A Review on Text Mining

Yu Zhang, Mengdong Chen and Lianzhong Liu
School of Computer Science and Engineering
BeihangUniversity
Beijing, China
{yzhang & cmd & lz_liu}@buaa.edu.cn

Abstract—**Because of large amounts of unstructured text data generated on the Internet, text mining is believed to have high commercial value. Text mining is the process of extracting previously unknown, understandable, potential and practical patterns or knowledge from the collection of text data. This paper introduces the research status of text mining. Then several general models are described to know text mining in the overall perspective. At last we classify text mining work as text categorization, text clustering, association rule extraction and trend analysis according to applications.**

Keywords-text mining; data mining; knowledge discovery

I. INTRODUCTION

With the rapid development of information technology and the extensive application of network, the Internet has gradually become an indispensable part of people's life. Web pages and social network sites will generate large amounts of unstructured text data, such as blogs, forum posts, technical documentation, etc. These data showing people's behavior and thought intuitively, contains a lot of information, which is extremely difficult to deal with because of the huge number and various forms. But the demand of analyzing text data is rising. Therefore, how to acquire the information people need from large numbers of unstructured text data becomes the research hotspot in the field of data mining and information. Text mining came into being.

Text mining [1], also known as knowledge discovery in textual database(KDT) [2] or text data mining [3], of which new interesting knowledge is created, is defined as the process of extracting previously unknown, understandable, potential and practical patterns or knowledge from the collection of massive and unstructured text data or corpus.

As a branch of data mining, text mining is believed to have higher commercial value than data mining because 80% of a company's information is contained in text documents [4]. However, text mining is more complex as the unstructured text data. Text mining is a comprehensive research area, which involves in the fields of artificial intelligence, machine learning, mathematical statistics, database system, and so on.

This article introduces the history of text mining and research status. Then some general models are described in Section III. The fourth part is to classify text mining work according to application. Finally, it is summary.

978-1-4799-8353-7 /15/$31.00 ©2015 IEEE

II. HISTORY AND RESEARCH STATUS

In 1958, Hans Peter Luhn [5] published an article on the IBM Journal, which describes a business intelligence system that can realize the document automatic extraction and coding by using data processing machine, and implement document classification through the word frequency statistics. It is recognized as the earliest definition of Business Intelligence (BI), also as the prototype of text mining.

Subsequently, many scholars have carried out fruitful research work in this field. List some of the typical representatives here. In 1960, Maron published a paper reporting a novel technique for literature indexing and searching in a mechanized library system [6], which is the first paper about automatic classification. KDT is first proposed by Feldman Ronen et al. [2] at the 1st International Conference on Knowledge Discovery and Data Mining in 1995. Bjornar Larsen and Chinatsu Aone [7] describe an unsupervised, near-linear time text clustering system, which is fast and effective, offering a number of algorithm choices for each phase. They used F-Measure (a combination of precision and recall) to gauge the quality of the generated hierarchies.

There are also many other outstanding research work, including text representation [8] and models construction [1][9]-[12]; data dimensions reduction research in feature extraction [13]-[14]; research on mining algorithm of text classification [15]-[17] and clustering [18]-[20]; deep semantic mining based on natural language process [21]-[22]; and text mining applications in different fields, such as literature mining in molecular biology [23]-[24], stock prediction in the field of finance and securities [25], web mining on the internet [26]-[28], digital library [29] and so on.

Currently, text mining has entered into the practical stage from the experimental stage. There are many successful text mining systems, like IBM Intelligent Miner for Text [30], Text Miner [31], VisualText [32] etc.

III. TEXT MINING MODELS

Text mining is generally composed of three steps: text preprocessing, text mining operations, postprocessing. Text preprocessing tasks including data selection, classification and feature extraction generally convert the documents into intermediate forms, which should be suitable for different mining purpose. Text mining operations are the central part of a text mining system, and include clustering, association rule

discovery, trend analysis, pattern discovery and other knowledge discovery algorithms. Postprocessing tasks manipulate data or knowledge coming from text mining operations, such as evaluation and selection of knowledge, interpretation and visualization of knowledge.

So far, there are a lot of common text mining models. The earliest one is the KDT system (Knowledge Discovery in Text) [10] proposed by Feldman et al. The general architecture of the KDT system is shown in figure 1. The system takes two inputs: a collection of keyword-labeled documents, and a keyword hierarchy which is a directed acyclic graph (DAG) of terms, where each of the terms is identified by a unique name. In general, the keyword hierarchy given the hierarchical relationship between the main concepts involved in the application domain, is part of the domain background knowledge. Discovery operation module is used to mine the collection of keyword-labeled documents with the background knowledge, and obtain the pattern people need, which is shown in a friendly way like graphics in presentation module.

