A SURVEY ON TEXT CLASSIFICATION TECHNIQUES AND APPLICATIONS

CT.Vidhya, S.M.Nithya, T.Vishnu Priya

Abstract— **Text classification is the way of discovering knowledge from ubiquitous text data which are easily accessible over the Internet or the Intranet. Automated text classification has been considered as a vital method to manage and process a vast amount of documents in digital forms that are widespread and continuously increasing. In general, text classification plays an important role in information extraction and summarization, text retrieval, and question answering. This paper illustrates the text classification techniques and applications.**

Index Terms— **Text classification, feature reduction, document clustering, feature selection.**

I. Introduction

Text databases consist of large collections of documents from various sources such as news articles, research papers, digital libraries, e-mail messages and Web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic forms such as electronic publications which can be viewed as a huge, interconnected, dynamic text database. Automatic text classification has always been an important application and research topic since the inception of digital documents.

Text classification is a supervised learning task for assigning text documents to pre-defined classes of documents. It is used to find valuable information from a huge collection of text documents available in knowledge databases, the World Wide Web and company-wide intranets. Text categorization is the task of automatically sorting a set of documents into categories from a predefined set. Categorization often relies on a domain for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic.

In general, text classification includes topic based text classification and text genre-based classification. Topic-based text categorization classifies documents according to their topics. Genre is defined on the way a text was created, the way it was edited and the register of language it uses. Data for genre classification are collected from the web, through newsgroups, bulletin boards, and broadcast or printed news. They are multi-source, and consequently have different formats. Thus, Genre classification is based on heterogeneous data. Figure 1 gives the Graphical Representation of the Text Classification process.

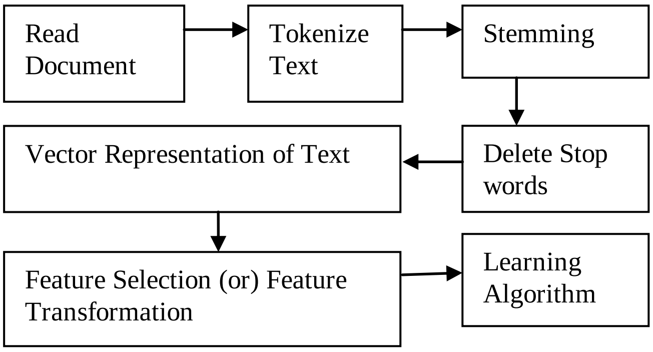
Read Document
Tokenize Text
Stemming
Delete Stop words
Vector Representation of Text
Feature Selection (or) Feature Transformation
Learning Algorithm

Fig 1. Text Classification Process

More formally, if di is a document of the entire set of documents D and {c1, c2…cn} is the set of all the categories, then text classification assigns one category cj to a document di. The task of text categorization is to train the classifier using these documents, and assign categories to new documents.

The rest of this paper is organized as follows: Section 2 presents the approaches to text classification. Section 3 describes the text classification techniques. Section 4 presents a brief description of the text mining applications. Finally, Section 5 concludes the work

II. Approaches of Text Classification

Text data set is created by processing spontaneous speech, printed text and handwritten text contains processing noise. The dataset is an unstructured dataset of documents which are pre-processed using the following three rules:

- Tokenize the file into individual tokens using space as the delimiter.
- Removing the stop word which does not convey any meaning.
- Use porter stemmer algorithm to stem the words with common root word.

There are various approaches to text mining, which can be classified from different perspectives, based on the inputs taken in the text mining system and the data mining tasks to be performed. The major approaches, based on the kinds of data they take as input, are as follows:

Manuscript received Dec, 2015*.*

CT.Vidhya*, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore, India, +91-9940815450*

S.M.Nithya*, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore, India, +91-9003458016*

T.Vishnu Priya*, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore, India, +91-8754511875,*

A. Keyword –Based Association Analysis

This approach collects sets of keywords or terms that occur frequently together as the input and then finds the association or correlation relationship among them. Association analysis first pre-processes the text data by parsing, stemming, removing stop words, and then invokes association mining algorithms. A set of frequently occurring consecutive or closely located keywords form a term or a phrase. The association mining process can help detect compound associations, domain dependant term or phrases or non-compound associations. Mining based on these associations is referred as term-level association mining. The problem of association mining in document databases is mapped to item association mining in transaction databases, where many interesting methods have been developed. This analysis can discover the relationship at a shallow level such as rediscovery of compound nouns or co-occuring patterns with less significance. It may not bring much deep understanding to the text.

B. Document Classification Analysis

Document classification has been used in automated topic tagging, topic directory construction, identification of the document and classifying the purpose of hyperlinks associated with a set of documents. A general procedure is as follows: First a set of pre-classified documents is taken as the training set. The training set is then analysed to derive the classification scheme. The so-derived classification scheme can be used for classification of other online documents. This tagging approach may rely on tags obtained by manual tagging which is costly and unfeasible for large collections of documents or by some automated categorization algorithm which may process a relatively small set of tags and require defining the categories beforehand.

