

The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. The Text-base navigation enables users to move about in a document collection by relating topics and Categorization is the operation that we use, when we want to classify documents into predefined categories.

Due to this, we are able to identify the main topics of a document collection. The categories are either preconfigured (by the programmer) or left for the user to specify. 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. It is very useful when we are exploring a new document collection and want to get an overview of the collection. The most important factor in a Clustering algorithm is the similarity measure. All Clustering algorithms are based on similarity measures. Summarization is the operation that reduces the amount of text in a document while still keeping its key meaning. With this operation the user usually is allowed to define a number of parameters, including the number of sentences to extract or a percentage of the total text to extract.

Text categorization plays a very important role in information retrieval, machine learning, text mining and it have been successful in tackling wide variety of real world applications. Key to this success have been the ever-increasing involvement of the machine learning community in text categorization, which has lately resulted in the use of the very latest machine learning technology within text categorization applications.

REFERENCES

[1] Berry Michael W., (2004), “Automatic Discovery of Similar Words”, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, 24-43.

[2] Navathe, Shamkant B., and Elmasri Ramez, (2000), “*Data Warehousing and Data Mining*”, in “*Fundamentals of Database Systems*”, Pearson Education pvt Inc, Singapore, 841-872.

[3] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), “*Tapping into the Power of Text Mining*”, Journal of ACM,Blacksburg.

[4] Sergio Bolasco, Alessio Canzonetti, Francesca DellaRatta-Rinald and Bhupesh K. Singh, (2002), “Understanding Text Mining: a Pragmatic Approach”, Roam, Italy.

[5] Liu Lizhen, and Chen Junjie, China (2002), “ Research of Web Mining”, Proceedings of the 4th World Congress on Intelligent Control and Automation, IEEE, 2333-2337.

[6] Haralampos Karanikas and Babis Theodoulidis Manchester, (2001), “Knowledge Discovery in Text and Text Mining Software”, Centre for Research in Information Management, UK Barzilay, M. Kan, and K.R. McKeown, “SIMFINDER: A Flexible Clustering Tool for Summarization,” Proc. NAACL Workshop Automatic Summarization, pp. 41-49, 2001.

[7] H. Zha, “Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering,” Proc. 25th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120, 2002.

[8] D.R. Radev, H. Jing, M. Stys, and D. Tam, “Centroid-Based Summarization of Multiple Documents,” Information Processing and Management: An Int’l J., vol. 40, pp. 919-938, 2004.

[9] R.M. Aliguyev, “A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization,”Expert Systems with Applications, vol. 36, pp. 7764- 7772, 2009.

[10] R. Kosala and H. Blockeel, “Web Mining Research: A Survey,” ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.

[11] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison- Wesley, 1989.