Research Journal of Applied Sciences, Engineering and Technology 11(3): 308-314, 2015
DOI: 10.19026/rjaset.11.1721
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: February 3, 2015 Accepted: March 20, 2015 Published: September 25, 2015

Research Article

Presentation Mining: An Overview of Information Extraction Systems

[1]Vinothini Kasinathan and [1, 2]Aida Mustapha
[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia Parit Raja, 86400 Batu Pahat, Johor, Malaysia

Abstract: In education, scanning through endless slides in PowerPoint presentation is highly ineffective especially for the Digital Natives due to their multi-modal learning style. In order to cater for the high volume of information emerging from printed alphabets to digital images, this study proposes a text mining approach to extract keywords from a collection of presentation slides in a similar topic. This approach is to support the existing architecture of presentation mapping, whereby the keywords extracted would then be reconstructed visually in the form of visual knowledge display. In achieving this, this study provides a general discussion of text mining technologies available and later focuses on different keyword extraction systems. Finally, this study introduces the frontier method of this field, which is presentation mining.

Keywords: Natural language processing, powerpoint, text mining

INTRODUCTION

The emergence of Internet and wide-spread use of electronic documents sheds a significant impact in managing information, as the electronic documents are becoming primary means for storing and accessing written communication (Fan *et al*., 2005). However, despite the technological advancement, the main issue still remains, which is the laborious task to search for the most relevant information as needed (Gupta and Lehal, 2009). Scanning through information in digital format is time-consuming (Thakkar *et al*., 2010) and not many people have the luxury of time to read and analyze the information. Information is also unstructured in nature and scattered throughout different places, hence making it almost difficult to retrieve the desired information without having to scan through the entire source (Zhang *et al*., 2009).

Text mining is introduced to address this issue in particular. It is the process of extracting information from different sources, recombining them to identify patterns and deriving information from such digital sources (Grobelnik *et al*., 2002). Text mining applications rise to support extraction and interpretation of unstructured text format, which in its present form, does not make a suitable input to automatic processing tasks such as information retrieval, document indexing, clustering and text classification (Chiang *et al*., 2008). Text mining is an intuitive choice of technology particularly in the research and educational field due to its ability to discover new hidden relationships from complex and voluminous body of knowledge published in the literature (Grobelnik *et al*., 2002), whether in related or non-related field of research. Among the technologies that text mining could offer include Question and Answering, Information Extraction, Topic Tracking, Text Summarization, Text Categorization, Text Clustering, Concept Linkage and Information Visualization.

Question and Answering technology focuses on searching the best answer available in its database to produce the answer. Information Extraction identifies key phrases and relationship within the text. The algorithm looks for a predefined sequence of words and then extracts the words based on some pattern matching concepts. Topic Tracking identifies and keeps user profile based on the documents the user views and is able to predict other documents which might be the interest of the user. Next, text Summarization attempts to reduce a text document in creating a summary that retains the most important points of the original document by a computer program. There are two methods of performing summarization, which is by extraction and abstraction. However, research of the former extraction method is mostly commonly used.

Text Categorization identifies the main themes of a document by placing the document into a pre-defined set of topics. This technology relies on thesaurus by ranking them into categories such as broad term, narrower term, synonyms and related terms. Next

Corresponding Author: Vinothini Kasinathan, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

technology, Text Clustering methods which can be used to automatically group the retrieved texts into a list of meaningful categories which are predefined topics. Two other technologies in text mining are Concept Linkage and Information Visualization. Concept Linkage connects related documents by identifying their common shared concepts and help users find information that they wouldn't have found using traditional search method. This promotes browsing information and is widely used in biomedical field. Lastly, Information Visualization produces output showing large textual sources in a visual hierarchy or map and provides browsing capabilities as well as simple search.

In English teaching and learning, Grobelnik *et al*. (2002) capitalized on text mining techniques to build a taxonomy or ontology from a database of documents from a huge collection of educational materials in different formats and at different educational level. The output is a uniformly formatted database of education materials based on uniform ontologies. Similar research on ontology modeling from unstructured documents that relies on text mining technologies include document categorization in biomedical informatics and information extraction for user profiling and web access analysis (Qi *et al*., 2009).