C. Document Clustering Analysis

This is one of the most crucial techniques for organizing the documents in an unsupervised manner. Due to the curse of dimensionality, it first projects the documents into a lower dimensional sub space in which semantic structure of the document space becomes clear. In low-dimensional semantic space, the traditional clustering algorithms can be applied. To this end, spectral clustering, mixture model clustering, clustering using Latent Semantic indexing and clustering using Locality preserving indexing are the most well known techniques. This analysis inputs semantic information such as events, facts or entities uncovered by information extraction. This approach is more advanced and may lead to the discovery of deep knowledge but it requires semantic analysis of text by natural language understanding and machine learning methods.

III. TEXT CLASSIFICATION TECHNIQUES

A. Information Extraction (IE)

Information Extraction is the process of automatic extraction of structured information such as entities, relationship between entities and attributes describing entities from unstructured texts. Mostly useful information such as names of people, places or organization mentioned in the text is extracted without a proper understanding of the text.IE identifies useful relevant text in a document. Useful information is defined as text segment and its associated attributes.

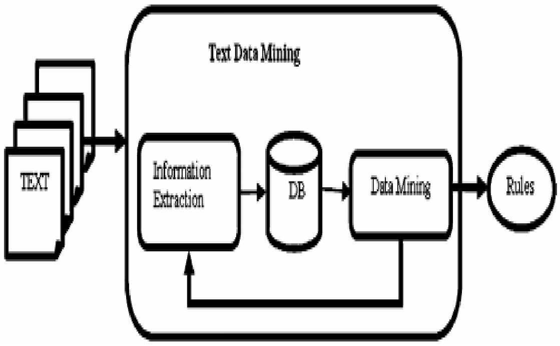

Fig 2. IE based TM framework

IE is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the extraction process. Tasks that IE systems can perform include:

- Term analysis, identifies the terms in a document, where a term may consist of one or more words. This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers.
- Named-entity recognition identifies the names in a document, such as the names of people or organisations. Some systems are also able to recognise dates and expressions of time, quantities and associated units and percentages.
- Fact extraction identifies and extracts complex facts from documents. Such facts could be relationships between entities or events.

B. Information Retrieval

Information Retrieval (IR) is finding a document of an unstructured nature usually text, that satisfies an information need within large collections of databases stored on computers. A typical information retrieval problem is to locate relevant documents in a document collection based on a user’s query. In such a search problem, user takes the initiative to “pull” the relevant information out from the collection. When the user has a long term information need ,a retrieval system may also take an initiative to “push” any newly arrived information item to a user ,if the item is judged as being relevant to the user’s information need. Such an information access process is called information filtering and the corresponding systems are called filtering systems or recommender systems.

Due to the abundance of text information, IR has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems and the web search engines. Some IR systems also use multi-word phrases (information retrieval) as index terms. Since phrases are considered more meaningful than individual words, a phrase match in the document is considered more informative than single word matches.

There are many models available for IR process which can be broadly classified as:

- Classical models of IR based on mathematical knowledge that was easily recognized and well understood simple, efficient and easy to implement. Boolean, vector and probabilistic models are the three basic classic information retrieval models.
- Non-Classical models of IR are based on principles other than similarity, probability, Boolean operations etc on which classical retrieval models are based on information logic model, situation theory model and Interaction model.

Alternative models of IR are enhancements of classical models making use of specific techniques from other fields. Example: Cluster model, fuzzy model and latent semantic indexing (LSI) models.

C. *Natural Language Processing (NLP)*

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. The goal of NLP is to accomplish human-like language processing.

An NLP system must be able to:

- Paraphrase an input text
- Translate the text into another language
- Answer questions about the contents of the text
- Draw inferences from the text

The role of NLP in text mining is to provide the systems in the information extraction phase with linguistic data that they need to perform their task. This is done by annotating documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools. Natural language processing approaches can be applied both to feature extraction and feature reduction phases of the text classification process.

D. *Text Preprocessing*

The first step in most retrieval systems is to identify the keywords for representing documents and this is referred to as tokenization. A text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. The set of different words obtained by merging all text documents of a collection is called the dictionary of a document collection. In order to reduce the size of the dictionary and thus the dimensionality of the description of documents within the collection, filtering and lemmatization or stemming methods can be applied.

E. *Feature Selection*

The aim of feature-selection methods is the reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification. This transformation procedure presents a number of advantages, including smaller dataset size, smaller computational requirements for the text categorization algorithms and shrinking of the search space. The goal is the reduction of the curse of dimensionality to yield improved classification accuracy and to reduce over fitting and therefore to increase the generalization. The feature selection phase contains three steps: (1) generating a candidate set containing a subset of the original features via certain research strategies; (2) evaluating the candidate set and estimating the utility of the features in the candidate set. Based on the evaluation, some features in the candidate set may be discarded or added to the selected feature set according to their relevance; and (3) determining whether the current set of selected features are good enough using certain stopping criterion. If it is, a feature selection algorithm will return the set of selected features, otherwise, it iterates until the stopping criterion is met.