Tan [1] put forward a general framework consisting of two components, which is a representative text mining model. The two components are: text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form (cf. fig. 2). Intermediate form (IF) can be semi-structured such as the conceptual graph representation, or structured such as the relational data representation. Intermediate form can be document-based wherein each entity represents a document, or concept-based wherein each entity represents an object or concept of interests in a specific domain. Mining a document-based IF deduces patterns and relationship across documents. Document clustering/visualization and categorization are examples of mining from a document-based IF. Knowledge distillation from a concept-based IF derives pattern and relationship across objects or concepts. Text mining operations, such as predictive modeling and associative discovery, fall into this category. A document-based IF can be transformed into a concept-based IF by realigning or extracting the relevant information according to the objects of interests in a specific domain. It follows that document-based IF is usually domain-independent and concept-based IF is domain-dependent.

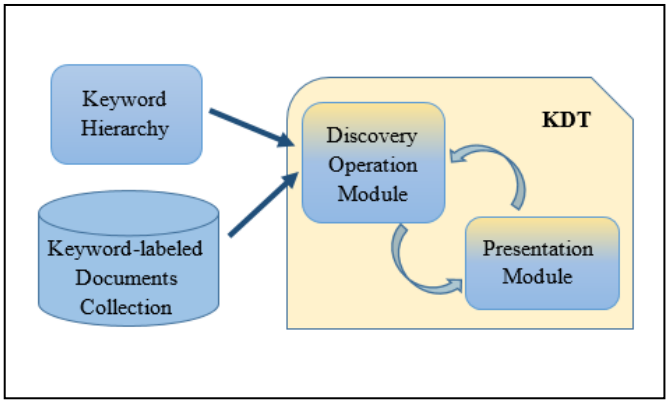

Figure 1. KDT system architecture

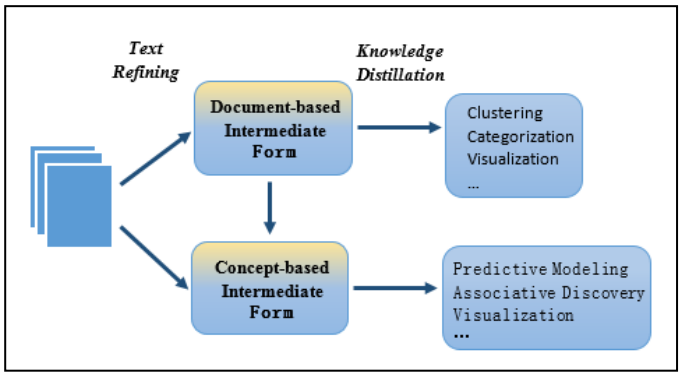

Figure 2. A text mining framework

Mothe et al. [11] add document warehouse on the basis of the general model. Figure 3 provides an overview of the steps needed to create a document warehouse. The information selection has in charge to gather the information related to the domain of interest, which is described through an information need. The information reformatting is to modify the format of the information, because documents are not necessarily structured. The information cleaning is used to decide what the information that has to be kept is. That includes the dimensions to take into account and eventually the values to be considered as well as solving some syntax and semantic problems (e.g. synonyms). The information summarization corresponds mainly to aggregation functions done on numerical data according to the value of some of the attributes. Classification, clustering, factorial analysis (principal component analysis, correspondence analysis) are easily performed on document warehouse.

In addition to general text mining models, many models have been proposed for specific application field. For instance, Shehata et al. propose a novel concept-based mining model [33] for text clustering. The model, whose input is a raw text document, consists of concept-based term analysis and concept-based similarity measure. They extend the model in [34]. The advanced concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept analysis, and concept-based similarity measure, as depicted in Fig. 4. The model can efficiently find significant matching concepts between documents, according to the semantics of their sentences.

