# A Variety of Text Mining Technology and Tools Research

Mingyang Jiang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China
e-mail: jiang_ming_yang@163.com

Xiaojing Fan
College of Mechanican Engineering
Inner Mongolia University for the Nationalities
Tongliao, China

Xinhong Zhang
Department of Neurology
The Affiliated Hospital of Inner Mongolia University for Nationalities
Tongliao, China

Lian Jie
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Yuxin Zhou
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

QiangHu Wang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

ZhiFeng Zhang
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China

Zhili Pei*
College of Computer Science and Technology
Inner Mongolia University for the Nationalities
Tongliao, China
e-mail: zhilipei@sina.com

**Abstract—Text mining is a process with rich semantics of the text were analyzed to understand the content and meaning it contains. Its in-depth research is bound to greatly improve people's ability to extract information from vast amounts of textual data, there is a high commercial value. This paper introduces the case of text data mining, and then gives a framework for text mining, text mining to extract text mining technology and related technical information used in the assessment method and so made a detailed introduction. Many text mining techniques based on data mining technology-based, can be seen from the name, text mining and data mining on the purpose is the same, are trying to extract knowledge from large amounts of information in data mining is from the original data extract, and text mining is extracted from the text material. Open source tools usually do as a business tool as data on a variety of formats provide good support, but there will be a certain format restrictions, or even require their own proprietary data formats. Text mining tools is for the four typical open-source tools, including data format features three modules and user experience.**

***Keywords- text mining; text mining tools; information extraction; information retrieval; data mining***

## I. INTRODUCTION

Text mining or document mining is a process of interest to the user or to obtain useful patterns from unstructured text information. Text mining covers a variety of technologies, including information extraction, information retrieval, natural language processing and data mining techniques[1][2]. Its main purpose is to extract knowledge from the unknown without the use of the original text. However, text mining is a very difficult task, because it must deal with those already vague and unstructured text data, so it is a multidisciplinary field of mixed, covering IT, text analysis, pattern recognition, statistical , data visualization, database technology, machine learning and data mining techniques.

Data mining tools are multi-language, multi-format most commercial text provides a good support, and data pre-processing functions are more perfect, supports structured and semi-structured data analysis process is completely unstructured text mining tools are generally open-source have their own inherent format requirements, foreign source text mining tool for Chinese support for the poor, and most are still stuck in the open source tool only supports structured and semi-structured data phase[3-6]. Text mining is currently still in the exploratory stage of development, including the development of commercial text mining tools to be faster than the open-source text mining tools. However, everything has its two sides, most commercial software because of its quality and scarcity and very expensive, not suitable for small businesses and research institutions excellent open source text mining tool is able to meet the relevant requirements to the maximum extent, and also to support for loading user's own expansion algorithm, or directly embedded into the user's own program were to go.

## II. TEXT MINING TECHNOLOGY FRAMEWORK

Data mining technology itself is a new development in the field of current data technology, text mining, and development history is shorter. Traditional information retrieval techniques for dealing with massive data is not satisfactory, they increasingly important text mining, text mining technology is visible from information extraction and related technical field slowly evolved from. Text mining and information retrieval with information extraction closely and fully consider using components to complete these tasks[4-9]. The best should be a text mining system in accordance with the process must be executed sequentially, some similar to the data mining process, which also describes a process for the extraction of knowledge, except that the information retrieval and information extraction into a pre-procedure.

Process of text mining technologies is shown in Fig .1.