In education, text mining enables students and educators to find accurate information in specialized topic area, citations analysis or a collection of frequently asked questions (FAQ) as compared to performing the traditional ad-hoc search. Text mining is also used for analyzing service learning activities in order to discover students' learning outcome as a reflection of the service learning activities (Hsu and Chang, 2012). Similarly, it is useful in monitoring the network education public sentiment for decision-making support, which is a collection of common views related to education that are expressed openly by public on the Internet (Li *et al*., 2010). However, as text mining is considered to be a relatively new interdisciplinary field (Gupta and Lehal, 2009; Grobelnik *et al*., 2002), there are no known standards to be considered as generic to text mining applications.

One of the most recent applications in education is presentation mapping (Kasinathan *et al*., 2013), which maps keywords from presentation slides into a mind map. Current approach in presentation mapping is based on structural information of the slides that makes it rigid and not robust to a big collection of input slides. To improve the approach, this study will review information extraction approaches in text mining to improve the existing structural-based mapping into automatic keyword and keyphrase extraction. The study will also present three case studies on existing information extraction systems, which are KEA, GenEx and Text Rank.

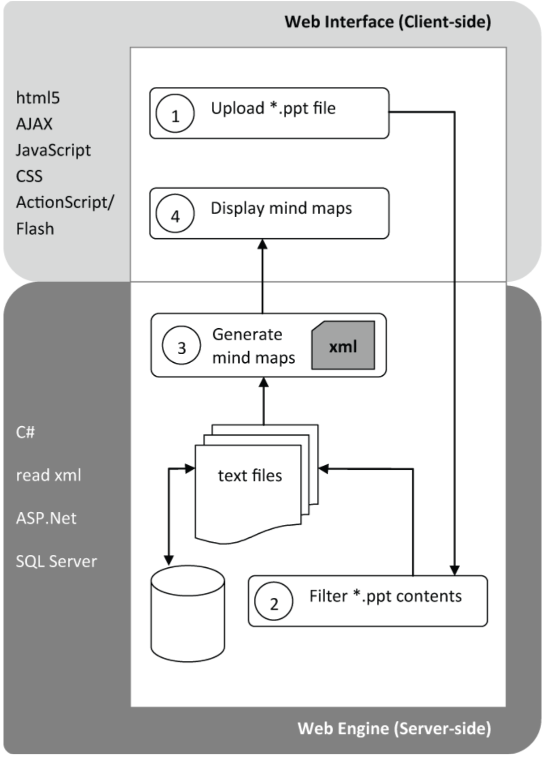

Fig. 1: Architecture for presentation mapping

Presentation mining: In education, lecture materials are often presented in the form of PowerPoint slides. While this technique is a significant leap from the chalk-and-board in-class presentation, the technique has grown old to the eyes of digital natives. The term Digital Natives was coined by Prensky (2001, 2004). In his study, he stated that Digital Natives has currently invented daily activities based on technology usage. Digital natives are more into environmental friendly with paperless communication. With regards to their learning style, copying notes would not be in their interest area, as they would attend classes without notebooks and paper but with technology devices. Digital Natives communications are always in real time communication. Being part of technology-driven and Internet-connected society, their learning style has evolved beyond slide presentation. These students struggle to organize all the lecture slides that inundate their lives and give no room for text-laden or oversimplified slides. They want their slides to be organized and prioritized and eventually, they need tools that enable them to use their lecture materials more creatively and effectively. PowerPoint style is said to routinely disrupts, dominates and trivializes the presentation content (Tufte, 2003).

In effort to cater the need of Digital Natives, Kasinathan *et al*. (2013) proposed the term Presentation Mapping, whereby keywords or key phrases are extracted from a collection of presentation slides and they are then are mapped into a visual knowledge display. Similar to a mind map or a concept map, the visual knowledge display will visually arrange the keywords and key phrases based on the content in a presentation file in accordance to the flow of individual slides. Figure 1 shows the architecture of Presentation Mapping as adapted from Kasinathan *et al*. (2013).

Based on this architecture, the title file will be captured as the main node for easy access to students. This will help students to capture the ideas from the important keywords and key phrases of a particular topic in a quicker manner. The architecture is also able to automatically generate the visual knowledge display with only a click of a button.

