Review of Text Mining Techniques

Priya Bhardwaj*
Priyanka Khosla**

Abstract

Data mining is a process of discovering potential and practical, previously unknown patterns from large pre existing databases. Text mining is a realm of data mining in which large amount of structured and unstructured text data is analyzed to produce information of high commercial value. Analyzing textual data requires context analysis. This paper represents the current research status of text mining. Association rules, a novel technique in text mining is gaining increasing currency among research scholars is discussed. Based on studied attempts, the potential future research activities have been proposed.

Keywords: component; formatting; style; styling; insert (key words)

I. Introduction

With the evolution of internet and rapid developments in information technology enormous amount of textual data is generated in the form of blogs, tweets and discussion forums. The data potentially has a lot of hidden information which can intuitively predict human behavior. The major challenge is to uncover relationships and associations in the data which is in various formats i.e. unstructured data [1]. Text mining aims at revealing the concealed information by using various techniques that are capable of coping up with large amount of structured data on one hand and handling the vagueness, fuzziness and uncertainty of the unstructured data on the other. Text mining or knowledge discovery from text (KDT) — for the first time mentioned in Feldman et al. [2] — deals with the computational analysis of textual data. It is an interdisciplinary field involving techniques from information extraction, information retrieval as well as Natural Language Processing (NLP) and integrates them with the algorithms and methods of data mining, statistics and machine learning.

The most convenient way of storing information is believed to be text. In the recent surveys it is considered that 80% of company's information is contained in text [4] and analysis of this information is required for making strategic decisions.

This paper introduces the current research status of text mining. Section III describes some general models used for mining text. The applications of text mining and the related techniques are discussed in Section IV followed by a conclusion.

II. State of the Art

Hans Peter Luhn[6] in 1958, published an article in journal of IBM which discusses about the automatic extraction by data processing machine and classifies the document on the word frequency statistics. This was considered to be one of the primitive definitions of business intelligence.

The research in the field of text mining continued and many scholars carried prolific research in the field. In the 1st International Conference on Data Mining and Knowledge Discovery in 1995 Feldman et al. [5] proposed Knowledge Discovery in Database (KDT). Supervised [7] and Unsupervised [8][9] learning algorithms are used to uncover hidden patterns in the textual documents.

Subsequently, other outstanding work done is in the field including dimensionality reduction on the basis of correlation in feature extraction [13]-[14]; soft set approach using association rule mining [15] by introducing SOFTAPRIORI that discovers relationships more accurately; sentiment analysis for online forums hotspot detection and forecast [16];

Priya Bhardwaj*
Assistant Professor
Institute of Information Technology and Management, Delhi, India
Priyanka Khosla**
Assistant Professor
Institute of Information Technology and Management, Delhi, India

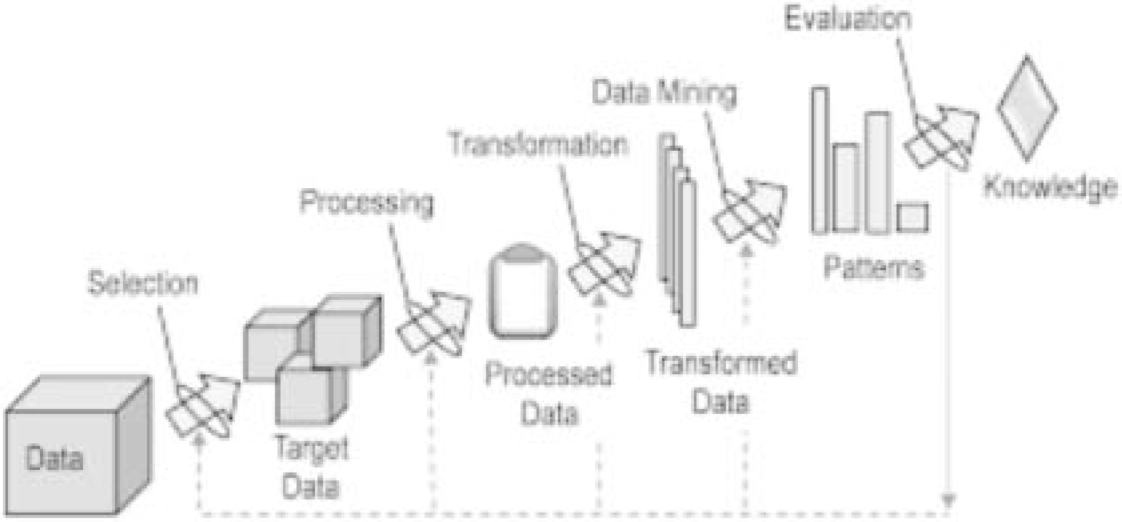

Figure 1: Knowledge Discovery Process

sentiment analysis using self organizing maps and ant clustering [17]; and text mining in various other fields such as stock prediction [18], web mining [19], digital library [20] and so on.