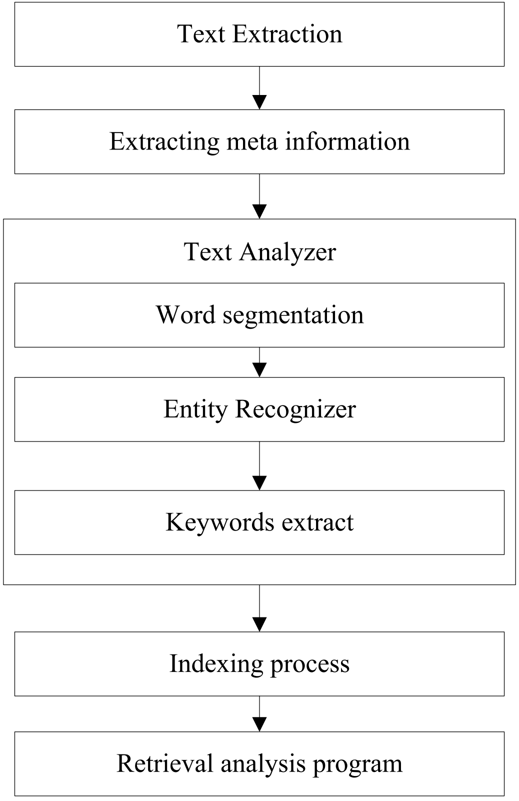

Figure 1. Process of Text Mining Technologies

1. Information Retrieval: Find and retrieve all those who are considered likely related to the current working text. In general, the system user can define text set, but still need a filtration system for the relevant text.

2. Information extraction: extracting information from text selection after the. This extraction process is generally filled with a user-defined process to get the desired information model.

3. Data Mining: Once filled with entries for each text, entered the standard database mining stage, you can expect discover some useful knowledge model.

4. Explanation: The interpretation placed on the mining phase derived from the pattern of course, the best interpreter can understand natural language format.

## III. MAIN TECHNOLOGY OF TEXT MINING

Many text mining techniques based on data mining technology-based, can be seen from the name, text mining and data mining on the purpose is the same, are trying to extract knowledge from large amounts of information in data mining is from the original data extract, and text mining is extracted from the text material[10][11]. If the concept of data generalization, it can be seen as a kind of text mining data mining data mining, but tend to be very precise and structured, most of the studies considered only from extracting knowledge database. For this reason, many techniques can not be applied to text mining areas freely, many data mining researchers also did not take into account the text input, the discussion in the text mining methods can be seen.

### A. *Affairs and Rules*

A transaction is the concept of data mining is used to temporarily assign values to the data item, it describes a series of feature vectors record and a temporary storage location used to describe the index include when it is applied to the text, they can use to record and representatives of the frequency of occurrence of each word and its location, however, is not limited to a particular feature of a word, it can be a phrase, or a punctuation mark. affairs and deal with their technology, and can not understand all the text and just trying to find this pattern: the simultaneous occurrence of words, phrases will be in a fixed list or use grammar rules in a particular text.

### B. *Conceptual Level*

Conceptual level can be described as the relationship between the concept of a directed graph, a father-son relationship with respect to his son, said his father is a more general concept hierarchy can have appropriate definition is entirely user-based, so this requires specific domain knowledge , which is a dedicated technical staff say there is a certain difficulty. conceptual level is also used to mark words and phrases related concepts will mark down, clearly mark all the ancestors of the concept of this particular concept .Feldman no fixed one way that many methods are available. when generating a collection of many of the concepts, they can begin to apply data mining technology[12].

### C. *Neural Networks*

Neural network is very suitable for use in noisy, and has a data structure and changing the properties of hard to understand, which is a common phenomenon in a text message. A self-organizing map is an unsupervised artificial neural networks.

## IV. Information Extraction and Related Technologies

Information extraction is among the most important text mining module, in fact, many articles have these two as the same concept, of course, in fact, they are not equal. Purpose is to scan information extraction from text and extracting facts needed. In information extraction, the dictionary to discover the relationship between the facts and their order information extraction, there are many different techniques to obtain dictionary particular field. Reference[13] provides an overview of information extraction technology, lists three basic stages of information extraction.

### A. Fact Extraction

At this stage, the concern is how to find independent facts in the text, so the domain knowledge is very important. According to the text in fact possible pattern recognition system can be built. The fact that the main extraction techniques have pattern matching, lexical analysis, syntactic and semantic structures.

### B. Facts Integration

The main problem to be solved by the fact that the integration is to prevent mutual explanation. Each one needs to look at the facts independently, and then see if they are mixed together will constitute the expression of meaning. Solve the problem of a sentence repeated explanation is relatively low stage. Relatively high level in terms of it should be integrated, you can use the concept of fusion events. But such treatment is actually for coders is very difficult, will involve many fuzzy recognition problems.

### C. Knowledge Representation

This is the third phase of information extraction, information extraction technology but also an important part. Author’s article only talks about how to fill the template stored in the database so that it can be a problem.