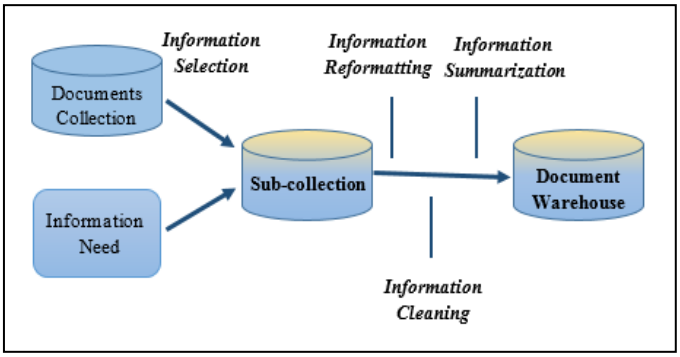

Figure 3. Overview of the document warehouse creation

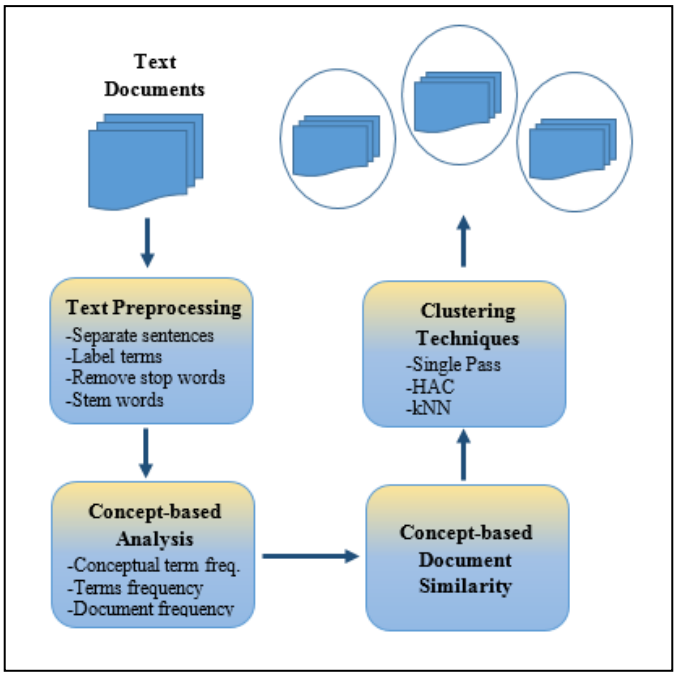

Figure 4. Concept-based mining model system

IV. THE CLASSIFICATION OF TEXT MINING

Although text mining is an emerging field, there already has been a great deal of research work involving a lot of application area. According to these application areas, text mining can be classified as text categorization, text clustering, association rule extraction and trend analysis.

A. Text Categorization

As an important text mining application, text categorization is a supervised learning process. Text categorization (TC) is the process of automatically determining the text category according to the text content under the given classification system. The study of TC dates back to 1960s, mainly for the index of scientific literature. To the 1990s, because of the increasing number of text data and the strong demand for processing text, TC fully developed. In recent years, TC has applied to many fields, from automatic or semi-automatic text indexing to spam filtering, metadata generation, word sense disambiguation, hierarchical categorization of web pages, genre detection, etc.

Word sense disambiguation may be seen as a TC task [35]-[36] once we view word occurrence contexts as documents and word senses as categories. Gale et al. [35] propose a method to disambiguate senses that are usually associated with different topics. They use the class of Bayesian decision models that has been applied successfully in related tasks such as author identification and information retrieval.

Sentiment analysis is also an important research area of text categorization. [37] calculates the orientation similarity of words in the phrase based on the sentiment weight priority and puts forward the concept of center word to calculate the orientation of the phrase according to the combination of the words in the phrase. Lillian Lee et al. [38] use standard bag-of-features framework and employ three machine learning methods (Naive Bayes, maximum entropy classification, and support vector machines) to classify sentiment, using movie reviews as data.

Other applications are e-mail filters to discard "junk" mail [39], language identification for texts of unknown language [40], the organization of patents into categories for making their search easier [41], and so on.

B. Text Clustering

Text clustering is one of the earliest and most mature fields in text mining. It is an unsupervised process through which objects are classified into groups without predefined categories. Text clustering is based on the well-known cluster hypothesis: relevant documents tend to be more similar to each other than to nonrelevant ones. Clustering is useful in a wide range of data analysis fields, including data mining, document retrieval, image segmentation, and pattern classification. And it is mostly used to improve the precision rate and recall rate of information retrieval system [42]-[44].

Text clustering can find topic in large scale text data, and is the powerful tool for text theme analysis. [45] gives a topic analysis method. First extract the name entities from documents, then look for frequent itemsets: groups of named entities that commonly occurred together, next perform clustering of the named entities grouped by the frequent itemsets using a hypergraph-based method [46]. Each cluster is represented as a set of named entities and corresponds to an ongoing topic in the corpus.

