INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND CREATIVE ENGINEERING (ISSN:2045-8711)
VOL.6 NO.4 APRIL 2016

A Survey on Data Mining and Text Categorization Technique

V.Kumthavalli
Associate Professor, Department of Computer Applications,
Sri Parasakthi College for Women, Courtallam, Tirunelveli, Tamil Nadu, India.
Email: saikumuthavalli@gmail.com
Dr.V.Vallimayil
Associate Professor & Head, Department of Computer Science & Applications,
Periyar Maniyammai University, Vallam, Thanjavur, Tamil Nadu, India.
Email: vallimayilv@gmail.com

Abstract- The data text mining gaining more recently because of the availability of the increasing number of the electronic documents from a variety of sources. In the current scenario, text classification gains lot of significance in processing and retrieval of text. The classification and knowledge discovery from these resources area for research. Automated document classification becomes a key technology to deal and organize huge volume of documents and its frees organizations from the need of manually organizing document bases. In this paper surveyed about data mining and text categorization technique.

Keywords- Text Categorization, DM,NN,RD,BP,RBF.

1. INTRODUCTION

Data Mining is in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data. These needs are automatic summarization of data, extraction of the "essence" of information stored and the discovery of patterns in raw data.

2. REVIEW OF LITERATURE

In this research work, survey the feature extraction and document classification. In order to propose this works are have analyzed various literatures are analyzed are presented in the following section.

Feature extraction is an important process for text classification. In this process, it is to determine the features which are most relevant to the classification process. This is because some of the words are much more likely to be correlated to the class distribution than others. Therefore, a wide variety of methods has been proposed in the literature in order to determine the most important features for the purpose of classification.

In order to classify the documents, it is necessary to find a way to represent documents in a way it preserves as much of the original information as possible and also is simple enough from a computational point of view. Different ways of representing documents that reflect different needs of their users has been proposed.

There have many feature extraction methods and text classifiers using machine learning techniques were already been reported by the researchers. The various literature which are used to propose the new method are as follows.

Many different techniques for removing 'less descriptive' terms has been developed in the area of information retrieval and text mining. These methods usually use some knowledge about the domain as well as some heuristics to obtain relatively good results. It also turns out the knowledge and using some "classical" pattern recognition methods in the text domain, the classification results can be improved. As far as different preprocessing techniques are concerned, the usage of nouns only did not affect the overall performance in a significant manner.

Text categorization is a popular research area, there has been a lot of previous research were undergone on it. There are two types of approaches to text categorization: Rule based and Machine learning algorithm based. In the former type of approaches, classification rules are built manually and unseen documents are classified based on these rules. In the latter type, such rules are built automatically by analyzing a sample of labeled documents statistically. Rule based approaches have high precision but very low recall because of the lack of flexibility, while machine learning based approaches show a greater balance between precision and recall. This is because

of their greater flexibility, although such approaches have lower precision than rule based approaches.

XindongWu et al. [XIND08] presented the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM): C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community With each algorithm, the provide a description of the algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm.

Ada et al. [ADA13] has presented the some data mining classification techniques such as neural network & SVMs for detection and classification of Lung Cancer in X-ray chest films. Due to high number of false positives extracted, a set of 160 features was calculated and a feature extraction technique was applied to select the best feature. They classify the digital X-ray films in two categories: normal and abnormal. The normal or negative ones are those characterizing a healthy patient. Abnormal or positive ones include types of lung cancer.

Esbrook et al. [ESBR04] proposed a multiple sampling methods combining oversampling and undersampling and validated its best result through both toy experiment and practical experiment such as text categorization. Drummond et al. [DRUM03] insisted that undersampling is more effective than oversampling in applying decision tree, C4.5, to pattern classification.

An [AN96] proposed the method of generating artificial training examples by injecting normalized random values called noise into each input vector of existing training examples and validated his approach through two toy experiments, the approximation of nonlinear function and the digit recognition.

Saito et al. [SAIT88] proposed a medical diagnosis expert system based on a multilayer Artificial Neural Network. They treated the network as a black box and used it only to observe the effects on the network output caused by change the inputs. Two methods for extracting rules from Artificial Neural Network are described by Towell et al. [TOWE93]. The first method is the subset algorithm proposed by Fu [FU91], which searches for subsets of connections to a node whose summed weight exceeds the bias of that node. The major problem with subset algorithms is that the cost of finding all subsets increases as the size of the Artificial Neural Network increases. The second method, the MofN algorithm proposed by Towell GG et al. [TOWE94], is an improvement of the subset method that is designed to explicitly search for M-of-N rules from knowledge based Artificial Neural Networks. Instead of considering an ANN connection, groups of connections are checked for their contribution to the activation of a node, which is done by clustering the Artificial Neural Network connections.