## V. Comparison of Open Source Text Mining Tools

Four representative source text mining tools for detailed analysis in the data format features three modules and user experience. Weka comprehensive algorithm which has been favored by many data mining staff, LingPipe is specifically developed for natural language processing toolkit, LIBSVM is SVM pattern recognition and regression toolkit, ROSTCM major colleges and universities in the face of very wide application of Chinese the support is best.

### A. Data Format

Open source tools usually do as a business tool as data on a variety of formats provide good support, but there will be a certain format restrictions, or even require their own proprietary data formats. When selecting tools, you should first consider whether the data meets or after conversion tool can meet the requirements, while, if the results of the analytical tools but also for subsequent processing, it should also take into account the output format previously used tools are common or can NO is converted to a common format, to support the work of the late. Weka input formats include ARFF, CSVXRFF and C4.5, output formats including ARFF, CSV, stored in the database via JDBC. LingPipeinput formats include XML, HTML and Text, output formats including XML.

Four open source tool has its own fixed format requirements, the need for data collection to make formatting. Although Weka support for common CSV format, but the effect of making the document more time ARFF format for later analysis, generally using its own tools will convert ARFF.Weka CSV txt format does not support document, requires the user to use another tool or write your own code format conversion. LIBSVM data output format requires a special tool to open the view, difficult to integrate other applications into the data output format three other open source tools easier expansion[14][15].

### B. Function Module

Function module is the most important tool when developing, but not the most versatile is the best, because often lead to full plain, but not deep enough, not enough analysis is the use of professional staff unwilling to see. Should be based on the actual situation, targeted to select the most appropriate tools to complete the analysis, which can achieve a multiplier effect. Therefore, the function module tool meets their requirements, usually directly about the user's selection will. Of four open-source text mining tools from text preprocessing steps, text categorization, and a variety of common algorithms return, text clustering and association rules, and can access the database, model evaluation and secondary development interface and other aspects of a more detailed comparison.

Text is text mining preprocessing crucial step process, which directly affects the late job classification, clustering, association rules, such as the effect of the more conventional operations text word, to stop words, word frequency analysis, text feature extraction is also text preprocessing the core content.

Text classification is preprocessed data, the process of selecting classifier training, evaluation and feedback of results. In this paper, a common classification algorithm TF-IDF classification, NaiveBayes classification, Knn classification, decision tree classification, neural network classifiers and support vector classification. Classification does not exist merits of each set of data has its appropriate classifier, so when training classification model, you need to try different classifiers and different parameters in order to achieve the optimization model.

Text clustering based division includes clustering, based on hierarchical clustering, density-based clustering, grid-based and model-based clustering. Clustering based division includes K-means, X-means, K-medoid and ISODATA, where X-means is to improve the K-means algorithm. Based on hierarchical clustering include BirchClusterer, CureClusterer, SingleLinkClusterer, CompleteLinkClusterer and AverageLinkClusterer. Density-based clustering include DBScan and Optics grid-based clustering include Sting Clusterer and CliqueClusterer, Cobweb belong clustering model. Regression analysis was used to determine between two or more variables quantitative relationship of interdependence, the use of a wide range of general regression analysis will be incorporated into the text classification category, but this article in order to more clearly compare different tools each regression algorithm, so alone out analysis and comparison. addition to the above features, the scope of application tools and scalability supports secondary development interface to access the database, and to assess the classification clustering training model and other factors in the selection of tools also must be considered.

### *C. Tool Evaluation*

In order to better understand the operation of text mining tools and their differences in the realization of the same algorithm, combined with the actual situation, select the Weka, LingPipe experimental evaluation. Taking into account the text mining achieve several major functional modules classification algorithm and clustering algorithm is the most mature, the paper chosen a more classic and two tools are contained Naive Bayes classifier in the testing process. Both tools text classification comparative results are shown in the table below.