Topic tracking of dynamic text data is also an interesting subject of text clustering. Montes-y-Gómez et al. [47] proposes a text mining method to get topics from online news and analyze the influence of the peak news topics over other current news topics. A common phenomenon in news reports is the influence of a peak news topic, i.e., a topic with one-time short-term peak of frequency, on the other news topics. This kind of influence is called an ephemeral association. They propose a technique with which the observable associations are detected by simple statistical methods.

C. Association Rule Extraction

Association rule extraction proposed by Agrawal [48], is an important topic in text mining. It is to find out the association relationship between different feature words from the text collection. The formal description of association rule extraction is given in [49], as below:

Given a collection of indexed documents $D=\{d_1,d_2,...,d_n\}$ and a set of items $A = \{w_1, w_2, \ldots, w_n\}$, which composed of keywords, term, phrase, or concept. Let W_i be a set of items. A document d_i is said to contain W_i if and only if $W_i \subseteq d_i$. An association rule is an implication of the form $W_i \Rightarrow W_j$ where $W_i \subset A$, $W_j \subset A$ and $W_i \cap W_j = \Phi$. There are two important basic measures for association rules, support and confidence. The rule $W_i \Rightarrow W_j$ has support s in the collection

of documents D if s % of documents in D contain $W_i \cup W_j$. The support is calculated by the following formula:

$$Support\left(W_iW_j\right) = \frac{Support\ count\ of\ W_iW_j}{Total\ number\ of\ documents\ D} \quad (1)$$

The rule $W_i \Rightarrow W_j$ holds in the collection of documents D with confidence c if among those documents that contain W_i, c % of them contain W_j also. The confidence is calculated by the following formula:

$$Confidence\left(W_i \diagdown W_j\right) = \frac{Support\left(W_iW_j\right)}{Support\left(W_i\right)} \quad (2)$$

An association rule extraction is broken into two steps: 1) generate all the itemsets whose support is greater than the user specified minimum support (called minsupp). Such sets are called the frequent itemsets and 2) use the identified frequent itemsets to generate the rules that satisfy a user specified minimum confidence (called minconf).

Text mining system often produces a large number of association rules, but few of them are the knowledge users interested in. So evaluate and choose the association rules are very important to a practical text mining system. Reference [50] estimates the novelty of text-mined rules using semantic distance measures based on WordNet [51]. If the semantic distance is short, then the rule may be useless.

D. Trend Analysis

If considering the time dimension of text data, it can reflect the changing rules of text topics or predict the development trend of objects [52]. Now the research on trend analysis mainly aims at Current news, financial reports, scientific literature, business reports and other scheduling text data [53].

[54] proposes general probabilistic approaches to discover and summarize the evolutionary patterns of themes in a text stream through discovering latent themes from text, constructing an evolution graph of themes, and analyzing life cycles of themes. To discover the evolutionary theme graph, their method would first generate word clusters (i.e., themes) for each time period and then use the Kullback-Leibler divergence measure to discover coherent themes over time. The evolution graph can reveal how themes change over time and how one theme in one time period has influenced other themes in later periods. They also propose a method based on hidden Markov models for analyzing the life cycle of each theme. This method would first discover the globally interesting themes and then compute the strength of a theme in each time period. This allows us to not only see the trends of strength variations of themes, but also compare the relative strengths of different themes over time.

Brian Lent et al. [55] find the changes of all kinds of patents over the years through analyzing the relevant patent database. Victor Lavrenko et al. [56] predict the trend of stock prices based on news of relevant quoted companies and historical data of stock prices. Montes-y-Gómez et al. [57] present a method for the trend analysis of news to find the current social hot spot and its changing tendency.

At present, work in this area mainly adopts methods based on statistics [10][54]. Feldman et al. [10] use keyword distributions to label documents, and calculate the distance between keyword distributions for collections from different points in time to find the changing trend of the text topics.

V. CONCLUSION

We have provided a very brief introductions to the text mining and its research status. Then several general models are described to know text mining in the overall perspective. At last we classify text mining work as text categorization, text clustering, association rule extraction and trend analysis according to applications. Text mining is a new direction of artificial intelligence, and with the continuous improvement of the text mining technology, its application areas will be growing.

REFERENCES

[1] Tan, Ah Hwee, et al. "Text Mining: The state of the art and the challenges." Proceedings of the Pakdd Workshop on Knowledge Disocovery from Advanced Databases(2000):65--70.

[2] Feldman, Ronen, and I. Dagan. "Knowledge Discovery in Textual Databases (KDT)." In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95(1995):112--117.

