Techniques, Applications and Challenging Issue in Text Mining

Shaidah Jusoh [1] and Hejab M. Alfawareh[2]

[1,2] College of Computer Science & Information Systems, Najran University
P.O Box 1988, Najran, Saudi Arabia

Abstract

Text mining is a very exciting research area as it tries to discover knowledge from unstructured texts. These texts can be found on a computer desktop, intranets and the internet. The aim of this paper is to give an overview of text mining in the contexts of its techniques, application domains and the most challenging issue. The focus is given on fundamentals methods of text mining which include natural language possessing and information extraction. This paper also gives a short review on domains which have employed text mining. The challenging issue in text mining which is caused by the complexity in a natural language is also addressed in this paper.

Keywords: *-; text mining, information extraction, natural language processing, ambiguity.*

1. Introduction

In this modern culture, text is the most common vehicle for the formal exchange of information. Although extracting useful information from texts is not an easy task, it is a need of this modern life to have a business intelligent tool which is able to extract useful information as quick as possible and at a low cost. Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool. The tool is a text mining system which has the capability to analyze large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information [1]. The aim of text mining tools is to be able to answer sophisticated questions and perform text searches with an element of intelligence.

Technically, text mining is the use of automated methods for exploiting the enormous amount of knowledge available in text documents. Text Mining represents a step forward from text retrieval. It is a relatively new and vibrant research area which is changing the emphasis in text-based information technologies from the level of retrieval to the level of analysis and exploration. Text mining, sometimes alternately referred to as text data mining, refers generally to the process of deriving high quality information from text. Researchers like [2], [3] and others pointed that text mining is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT). According to [4] the boundaries between data mining and text mining are fuzzy. The difference between regular data mining and text mining is that in text mining, the patterns are extracted from natural language texts rather than from structured databases of facts.

Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics, and computational linguistics. Preprocessing of document collection (text categorization, information extraction, term extraction), storing the intermediate representations, analysing the intermediate representation using a selected technique such as distribution analysis, clustering, trend analysis, and association rules, and visualizing the results are considered necessary processes in designing and implementing a text mining tool. Among the features of text mining systems/tools are:

- a user centric process which leverages analysis technologies and computing power to access valuable information within unstructured text data sources ;
- text mining processes are driven by natural language processing and linguistic based algorithm
- eliminate the need to manually read unstructured data sources.

Research in text mining has been carried out since the mid-80s when the US academic, Prof Don Swanson, realized that, by combining information slice from seemingly unrelated medical articles, it was possible to deduce new hypotheses [5]. In the early years of text mining research, text mining systems were aimed at information specialists. They typically require a combination of domain and informatics expertise to configure. Today, work on text mining has been carried out by researchers for different various type of domains. The aim of this paper is to give an overview of text mining system. The paper is organized as follows. Section 2 presents fundamental techniques in text mining. Section 3 reviews text mining work which has been conducted for a specific domain. Section 4 addressed the challenging issue in developing a robust text mining. Section presents a summary of the paper.

2. Techniques

Researchers in the text mining community have been trying to apply many techniques or methods such as rule-based, knowledge based, statistical and machine-learning-based approaches. However, the fundamental methods for text mining are natural language processing (NLP) and information extraction (IE) techniques. The former technique focuses on text processing while the latter focuses on extracting information from actual texts. Once extracted, the information can then be stored in databases to be queried, data mined, summarized in a natural language and so on. The use of natural language processing techniques enables text mining tools to get closer to the semantics of a text source [6]. This is important, especially when the text mining tool is expected to discover knowledge from texts.

2.1 Natural Language Processing (NLP)

NLP is a technology that concerns with natural language generation (NLG) and natural language understanding (NLU). NLG uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. Most NLG systems include a syntactic reliazer to ensure that grammatical rules such as subject-verb agreement are obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently. The most well known NLG application is machine translation system. The system analyzes texts from a source language into grammatical or conceptual representations and then generates corresponding texts in the target language. NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic. NLU consists of at least of one the following components; tokenization, morphological or lexical analysis, syntactic analysis and semantic analysis. In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol such an exclamation mark. Morphological or lexical analysis is a process where each word is tagged with its part of speech. The complexity arises in this process when it is possible to tag a word with more than one part of speech. Syntactic analysis is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence. It determines, for instance, how a sentence is broken down into phrases, how the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words used [7].