F. *Feature Reduction*

Feature reduction operation involves a combination of three general approaches namely stop words, stemming and statistical filtering. The idea of stop word filtering is to remove words that bear little or no content information, like articles, conjunctions, prepositions. Stop word lists or the stop.

IV. Applications of Text Mining

The major Text Mining Techniques are 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.

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 word can be determined from their frequency, which is said to be more efficient and language independent. Second way of traditional feature reduction is the use of stemming to reduce frequencies of words with a common root to a single feature. Stemming methods try to build the basic forms of words. A stem is a natural group of words with equal or very similar meaning. After the stemming process, every word is represented by its stem. Statistical filtering practices are used to select those words that have higher statistical significance.

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, all applications find an answer, from market analysis to human resources management, from spelling correction to customer opinion survey.

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 details of the score cards 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

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. As most information is stored as text, text mining is believed to have a high commercial potential value. In this paper, the basic approaches of text classification are illustrated. A detailed description of the text classification techniques and the applications of text mining are described.

Traditional information retrieval methods become inadequate for increasing vast amount of data. Without knowing what could be in the documents; it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need different tools to compare the documents, rank the importance and relevance of documents, or find patterns and trends across multiple documents.

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.

Acknowledgment

Our completion of this paper could not have been accomplished without the support of staff members those who are working with us. We are very much thankful to them. For the reference, we refer many articles and research papers of many authors and institutions those are available in online and offline. We offer our sincere appreciation for the learning opportunities provided by those authors and institutions. At last but not least, our heartfelt thanks to all our family members for caring us and for the huge support.

References

[1] F. Sebastian, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[2] B.Y. Ricardo and R.N. Breathier, Modern Information Retrieval. Addison Wesley Longman, 1999.

[3] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. 14th Int'l Conf. Machine Learning, pp. 412-420, 1997.

[4] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[5] Jiawei Han and Michelin Kamber, "Data Mining concepts and techniques".

[6] Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods," Proc. ACM SIGIR, pp. 42-49, 1999.

[7] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. SIAM Review, 41:335–362, 1999.

[8] H.Almuallim & T.G.Dietterich, (1991), Learning with many irrelevant features, in Ninth National Conference on Article Intelligence", MIT Press, pp. 547{552}.

[9] R.Fano, Transmission of information. MIT Press, Cambridge, MA, 1961.

[10] T. Joachims, "Text Categorization with Support Vector Machine Learning with Many Relevant Features," Technical Report LS-8-23, Univ. of Dortmund, 1998.

[11] M. Ikonomakis, S. Kotsiantis, V. Tampakas, " Text Classification using Machine Learning Techniques" vol.4, 2005.

[12] J. Brank, M.Grobelnik, N.Milic-Frayling, D.Mladenic, "Interaction of Feature Selection Methods and Linear Classification Models", Proc. of the 19th International Conference on Machine Learning, Australia, 2002.

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

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

[15] G.Forman, an Experimental Study of Feature Selection Metrics for Text Categorization. Journal of Machine Learning Research, 3 2003, pp. 1289-1305

Authors Profile

CT.Vidhya is currently working as Assistant Professor in the Department of Computer Science and Engineering, Avinashilingam Institute of Home Science and Higher Education for Women University. From April 2012 to February 2015 She worked as Assistant Professor in the Department of Computer Science and Engineering, Adithya Institute of Technology, Coimbatore. She received her Master Degree from the Department of Computer Science and Engineering in 2012. She received her Bachelor degree from the Department of Computer Science and Engineering, Avinashilingam University, Coimbatore in 2010. Her research interest is in data mining, network security and information security.

S.M.Nithya is currently working as Assistant Professor in the Department of Computer Science and Engineering, Avinashilingam Institute of Home Science and Higher Education for Women University. From June 2011 to 2013, She worked as Assistant Professor in the Department of Information Technology, Maharaja Engineering College, Avinashi. . She received her Master Degree in the Department of Information Technology in Anna University of Technology, Coimbatore in June 2011. She received her Bachelor degree in the Department of Information Technology, Avinashilingam University, Coimbatore in May 2009. Her research interest is in data mining and Bioinformatics.

T.Vishnu Priya is currently working as Assistant Professor in the Department of Computer Science and Engineering, Avinashilingam Institute of Home Science and Higher Education for Women University, Coimbatore. From May 2014 to Feb 2015, She worked as Assistant Professor in the Department of Computer Science and Engineering, Kalasalingam University. She received her Master Degree from the Department of Information Technology in Veltech Multitech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai in 2013. She received her Bachelor degree in the Department of Information Technology, Avinashilingam University, Coimbatore in 2011. Her research interest is in Image Processing ,Data Mining and Web security.