Liu et al. proposed [LIU95] X2R, a simple and fast algorithm that can be applied to both numeric and discrete data, and generate rules from datasets. It can generate perfect rules in the sense that the error rate of the rules is not worse than the inconsistency rate found in the original data. The problem of the rules generated by X2R, are order sensitive, i.e., the rules should be fired in sequence. Liu described a family of rule generators in [LIU98] that can be used to extract rules in various applications. It includes versions that can handle noise in data, produce perfect rules, and can induce order independent or dependent rules. The basic idea of the algorithm is simple: using first order information in the data to determine shortest sufficient conditions in a pattern that can differentiate the pattern from patterns belonging to other classes.

Taeho Jo [TAEH10] propose a string vector based text categorization model using neural network. Its property is shared from a linear classifier perceptron, which is an early neural network. It also develop the neural network based approach for train the classifier to categorize the document, even though for implementing a text categorization system where feature dimensionality still quite large.

Sunita Beniwal et al. [SUNI12] used various methods for classification exists like bayesian, decision trees, rule based neural networks etc. Before applying any mining technique, irrelevant attributes needs to be filtered. Filtering is done using different feature selection techniques like wrapper, filter, embedded technique. Also provide a survey of various feature selection techniques and classification techniques used for mining.

3. DATA MINING TECHNIQUES

3.1 Minable Data

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. Algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files.

3.2 Flat Files

Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

3.3 Relational Databases

A relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. The most commonly used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count. For instance, an SQL query to select the videos grouped by category would be: SELECT count(*) FROM Items WHERE type=video GROUP BY category. Data mining algorithms using relational databases can be more versatile than data mining algorithms

specifically written for flat files, since they can take advantage of the structure inherent to relational databases. While data mining can benefit from SQL for data selection, transformation and consolidation, it goes beyond what SQL could provide, such as predicting, comparing, detecting deviations, etc.

3.4 Transaction Databases

A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. For example, in the case of the video store, the rentals table such the transaction database. Each record is a rental contract with a customer identifier, a date, and the list of items rented (i.e. video tapes, games, VCR, etc.). Since relational databases do not allow nested tables (i.e. a set as attribute value), transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and one for the transaction items. One typical data mining analysis on such data is the so-called market basket analysis or association rules in which associations between items occurring together or in sequence are studied.

3.5 Multimedia Databases

Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.

3.6 Spatial Databases

Spatial databases are databases that, in addition to usual data, store geographical information like maps and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.

3.7 Time-Series Databases

Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

3.8 World Wide Web

The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, which encompasses documents available the structure of the Web, which covers the hyperlinks and the relationships between documents; and the usage of the web, describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web or web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining.

3.9 Categorization of Data Set

There are many data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and comprehensive. Classification according to the type of data source mined: this classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc. Classification according to the data model drawn on: this classification categorizes data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional, etc. Classification according to the king of knowledge discovered: this classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together. Classification according to mining techniques used: Data mining systems employ and provide different techniques. This classification categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options and offer different degrees of user interaction.

3.10 Performance Issues

Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

3.11 Data Source Issues

There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the

advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. The storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic.

3.12 Pattern Mining

Pattern mining consists of developing data mining algorithms to discover interesting, unexpected and useful patterns in databases. Pattern mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs, etc. Pattern mining algorithms can be designed to discover various types of patterns: sub graphs, associations, indirect associations, trends, periodic patterns, rules, lattices, sequential patterns, etc.

There are several definitions. Define an interesting pattern as a pattern that appears *frequently* in a database. Other to discover *rare patterns*, patterns with a high *confidence*, the top patterns, etc. There are two types of pattern mining which are follows

Frequent Pattern Mining
Sequential Pattern Mining

3.13 Frequent Pattern Mining

The most popular algorithm for pattern mining is without any doubt is Apriori Algorithm. It is designed to be applied on a transaction database to discover patterns in transactions made by customers in stores. But it can also be applied in several other applications. A transaction is defined a set of distinct items. Apriori Algorithm takes as input (1) a dataset set by the user and (2) a transaction database containing a set of transactions. Apriori Algorithm outputs all frequent item sets, groups of items shared by no less than the dataset in the input database. For example, consider the following transaction database containing four transactions. Given an example of two transactions, frequent item sets are "bread, butter", "bread milk", "bread", "milk" and "butter".

T1:bread,butter,spinach
T2:butter,salmon
T3:bread,milk,butter
T4: cereal, bread milk

Apriori algorithm can also apply a post-processing step to generate "association rules" from frequent item sets. The Apriori algorithm has given rise to multiple algorithms that address the same problem or variations of this problem such as to (1) incrementally discover frequent item sets and associations , (2) to discover frequent subgraphs from a set of graphs, (3) to discover subsequence's common to several sequences, etc.

3.14 Sequential Pattern mining

The second method is the sequential pattern mining. A sequence pattern mining is defined as a set of sequences. A sequence is a list of transactions. For example in the first part of the following figure a sequence database containing four sequences is shown. The first sequence contains item *a* and *b* followed by *c*, followed by *f*, followed by *g*, followed by *e*.