Semantic analysis is a process of translating a syntactic structure of a sentence into a semantic representation that is precise and unambiguous representation of the meaning expressed by the sentence. A semantic representation allows a system to perform an appropriate task in its application domain. The semantic representation is in a formally specified language. The language has expressions for real world objects, events, concepts, their properties and relationships, and so on. Semantic interpretation can be conducted in two steps: *context independent interpretation* and *context interpretation*. Context independent interpretation concerns what words mean and how these meanings combine in sentences to form sentence meanings. Context interpretation concerns how the context affects the interpretation of the sentence. The context of the sentence includes the situation in which the sentence is used, the immediately preceding sentences, and so on.

2.2 Information Extraction (IE)

IE involves directly with text mining process by extracting useful information from the texts. IE deals with the extraction of specified entities, events and relationships from unrestricted text sources. IE can be described as the creation of a structured representation of selected information drawn from texts. In IE natural language texts are mapped to be predefine, structured representation, or templates, which, when it is filled, represent an extract of key information from the original text [8], [9]. The goal is to find specific data or information in natural language texts. Therefore the IE task is defined by its input and its extraction target. The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists.

Using IE approach, events, facts and entities are extracted and stored into a structured database. Then data mining techniques can be applied to the data for discovering new knowledge. Unlike information retrieval (IR), which concerns how to identify relevant documents from a document collection, IE produces structured data ready for post-processing, which is crucial to many text mining applications. Figure 1 illustrates how IE can play a part in a knowledge mining process. Furthermore, IE allows for mining the actual information present within the text, rather than the limited set of tags associated to the documents. The work of [9], [10], have presented how information extraction is used for text mining. According to [11] and [12] typical IE are developed using the following three steps:-

- text pre-processing; whose level ranges from text segmentation into sentences and sentences into tokens, and from tokens into full syntactic analysis;

- rule selection; the extraction rules are associated with triggers (e.g. keywords), the text is scanned to identify the triggering items and the corresponding rules are selected;

- rule application, which checks the conditions of the selected rules and fill in the form according to the conclusions of the matching rules.

Furthermore [13] and [14] emphasized that information extraction is based on understanding of the structure and meaning of the natural language in which documents are written, and the goal of information extraction is to accumulate semantic information from text. Technically, extracting information from texts requires two pieces of knowledge: lexical knowledge and linguistic grammars. Using the knowledge we are able to describe the syntax and semantic of the text[15]. A common approach to information extraction is to use patterns which match against text and identify items of interest. Patterns are applied to texts which have undergone various levels of linguistic analysis, such as phrase chunking [16] and full syntactic parsing [17]. The approaches may use different definition of what constitutes a valid pattern. For example, [18] use subject-verb-object tuples derived from a dependency parse, followed by [19] uses patterns which match certain grammatical categories, mainly nouns and verbs, in phrase chunked text. Reference [20] reported in identifying the parts of a person name through analysis of name structure. For example, the name Doctor Paul R. Smith is composed of a person title, a first name, a middle name, and a surname. It is presented as a preprocessing step for an entity recognition and for the resolution of co-references to help determine, for instance, that John F. Kennedy and President Kennedy are the same person, while John F. Kennedy and Caroline Kennedy are two distinct persons.

Research work in [21] applied IE for detecting events in text. Event detection consists of detecting temporal entities in conjunction with other entities. For example, conferences are usually made up of four parts: one conference name, one location, and two dates (e.g., name: “AAAI,” location: “Boston,” start date: “July 16th 2006,” end date: “July 20th 2006”). A person birth or death is a person name and date pair (e.g., name: “John Lennon,” date: “December 8th, 1980”). Smith used event detection to draw maps where war locations and dates are identified.

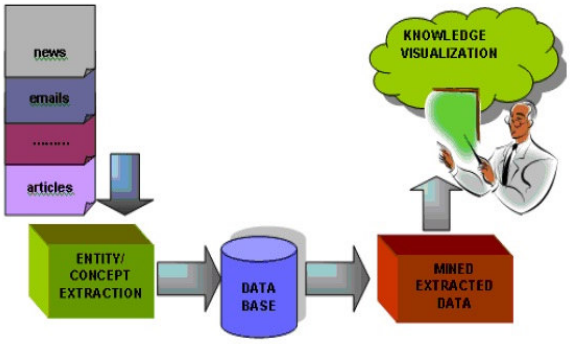

Fig. 1 A diagram shows important entities are extracted and stored in a database. Data mining approach is used to mined the stored data. Hidden knowledge is then visualized.