[3] Hearst, Marti A. "Untangling text data mining." University of Maryland1999:3--10.

[4] S. Grimes. "Unstructured data and the 80 percent rule." Carabridge Bridgepoints, 2008.

[5] Luhn, H. P. "A Business Intelligence System." Ibm Journal of Research & Development2.4(1958):314-319.

[6] Maron, M. E., and J. L. Kuhns. "On Relevance, Probabilistic Indexing and Information Retrieval.." Journal of the Acm7.3(1960):216-244.

[7] Larsen, Bjornar, and C. Aone. "Fast and effective text mining using linear-time document clustering." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data miningACM, 1999:16-22.

[8] Salton, G., A. Wong, and C. S. Yang. "A vector space model for automatic indexing." In Communications of the ACM 18(11), 1975: 613-620.

[9] Steinheiser, R., and C. Clifton. "Data Mining on Text." 2012 IEEE 36th Annual Computer Software and Applications ConferenceIEEE Computer Society, 1998:630.

[10] Feldman, Ronen, I. Dagan, and H. Hirsh. "Mining Text Using Keyword Distributions." Journal of Intelligent Information Systems10.3(1998):281-300.

[11] Mothe J., Chrisment C., Dkaki T., Dousset B., Egret D., (2001) "Information mining: use of the document dimensions to analyse interactively a document set", European Colloquium on IR Research: ECIR, 66-77.

[12] Ghanem, M., Chortaras, A., Guo, Y., Rowe, A., & Ratcliffe, J. (2005). "A Grid Infrastructure For Mixed Bioinformatics Data And Text Mining." Computer Systems and Applications, 2005. The 3rd ACS/IEEE International Conference on(Vol.29, pp.41-I).

[13] Karanikas, Haralampos, C. Tjortjis, and B. Theodoulidis. "An Approach to Text Mining using Information Extraction." Proc. Workshop Knowledge Management Theory Applications (KMTA 00(2000).

[14] Hu, Qinghua, 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, 2003:638-645.

[15] Tan, Songbo, et al. "Using dragpushing to refine centroid text classifiers." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.

[16] Masoud Makrehchi and Mohamed S. Kamel. "Text Classification Using Small Number of Features.." Lecture Notes in Computer Science(2005):580-589.

[17] Jiang, Chuntao, et al. "Text Classification using Graph Mining-based Feature Extraction." Knowledge-Based Systems23.4(2010):302–308.

[18] Hotho, Andreas, Steffen Staab, and Gerd Stumme. "Ontologies improve text document clustering." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.

[19] Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[20] Luo, Congnan, Y. Li, and S. M. Chung. "Text document clustering based on neighbors.." Data & Knowledge Engineering68.11(2009):1271-1288.

[21] Berendt, Bettina, Andreas Hotho, and Gerd Stumme. "Towards semantic web mining." The Semantic Web—ISWC 2002. Springer Berlin Heidelberg, 2002. 264-278.

[22] Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." Proceedings of the 12th international conference on World Wide Web. ACM, 2003.

[23] de Bruijn, Lambertus, and Joel Martin. "Literature mining in molecular biology." Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications. 2002.

[24] Tanabe, L., et al. "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling." Biotechniques 27.6 (1999): 1210-4.

[25] Mittermayer, Marc-André. "Forecasting intraday stock price trends with text mining techniques." System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. IEEE, 2004.

[26] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web." Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on. IEEE, 1997.

[27] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2.1 (2000): 1-15.

[28] Grace, L. K., V. Maheswari, and Dhinaharan Nagamalai. "Analysis of web logs and web user in web mining." arXiv preprint arXiv:1101.5668 (2011).

[29] Witten, Ian H., et al. "Text mining in a digital library." International Journal on Digital Libraries 4.1 (2004): 56-59.

[30] http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an&subtype=ca&htmlfid=897/ENUS298-061&appname=isource&language=enus#3pb

[31] http://www.sas.com/en_us/software/analytics/text-miner.html

[32] http://www.textanalysis.com/Products/VisualText/visualtext.html

[33] Shehata, S., F. Karray, and M. Kamel. "Enhancing Text Clustering Using Concept-based Mining Model." IEEE 13th International Conference on Data MiningIEEE Computer Society, 2006:1043-1048.

[34] Shehata, Shady, F. Karray, and M. S. Kamel. "An Efficient Concept-Based Mining Model for Enhancing Text Clustering." IEEE Transactions on Knowledge & Data Engineering22.10(2010):1360-1371.