A sequential rule has the form X –> Y where X and Y are two distinct non empty sets of items. The meaning of a rule is that if the items of X appear in a sequence in any order, they will be followed by the items of Y in any order. The support of a rule is the number of sequence containing the rule divided by the total number of sequences. The confidence of a rule is the number of sequence containing the rule divided by the number of sequences containing its antecedent. The goal of sequential rule mining is to discover all sequential rules having a support and confidence no less than two thresholds given by the user named "minsup" and "minconf". For example, on the right part of the following figure some sequential rules are shown for *minsup=0.5* and *minconf=0.5*, discovered by the Rule Growth algorithm.

ID	Sequence
Sequence 1	{a,b},{c},{f},{g},{e}
Sequence 2	{a,d},{c},{b},{a,b,e,f}
Sequence 3	{a},{b},{f},{e}
Sequence 4	{b},{f,g}

ID	Rule	Support info	Confidence
Rule 1	{a,b,c}={e}	0.5	1.0
Rule 2	{a}–> {c,e,f}	0.5	0.66
Rule 3	{a,b}–>{e,f}	0.5	1.0
Rule 4	{b} –>{e,f}	0.75	0.75
Rule 5	{a}–> (e,f}	0.75	1.0
Rule 6	{c}–>{f}	0.5	1.0

Pattern mining is often viewed as techniques to explain the past by discovering patterns. However, patterns found can also be used for prediction. As an example of application, the following paper shows how sequential rules can be used for predicting the next web pages that will be visited by users on a website, with a higher accuracy than using sequential patterns .

4. CONCLUSION

In this paper, studied about research on machine learning based approaches to text categorization has been surveyed systematically. In the previous works, dimension of numerical vectors should reserve, at least, several hundred for the robustness of document classification systems. In order to mitigate sparse distribution, a task of text classification was decomposed into binary classification tasks in applying one among the traditional approaches. This requires classifiers as many as predefined categories, and each text classifier judges

whether an unseen document belongs to its corresponding category or not. This research will be a successful attempt to solve the two main problems by encoding the documents into alternative structured data to numerical vectors and then competitive self organizing neural network which received string vectors as its input data because of its advantages.

REFERENCES

[ADA13] Ada, Rajneet Kaur. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 3, March 2013 ISSN: 2277 128X.

[AN96] G. An, "The Effects of Adding Noise During Backpropagation Training on a Generalization Performance", Neural Computation, Vol 7, 1996, pp643-647.

[DRUM03] C. Drummond and R.C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling", International Conference on Machine Learning, 2003, in Post-Workshop on Learning from Imbalanced Datasets II.

[ESBR04] A. Estabrooks, T. Jo, and N. Japkowicz, "A Multiple Resampling Method for learning from Imbalances Data Sets", Computational Intelligence, Vol 28, No 1, 2004, pp18-36.

[FRAN99] E. Frank, I. H. Witten, G.W. Paynter, KEA: Practical Automatic Keyphrase Extraction, In E. A. Fox, N. Rowe (eds.): Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries. 1999, ACM Press, Berkeley, CA , 254 – 255.

[FU91] Fu L. Rule learning by searching on adapted nets. Proceedings of National Conference on Artificial Intelligence; Anaheim, CA, USA. 1991. pp. 590–595.

[HACO03] Y. HaCohen-Kerner, Automatic Extraction of Keywords from Abstracts, In V. Palade, R. J. Howlett, L. C. Jain (eds.): KES 2003. Lecture Notes in Artificial Intelligence, 2003, Vol. 2773,Springer-Verlag, Berlin Heidelberg, 843 – 849.

[LIU95] Liu H, Tan ST. X2R: A fast rule generator. Proceedings of IEEE International Conference on Systems, Man and Cybernetics; Vancouver, BC, Canada. 22–25 October 1995; pp. 1631–1635.

[MLAD92] Mladenic, D., Grobelink, M. (1999). Feature Selection for unbalanced class distribution and Naïve Bayes. In: the Proceedings of International Conference on Machine Learning, p. 256-267.

[SUNI12] Sunita Beniwal, Jitender Arora. Classification and Feature Selection Techniques in Data Mining.International Journal of Engineering Research & Technology (IJERT). Vol. 1 Issue 6, August - 2012. ISSN:2278-0181.

[TAEH01] Taeho Jo and Jerry Seo, "Text Categorization Oriented Connectionist Model", The Proceedings of International Conference on Computer Processing of Oriental Languages, 2001, pp65-68.

[TOWE93] Towell GG, Shavlik JW. Extracting refined rules from knowledge-based neural networks. Mach. Learn. 1993;13:71–101.

[TOWE94] Towell GG, Shavlik JW. Knowledge - based artificial neural networks. Artif. Intell. 1994;70:119–165.

[XIND08] Xindong Wu, Vipin Kumar, J.Ross Quinlan, Joydeep Ghosh, Qiang Yang ,: Top 10 algorithms in data mining, Knowl Inf Syst (2008) 14:1–37.

[YANG99] Yang, Y. (1999). "An evaluation of statistical approaches to text categorization, Information Retrieval", 1 (1-2) 67-88.

[CHEN96] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[FAYY96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[FRAW91] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, Knowledge Discovery in Databases: An Overview. In G. Piatetsky-Shapiro et al. (eds.), Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

[HAN00] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[IMIE96] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.

[PIATE96] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.