3. Applications

Text mining application uses unstructured textual information and examines it in attempt to discover structure and implicit meanings hidden within the text [9]. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with. Through text mining, we can uncover hidden patterns, relationships, and trends in text. [14] argued that the benefits of using text mining is to get to decision points more quickly, at least 10x speedup over previous methods, and find information which is hidden. Reference [23] addressed that text mining enables organizations to explore interesting patterns, models, directions, trends, rules, contained in text in much the same way that data mining explores tabular or “structured” data.

3.1 Bioinformatics

Research work for IE has grown dramatically in a bioinformatics domain, where biomedical journal articles have become an important application area in the recent years. The motivation for this work comes primarily from biologists, who find themselves faced with an enormous increase in the number of publications in their field since the advent of modern genomics is too many; keeping up with the relevant literature is nearly impossible for many scientists [22]. In the bioinformatics domain, biomedical research literature has been a target for text mining. The first textbook on biomedical text mining with a strong genomics focus appeared in 2005 [14], where it has reported that industry has suggested that 90% of drug targets are derived from the literature. The goal of text mining in this area is to allow biomedical researchers to extract knowledge from the biomedical literature in facilitating new discovery in a more efficient manner [4], [24].

Most of the text mining research in this domain has been done in the context of MEDLINE. MEDLINE records consist of a title, an abstract, a set of manually assigned metadata terms. Various text mining approaches such as co-occurance based mining (Blake and Pratt, 2001), IR-based meta-data profiling [25] have been proposed for MEDLINE data.

In evaluating biomedical text mining, Hersh, [24] claimed that, most research in text mining still focuses on the development of specific functions or algorithms. Although some text mining systems have been developed, such as MedScan [26] and Textpresso [2], Hersh argued that none is really routine use by end-users. At the same time, most of the text-mining tools in biomedical domain have focused on test collections developed by individual research groups.

3.2 Business Intelligence

Of the major concerns in any business is to minimize the amount of guessing work involved in decision making. The risk of making wrong prediction should be reduced. Most of the data mining techniques are created to deal with prediction. The problem with data mining is that it can help only up to a certain point, since most of data are available in texts (reports, memos, emails, planning document, etc). Data mining and text mining techniques can complement each other. For example, data mining techniques may be used to reveal the occurrence of a particular event while text mining techniques may be used to look for an explanation of an event.

3.3 National Security

The use of text mining tool in national defence security domain has become an important issue. Government agencies are investing considerable resources in the surveillance of all kinds of communication, such as email, chats in chat rooms. Email is used in many legitimate activities such as messages and documents exchange. Unfortunately, it can also be misused, for example in the distribution of unsolicited junk mail, mailing offensive or threatening materials. Since time is critical and given the scale of the problem, it is infeasible to monitor emails or chat rooms normally. Thus automatic text mining tools offer a considerable promise in this area. Although not much work has been conducted in this area (compared to bioinformatics), text mining technology is becoming an emergence technology for national security defence. The work of [28] particularly focuses on investigating and determining the gender of the author based on the gender-preferential language used by the author. They claimed that men and women use language and converse differently even though they speak the same language. The work has been conducted by using the Support Vector Machine (SVM) developed by T. Joachims from the University of Dortmund. The work of [29] for example, has applied text mining techniques to existing medical literature to identify viruses which can be potentially be used as biological weapon, and where such capability is not yet recognized. Another example of text mining system is COPLINK system [30]. It was done at University of Arizona in Tucson to help the police to discover the link between agencies.

4. Challenging Issue

The major challenging issue in text mining arise from the complexity of a natural language itself. The natural language is not free from the ambiguity problem. Ambiguity means the capability of being understood in two or more possible senses or ways. Ambiguity gives a natural language its flexibility and usability, and consequently, therefore it cannot be entirely eliminated from the natural language. One word may have multiple meanings. One phrase or sentence can be interpreted in various ways, thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated for a specific domain. On the other hand, most of the IE systems that involve semantic analysis exploit the simplest part of the whole spectrum of domain and task knowledge, that is to say, named entities. According to [31], IE does a more limited task than full text understanding. He pointed that in full text understanding, all the information in the text is presented, whereas in information extraction, the semantic range of the output, the relations will be presented are delimited. However, the growing need for IE application to domains such as functional genomics requires more text understanding. Named entity recognition (NER) describes an identification of entities in free text. For example, in biomedical domain, entities would be gene, protein names and drugs. NER often forms the starting point in a text mining system, meaning that when the correct entities are recognized, the search for patterns and relations between entities can begin. Reference [4] also claimed that one of the major problems in NER is ambiguous protein names; one protein name may refer to multiple gene products.