[35] Gale, W. A., Church, K. W., and Yarowsky, D. (1993). “A Method for Disambiguating Word Senses in a Large Corpus.” Computers and the Humanities 26(5): 415–439.

[36] Escudero, Gerard, L. Marquez, and G. Rigau. "Boosting Applied to Word Sense Disambiguation." IN PROCEEDINGS OF THE 12TH EUROPEAN CONFERENCE ON MACHINE LEARNING2000:129--141.

[37] Dun LI, Fu-Yuan CAO, Yuan-Da CAO, Yue-Liang WAN. "Text Sentiment Classification Based on Phrase Patterns." Computer Science. 35.4(2008):132-134. DOI:10.3969/j.issn.1002-137X.2008.04.037.

[38] Pang, Bo, L. Lee, and S. Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques." Proceedings of Emnlp(2002):79--86.

[39] Androutsopoulos, Ion, et al. "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages." Proceedings of Annual International Acm Sigir Conference on Research & Development in Information Retrieval(2000):160--167.

[40] Cavnar, William B., and J. M. Trenkle. "N-Gram-Based Text Categorization." Proceedings of Int’l Symposium on Document Analysis & Information Retrieval Las Vegas Nv(2001):161--175.

[41] Larkey, Leah S. "A Patent Search and Classification System." Digital Libraries the Fourth Acm Conference on Digital Libraries(1999):79--87.

[42] Rijsbergen, Van. C.J. "Information Retrieval." 14th International Symposium on Methodologies for Intelligent Systems. Volume 2871., Maebashi City, Japan, LNCS, Springer-Verlag12.2-3(1989):95.

[43] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey "Scatter/Gather: a cluster-based approach to browsing large document collections." Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrievalACM, 1992:318--329.

[44] Oren Zamir, Oren Etzioni, Omid Madani, Richard M. Karp. "Fast and Intuitive clustering of Web documents." Proceedings of International Conference on Knowledge Discovery & Data Mining(1997):287--290.

[45] Clifton, Chris, and R. Cooley. TopCat: Data Mining for Topic Identification in a Text Corpus. Principles of Data Mining and Knowledge DiscoverySpringer Berlin Heidelberg, 1999:174-183.

[46] E.-H. S. Han, G. Karypis, and V. Kumar, “Clustering Based on Association Rule Hypergraphs,” Proc. SIGMOD’97 Workshop Research Issues in Data Mining and Knowledge Discovery, 1997.

[47] M. Montes-y-Gómez, A. Gelbukh, and A. López-López. "Discovering Ephemeral Associations among News Topics." 17th international joint conference on artificial intelligence ijcai-01, workshop on adaptive text mining 2001.

[48] Agrawal, Rakesh, T. Imieliński, and A. Swami. "Mining Association Rules Between Sets Of Items In Large Databases." SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data1993:207--216.

[49] Mahgoub, Hany, et al. "A Text Mining Technique Using Association Rules Extraction." International Journal of Computational Intelligence1(2008):21.

[50] Basu, Sugato, et al. "Using lexical knowledge to evaluate the novelty of rules mined from text." Proceedings of the NAACL workshop and other Lexical Resources: Applications, Extensions and Customizations. 2001.

[51] Fellbaum, C., and G. Miller. WordNet:An Electronic Lexical Database. MIT Press, 1998.

[52] Zhi-Qun Chen, and Guo-Xuan Zhang. "A Survey of Text Mining." Journal of Pattern recognition and artificial intelligence 18.1(2005):65-74. DOI:10.3969/j.issn.1003-6059.2005.01.012.

[53] Zhi-Qun Chen. " A Survey of Trend Mining for texts." Journal of Information Science 2(2010):316-320.

[54] Mei, Qiaozhu, and C. X. Zhai. "Discovering evolutionary theme patterns from text: an exploration of temporal text mining." Proceedings of Kdd ’(2005):198-207.

[55] Lent, Brian, Rakesh Agrawal, and Ramakrishnan Srikant. "Discovering Trends in Text Databases." KDD. Vol. 97. 1997.

[56] Lavrenko, Victor, et al. "Mining of Concurrent Text and Time Series." Proceedings of Acm Sigkdd Intl Conference on Knowledge Discovery & Data Mining Workshop on Text Mining(2000):37--44.

[57] Montes-y-Gómez, López –López and Gelbukh, "Text Mining as a Social Thermometer", Proc. Of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99, Stockholm, 1999.