TABLE I. TEXT CLASSIFICATION COMPARATIVE RESULTS

| Evaluation Index | Weka | LingPipe |
|---|---|---|
| Correct rate | 77.56 | 82.63 |
| Error rate | 22.44 | 17.37 |
| Accuracy | 78 | 83 |

## VI. SUMMARY

Text mining is the biggest motivation lurking in electronic form from the large amount of text data processing using data mining technology companies large amounts of text data, will bring tremendous business value addition for people interested in the reasons for Text Mining lies: people sometimes do not know what they're looking for in the end, and mining can extract a lot of useful information from the database needs to be done in the future is: how will the existing data mining and text mining technology applications in the field is good integration, so this article will be able to tap more effectively.In the existing data mining technology into the field of text mining, while also taking into account the special nature of text mining techniques, such as handling semantic relations, greater demand for unstructured text resources such as time and space. Can develop new text mining algorithms for these features to improve the results of satisfaction. Open source tools for data input format requires different application needs its own conversion program requires the user to have some foundation, so it should further support a variety of open source tools commonly used data formats.

## ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61163034, 61373067), the Grassland Excellent Talents Project of Inner Mongolia Autonomous Region(2013), the supported By Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region(NJYT-14-A09), the Inner Mongolia Natural Science Foundation(2013MS0911), the 321 Talents Project the two level of Inner Mongolia Autonomous Region(2010), the Inner Mongolia talent development fund(2011), the First Trade Leading Talent of Tongliao, and the Science Research Project of Inner Mongolia University(NMD1229, NMD1322). The Corresponding author is Zhili Pei, and the Corresponding email is zhilipei@sina.com.

## REFERENCES

[1] Helena Ahonen, Oskari Heinonen, Mika Klemettinen, Inkeri Verkamo A. Mining in the Phrasal Frontier[C]. In: Proceedings of PKDD 97-1st European Symposium on Principles of Data Mining and Knowledge Discovery. Norway: Trondheim, 1997.

[2] Oren Etzioni. The World-Wide Web:Quagmire or Gold Mine[J]. Communications of the ACM, 1996, 39(11): 65-68.

[3] Key Technology Research [D] Shozhong Tang. Text mining. Beijing Forestry University 2013.

[4] Wich Yvonne, Warschat Joachim, Spath, Dieter; Ardilio Antonino, Konig-Urban Kamilla, Uhlmann Eckart. Using a text mining tool for patent analyses: Development of a new method for the repairing of gas turbines[C]. 2013 Portland International Conference on Management of Engineering and Technology, PICMET 2013. 2013: 1010-1016.

[5] Iio Jun. Basic techniques in text mining using open-source tools. Proceedings of the 9th International Symposium on Open Collaboration, WikiSym + OpenSym. [C] 2013: 175-182

[6] Da Costa Pinho Isis, Epstein Daniel, Reategui Eliseo Berni, Corrêa Ygor. The use of text mining to build a pedagogical agent capable of mediating synchronous online discussions in the context of foreign language learning. 43rd IEEE Annual Frontiers in Education Conference, FIE 2013[C]. 2013: 393-399

[7] Samhaa R. El-Beltagy,Ahmed Rafea. KP-Miner: A keyphrase extraction system for English and Arabic documents[J]. Information Systems . 2008 (1)

[8] Juanzi Li,Qi'na Fan,Kuo Zhang. Keyword extraction based on tf/idf for Chinese news document[J]. Wuhan University Journal of Natural Sciences . 2007 (5)

[9] C.D. Manning,P. Raghavan,P.,H. Schutze.Introduction to Information Retrieval. Journal of Women s Health . 2008

[10] Computer and telecommunications Xiaoyan Zhang, Ying Hua. Text mining methods and applied research [J]. 2011(12):68-69

[11] Salton G,Wong A,Yang CS.A vector space model for automatic indexing. Communications of the ACM . 1975

[12] Fang Li.Text mining of certain key technologies [D]. Beijing University of Chemical Technology .2010.

[13] Ralph Grishman. Information Extraction: Techniques And challenges[J]. Information Extraction: A Multi disciplinary Approach to an Emergine Information Technology, 1997, 1299: 10-27.

[14] Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B. Rule extraction from support vector machines: an overview of issues and application in credit scoring. Studies in Computational Intelligence, 2008, 80: 33-63.

[15] Carter P H. The Rocchio classifier and second generation wavelets. In: Proceedings of the International Society for Optical Engineering. Orlando, USA: SPIE, 2007: 1-11.