The work of [32] have demonstrated an effort to resolve ambiguous terms using sense-tagged corpora and unified medical language system (UMLS) with the motivation that the UMLS has been used in natural language processing applications such as information retrieval and information extraction systems. In their work, machine-learning techniques have been applied to sense-tagged corpora, in which senses (or concepts) of ambiguous terms have been most manually annotated. Sense disambiguation classifiers are then derived to determine senses (or concepts) of those ambiguous terms automatically. However, they conclude that manual annotation of a corpus is an expensive task.

Consequently [33] have extended the previous work by mining is biological named entity tagging (BNET) that identifies names mentioned in text and normalizes them with entries in biological databases. They concluded that that names for genes/proteins are highly ambiguous and there are usually multiple names for the same gene or protein.

The previous work has shown that recognizing and classifying named entities in texts require knowledge on the domain entities. List entities are used to tag text entities, with the relevant semantic information; however exact character strings are often not reliable enough for precise entity identification [15]. Research work in [34] demonstrated on using possibility theory and context knowledge in resolving an ambiguous entity. The obtained results show that the approach was successful; however, the context of the texts should be defined by a user. As claimed by [4], the ambiguity is still the major "world problem" in text mining applications.

5. Summary

This paper has presented an overview techniques, applications and challenging issue in text mining. The focus has been given on fundamental methods for conducting text mining. The methods include natural language processing and information extraction. A brief review on application domains has been presented. The purpose of this section is to give an overview to a reader on how text mining systems can be used in real life. The paper also addressed the most challenging issue in developing text mining systems.

References

[1] F. Sebastiani, "Machine learning," *ACM Computing Surveys*, vol. 1, no. 34, pp. 1–47, 2002.

[2] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (kdt)," in *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 1995, pp. 112–117.

[3] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 3–10.

[4] R. Malik, "Conan: Text mining in biomedical domain," PhD thesis, Utrecht University, Austria, 2006.

[5] J. Nightingal, "Digging for data that can change our world," *the Guardian*, Jan 2006.

[6] S. Jusoh and H. M. Alfawareh, "Agent-based knowledge mining architecture," in *Proceedings of the 2009 International Conference on Computer Engineering and Applications*, IACSIT. Manila, Phillipphines: World Academic Union, June 2009, pp. 602–606.

[7] S.Jusoh and H.M. Alfawareh, "Natural language interface for online sales," in *Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007)*. Malaysia: IEEE, November 2007, pp. 224–228.

[8] R. Rao, "From unstructured data to actionable intelligence," in *Proceedings of the IEEE Computer Society*, 2003.

[9] H. Karanikas, C. Tjortjis, and B. Theodoulidis, "An approach to text mining using information extraction," in *Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference*, 2000.

[10] U. Nahm and R. Mooney, "Text mining with information extraction," in *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.

[11] R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson, "FASTUS: A cascaded finite-state transducer for extraction information from natural language text," in *Finite States Devices for Natural Language Processing*, E. Roche and Y. Schabes, Eds., 1997, pp. 383–406.

[12] J. Cowie and Y. Wilks, *Information extraction*, New York, 2000.

[13] N. Singh, "The use of syntactic structure in relationship extraction," Master's thesis, MIT, 2004.

[14] R. Hale, "Text mining: Getting more value from literature resources," *Drug Discovery Today*, vol. 10, no. 6, pp. 377–379, 2005.

[15] C. Nedellec and A. Nazarenko, "Ontologies and information extraction: A necessary symbiosis," in *Ontology Learning from Text: Methods, Evaluation and Applications*, P. Buitelaar, P. Comiano, and B. Magnin, Eds. IOS Press Publication, 2005.

[16] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Machine Learning*, vol. 34, pp. 233–272, 1999.

[17] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks, "Description of the lasie system as used for muc-6," in *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 1996, pp. 207– 220.

[18] R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen, "Automatic acquisition of domain knowledge for information extraction," in *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000*, 2000, pp. 940–946.

[19] R. Jones, R. Ghani, T. Mitchell, and E. Riloff, "Active learning for information extraction with multiple view feature sets," in *Proceedings of the 20th International*

Conference on Machine Learning (ICML 2003), August 21-24 2003.