III. Text mining Models

Generally text mining is a four step process which is text preprocessing, data selection, data mining and post processing..

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Data Cleaning

The textual data available for mining is generally collected over web from the tweets, discussion forums and blogs. The data set available from these sources is in various formats i.e. "unstructured". We need to "clean" the data by performing parsing of data, missing value treatment, removing inconsistencies. After performing the desired operations the data set should be consistent with the system.

B. Data selection and transformation

The textual data available for mining is generally collected over web from the tweets, discussion forums and blogs. The data set available from these sources is in various formats i.e. "unstructured". We need to "clean" the data by performing parsing of data, missing value treatment, removing inconsistencies. After performing the desired operations the data set should be consistent with the system.

C. Data Mining

After the document being converted into the intermediate form data mining techniques can be applied to different type of data according (structured, semi- structured and unstructured) to recognize relationships and patterns. The various data mining techniques are discussed in detail in section IV.

D. Data Post processing

It includes the tasks of evaluation and visualization of the knowledge coming out after performing text mining operations.

IV. Techniques Used in Data Mining

The progress of Information Technology has produced large amount of data and data repositories in diverse areas. The research made in databases has further given rise to the techniques used to store and process the data for decision making. Thus, Data mining is a process of finding useful patterns from large amount of data and is also termed as knowledge discovery process which states the knowledge mining or extraction from large amount of data.

Machine Learning Algorithms

- Unsupervised Machine Learning :It is a type of machine learning algorithm that is used to draw conclusion from datasets that consists of input data without the labeled responses. The most familiar unsupervised learning method is cluster analysis, that is used for exploratory data analysis to find hidden patterns or grouping in data.

- Supervised Machine Learning Algorithm: It is a type of machine learning algorithm that uses a identified dataset (called the training dataset) in order to make predictions. The training data set comprises of input data and response values. From this dataset, the supervised learning algorithm searches for a model that can make predictions of the response values for a new dataset. A test dataset is often used to validate the model. Using larger training datasets often yield models with higher predictive power that can generalize well for new datasets.

A. Classification Technique:

Classification is the commonly used data mining technique that employs training dataset or pre-classified data to generate a model that is used to classify records according to rules. This technique of data mining is used to find out in which group each data instance is related within a given dataset using the training dataset. It is used for classifying data into different classes according to some constraints. Credit Risk analysis and fraud detection are the application of this technique. This algorithm employs decision tree or neural network-based classification algorithms. Classification is a Supervised learning that involves the following steps:

Step 1: Rules are extracted using the learning algorithm from (create a model of) the training data. The training data are pre classified examples (class label is known for each example).

Step 2: Evaluation of the rules on test data. Usually split known data into training sample (2/3) and test sample (1/3).

Step 3: Apply the generated rules on new data.

Thus, the classifier-training algorithm uses the pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called as a classifier. Rules are generated from it that further helps in making decisions.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

B. Clustering RulesTechnique:

It is the task of grouping objects in such a way that objects in the same group or cluster are similar in one sense or another to each other than to those objects present in another groups. Thus it is an identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Types of clustering methods involves

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

C. Association Rules Technique:

Association is a data mining technique that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. These rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be “If a customer buys a dozen eggs, he is 80% likely to also purchase milk.” Therefore both eggs and milk together are associated with each other and are likely to be placed together to increase the sales of both the product. Thus association rules helps industries and businesses to make certain decisions, such as cross marketing, customer shopping, designing of catalogue etc. Association Rule algorithms should be able to generate rules with confidence values less than one. Although the number of possible

Table I. Tasks With Algorithms

Examples of tasks	**Algorithms to use**
Predicting a discrete attribute Flag the customers in a prospective buyers list as good or poor prospects. Calculate the probability that a server will fail within the next 6 months. Categorize patient outcomes and explore related factors.	**Decision Tree Algorithm** **Clustering Algorithm** **Neural Network Algorithm**
Predicting a continuous attribute Forecast next year's sales. Predict site visitors given past historical and seasonal trends. Generate a risk score given demographics.	**Decision Tree Algorithm**
Predicting a sequence: Perform click stream analysis of a company's Web site. Analyze the factors leading to server failure. Capture and analyze sequences of activities during outpatient visits, to formulate best practices around common activities.	**Clustering Algorithm**
Finding groups of common items in transactions: Use market basket analysis to determine product placement. Suggest additional products to a customer for purchase. Analyze survey data from visitors to an event, to find which activities or booths were correlated, to plan future activities.	**Association Algorithm** **Decision Tree Algorithm**
Finding groups of similar items: Create patient risk profiles groups based on attributes such as demographics and behaviors. Analyze users by browsing and buying patterns. Identify servers that have similar usage characteristics.	**Clustering Algorithm**