According to Xie *et al*. (2010), keywords provide an overview of a given document through the list of words. It also serves as representative summary of the document (Mishra and Singh, 2011), which helps the users grasp the content of the document quickly (Shi *et al*., 2008; Kongkachandra and Chamnongthai, 2008). However, this study argues that although the keywords extracted from the PowerPoint slides would give students an overall view that assists the understanding of the subject as a whole, the actual keywords on the concepts extracted is very domain-specific. Therefore, simply mapping the keywords from extracted from the presentation slides does not necessarily do justice to the domain knowledge. At the same time, certain concepts can be in the form of a keyword, while many are in phrases with varying length (i.e. natural language vs. natural language processing).

Identifying keywords in the sense of differentiating between scientific and specific knowledge requires skills, especially when the keywords are technical and the audiences are students (Ogrenci, 2012). Extracting the identified keywords is itself another challenge. To master the skill of understanding keywords is time consuming, laborious and inefficient (Huang and Wang, 2010). To date, finding and understanding keywords are usually performed using some pre-established thesaurus or manually tagged keywords. A more sophisticated keyword extraction technique is imperative to achieve the purpose in presentation mapping. From the perspective of data mining, keyword extraction from unstructured text documents has been made possible via text mining (Wang *et al*., 2008), in particular, via a concept called Information Extraction (IE). The following section will first introduce the notion of IE and present three case studies for IE systems, which are TextRank (Xie *et al*., 2010), KEA (Frank and Medelyan, 2009) and GenEx (Litvak *et al*., 2009).

INFORMATION EXTRACTION SYSTEMS: CASE STUDIES

Information Extraction (IE) enables user to extract relevant information quickly and accurately. The concept of IE originates from the field of Information Retrieval (IR), where its task is to distinguish set of documents based on a particular query (Milward and Thomas, 2000; Miner *et al*., 2012). Queries in IR, in turns, are usually in the form of keywords inputted by the users. On the other hand, IE differs from IR in the sense that it is used to extract specific information from documents, where the extracted information is analyzed for patterns or other meaningful representation (Miner *et al*., 2012). Researches on IE approaches can be broadly categorized into two; application of pre-established thesaurus or statistical techniques, used in conjunction with certain machine learning algorithms (Huang and Wang, 2010).

Pre-established thesaurus: This approach relies on the use of pre-established thesaurus as its source of keywords together with the application of machine learning algorithms for the process of keywords extraction. In this case, if some keywords of the given documents are not included in the thesaurus, it is difficult to extract all the keywords accurately from the given documents. Hence, its extensibility and portability is poor. KEA (Frank and Medelyan, 2009) is an example of keywords extractor system that is based on this approach.

Statistical approaches: This approach relies on the use of statistical information of the document content with similar application of machine learning algorithm for the keywords extraction. Because this approach does not rely on the pre-established thesaurus but the statistical information of the document instead, its extensibility and portability are better. Examples of systems that are based on this approach include TextRank (Xie *et al*., 2010) and GenEx (Litvak *et al*., 2011).

Machine learning algorithms, the base approach to either thesaurus-based or statistical-based IE systems, may be supervised or unsupervised (Xie *et al*., 2010). In supervised machine learning algorithms, the input is a set of training samples that consist of manually supplied keywords, which are required in order for the system to learn the keywords to be extracted. Hence, the supervised approaches rely heavily on the application of predictive modeling, where it is used to predict and extract the keywords based on the training model. The model is built as patterns, where the patterns are trained or derived from the training samples of the manually supplied keywords (Miner *et al*., 2012). This means, the more training samples the system has, the more accurate the keywords extracted by the system will be. However, despite its ability to produces highly accurate outputs, this approach requires a lot of training samples to achieve its accurate output (Huang and Wang, 2010).

On the contrary, unsupervised machine learning algorithms do not require training data in extracting the keywords (Xie *et al*., 2010). Instead, it directly uses the content structure or statistical information of given documents in extracting the keywords (Huang and Wang, 2010). Therefore, IE system that are based on unsupervised approaches works by identifying repeatable patterns in the given documents (Miner *et* *al*., 2012), whereby the patterns are used to identify and extract the keywords. This means, unsupervised approach relies heavily on the statistical information from the document content structure in order to analyze the given documents. Examples of unsupervised machine learning algorithms include the TFxIDF algorithm (Frank and Medelyan, 2009) and Word Co-occurrence Statistical Algorithm (Matsuo and Ishizuka, 2003).