[20] U. Charniak, "Unsupervised learning of name structure from coreference data," in *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001, pp. 48–54.

[21] D. Smith, "Detecting and browsing events in unstructured text," in Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, 2002.

[22] D. Jurafsky and J. H. Martin, Speech and Language Processing: An introduction to Natural language Processing, Computational Linguistics and Speech Recognition. United States of America: Prentice Hall, 2009

[23] S. Chakrabarti, Mining the Web: Analysis of Hypertext and Semi Structured Data. San Francisco, CA: Morgan Kaufman, 2000.

[24] W. Hersh, "Evaluation of biomedical text-mining systems: Lessons learned from information retreival," *Briefings in Bioinformatics*, vol. 6, no. 4, pp. 344–356, 2005.

[25] P. Srinivasan, "Meshmap: A text mining tool for medline," in *Proceedings the American Medical Informatics Annual Symposium*, 2001, pp. 642–646.

[26] T. Sekimizu, H. Park, and J. Tsuji, "Identifying the interactions between genes and gene products based on frequently seen verbs in medline abstract," S. Miyano and T. Takagi, Eds. Tokyo Japan: Universal Academy Press, 1998.

[27] H. M. Muller, E. Kenny, and Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature," *P.W PLoS Biol*, vol. 2, no. 11, Sep 21 2004.

[28] M. Corney, O. devel, A. Andersonand, and G. Mohay, "Genderpreferential text mining of e-mail discourse," in *Proceedings of the 18th Annual Computer Security Applications Conference*. Washington: IEEE Computer Society, December09-13 2002, pp. 51–63.

[29] D. R. Swanson, N. R. Smalheiser, and A. Bookstein, "Information discovery from complementary literatures: categorizing viruses as potential weapons. journal of the american society for information science," *Journal of the American Society for Information Science*, vol. 52, no. 10, pp. 797–812, 2001.

[30] P. Hu, C. Lin, and H. Chen, "User acceptance of intelligence and security informatics technology: A study of coplink," *Journal of The American Society for Information Science and Technology (JASIST)*, vol. 56, no. 3, pp. 235–244, 2005.

[31] R. Grishman, "Information extraction: Techniques and challenges," in *Proceedings of the SCIE*, 1997, pp. 207–220.

[32] H. Liu, S. B. Johnson, and C. Friedman, "Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS," *Journal of the American Medical Informatics Associations (JAMIA)*, vol. 9, pp. 621–636, 2002.

[33] H. Liu, Z. Hu, M. Torii, C. Wu, and C. Friedman, "Quantitative assessment of dictionary-based protein named entity tagging," *Journal of the American Medical Informatics Associations (JAMIA)*, vol. 13, pp. 497–507, 2006.

[34] H. M. Alfawareh and S. Jusoh, "Resolving ambiguous entity through context knowledge and fuzzy approach," *International Journal on Computer Science and Engineering (IJCSE)*, vol. 3, no. 1, pp. 410–422, 2011.

Shaidah Jusoh is currently an associate professor at the of College of Computer Science & Information Technology, Najran University, Saudi Arabia. She completed her PhD in Engineering (Engineering System and Computing, December 2005) from University of Guelph, Canada. Her PhD research is in the area of intelligent systems. She also received Master of Science in Computer Science with specialization in Distributed Information System from the University of Guelph, Canada, and Bachelor of Information Technology (with Honors) from Universiti Utara Malaysia, Malaysia. Previously she worked at Zarqa University, Jordan, Taibah University, Saudi Arabia, Universiti Utara Malaysia, Malaysia and University of Guelph, Canada. Dr. Shaidah has been an active member of editorial board committees, reviewer committees and technical program committees of international journals and proceedings. She has published numerous number of publications in international refereed journals and high quality international conference proceedings.

Hejab M. Alfawareh is currently an assistant professor at the the of College of Computer Science & Information Technology, Najran University, Saudi Arabia. Dr. Hejab Al Fawareh obtained his PhD in Information Technology with specialization in Artificial Intelligence from Universiti Utara Malaysia in 2009/2010. Dr. Hejab has taught 11 different courses in various university namely, Zarqa University, Al-Albayt University (both in Jordan), Taibah University and Saudi Arabia He is also actively involved with various committees at Zarqa University. His research interests include information extraction, ambiguity resolution, social networks, text mining and fuzzy applications. He has presented his research work in France, Malaysia, Philippines, Ukraine, and Jordan. He has published research articles in journals, book chapters, and international proceedings.