Association Rules for a given dataset is generally very large and among that a high proportion of the rules are usually of little value. Types of association rules are:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

V. Choosing an Algorithm by Task

To help you select an algorithm for use with a specific task, the following table provides suggestions for the types of tasks for which each algorithm is traditionally used.

VI. Conclusion

The paper has provided a concise introduction about the state of the art of text mining. In the next section the steps required to extract valuable information from the data set are described. Consequent section summarized various data mining techniques such as classification, clustering and association rule. Text mining gives a direction to the upcoming fields like artificial intelligence, therefore it needs the continuous improvement in order to grow its application areas.

References

1. Ah Hwee Tan et al., “Text Mining: The state of the art and the challenges”, *Proceedings of the Pakdd Workshop on Knowledge Disocovery from Advanced Databases*, pp. 65-70, 2000.
2. R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117, 1995.
3. Marti A. Hearst, Untangling text data mining, pp. 3-10, 1999, University of Maryland.
4. S.Grimes. “Unstructured data and 80 percent rule.” Carabridge Bridgepoints, 2008
5. H. P. Luhn, “A Business Intelligence System”, *Ibm Journal of Research & Development*, vol. 2, no. 4, pp. 314-319, 1958.
6. M. E. Maron, J. L. Kuhns, “On Relevance Probabilistic Indexing and Information Rctrieval”, *Journal of the Acm*, vol. 7, no. 3, pp. 216-244, 1960.
7. Larsen, Bjornar, and Chinatsu Aone. “Fast and effective text mining using linear-time document clustering.” Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999.
8. Jiang, Chuntao, et al. “Text classification using graph mining-based feature extraction.” Knowledge-Based Systems 23.4 (2010): 302-308.
9. Liu, Wei, and Wilson Wong. “Web service clustering using text mining techniques.” International Journal of Agent-Oriented Software Engineering 3.1 (2009): 6-26.
10. Ronen Feldman, I. Dagan, H. Hirsh, “Mining Text Using Keyword Distributions”, Journal of Incelligent Information Systems, vol. 10, no. 3, pp. 281-300, 1998.
11. J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret, “Information mining: use of the document dimensions to analyse interactively a document set”, European Colloquium on IR Research: ECIR, pp. 66-77, 2001.
12. M. Ghanem, A. Chortaras, Y. Guo, A. Rowe, J. Ratcliffe, “A Grid Infrastructure For Mixed Bioinformatics Data And Text Mining”, Computer Systems and Applications 2005. The 3rd ACS/IEEE International Conference, vol. 29, pp. 41-1, 2005.
13. Haralampos Karanikas, C. Tjortjis, B. Theodoulidis, “An Approach to Text Mining using Information Extraction”, Proc. Workshop Knowledge Management Theory Applications (KMTA 00, 2000.
14. Qinghua Hu et al., “A novel weighting formula and feature selection for text classification based on rough set theory”, *Natural Language Processing and Knowledge Engineering 2003. Proceedings. 2003 International Conference on IEEE*, pp. 638-645, 2003.
15. Nahm, Un Yong, and Raymond J. Mooney. “Mining soft-matching association rules.” Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.
16. Li, Nan, and Desheng Dash Wu. “Using text mining and sentiment analysis for online forums hotspot detection and forecast.” Decision support systems 48.2 (2010): 354-368.
17. Chifu, Emil ªt, Tiberiu ªt Leþia, and Viorica R. Chifu. “Unsupervised aspect level sentiment analysis using Ant Clustering and Self-organizing Maps.” Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on. IEEE, 2015.
18. Nikfarjam, Azadeh, Ehsan Emadzadeh, and Saravanan Muthaiyah. “Text mining approaches for stock market prediction.” Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. Vol. 4. IEEE, 2010.
19. Kosala, Raymond, and Hendrik Blockeel. “Web mining research: A survey.” ACM Sigkdd Explorations Newsletter 2.1 (2000): 1-15.
20. Fuhr, Norbert, et al. “Digital libraries: A generic classification and evaluation scheme.” International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 2001.