TFxIDF is an acronym for Term Frequency Inverse Document Frequency (Zhang *et al*., 2008), which are extensively used in applications of IE systems. This measurement is also commonly used in conjunction with other algorithms such as term occurrences, length and/or distance of a word, as well as other IE algorithms such as KEA++ (Xie *et al*., 2010; Litvak *et al*., 2011). The main idea of TFxIDF is that if the term frequency of a particular word or phrase in a document is high and rarely appears in other documents, then the word or phrase is considered as important (Wang *et al*., 2012). Equation 1 shows the classical formula for TFxIDF (Yong-Qing *et al*., 2008):

$$W = TF \cdot IDF = TF \cdot \frac{1}{DF} \tag{1}$$

where W is the weight or value of a word, TF is the term frequency of a word, IDF is the inverse document frequency and DF is the document frequency. TFxIDF algorithm assumes that a term or a word represents the characteristics of documents in a particular class only if the term occurs frequently in the documents of that particular class, while occur less frequently in other documents in different class (Qu *et al*., 2008). Due to this property, it is possible that less frequent keywords get left out. Furthermore, TFxIDF also suffers from its huge computation requirement to process the value or weight of each terms occurring throughout the documents (Zhang *et al*., 2008).

Meanwhile, Word Co-occurrence Statistical algorithm (Matsuo and Ishizuka, 2003) assumes that distribution for term co-occurrence reflects the importance of a term. This means, if the probability distribution of co-occurrence between term A and frequent term is biased to a particular subset of frequent terms, therefore the term A will be considered as a keyword because the two terms in a sentence are considered to co-occur once. The importance of a term is calculated using the probability distribution as Equation 2:

$$P_g = \frac{n_g}{N_{Total}} X^2(w) = \sum_{g \in G} \frac{\left(freq(w,g) - n_w P_g\right)^2}{(n_w P_g)}$$

$$X'^2(w) = X^2(w) - \max_{g \in G} \left\{ \frac{(freq(w,g) - n_w P_g)^2}{(n_w P_g)} \right\} \tag{2}$$

where N_{Total} is the total number of different terms, n_g is the total number of terms in sentences where g appears, n_w is the total number of terms in sentences where w appears, P_g is the expected probability of G, $freq(w, g)$ is the frequency of co-occurrence of term w and term g, $X^2(w)$ is the preliminary degree of bias for co-occurrence of term w and finally $X'^2(w)$ is the final degree of bias for co-occurrence of term w.

Two established keyword extraction systems, which are the Keyphrase Extraction Algorithm (KEA) system (Wang *et al*., 2008) and Gen, Exkeyphrase extraction system (Kongkachandra and Chamnongthai, 2008) are designed based on supervised machine learning approach that requires training samples in the form of documents with manually supplied keywords. In addition, KEA also relies on the use of pre-establish thesaurus to support the extraction. Another keywords extraction system called the TextRank (Xie *et al*., 2010) is designed based on heuristics approaches, in which it applies graph-based ranking algorithm to extract the keywords.

Graph-based ranking algorithms model documents in the form of graph, which consists of text units such as term or word as the vertex and is connected by the relationship between the vertices or text units as the edge (Wei *et al*., 2008). The importance of each term or word as represented by the vertex is calculated through the use of graph-based algorithm (Zhou *et al*., 2009). One of the most common algorithms used for graph-based ranking is the TextRank algorithm, which is based on word co-occurrence concepts. In TextRank, if a vertex is linked to another vertex, that vertex is basically recommending or co-occurring with the other vertex as linked to it (Thakkar *et al*., 2010). However, the connection or co-occurrence has to be within a window of maximum *N* words, where *N* can be set between 2 to 10 words (Litvak *et al*., 2011). Therefore, the importance of a vertex is determined by the number of vertices recommending that particular vertex.

The general processes of graph-based methods are as follows. First, the text units are identified and represented in the form of vertices in the graph. Next, the relations between vertices are mapped into the graph. Third, the importance of the vertices is calculated through the implementation of graph-based ranking algorithm. Finally, the vertices are sorted based on the value calculated in the previous steps (Thakkar *et al*., 2010). In this way, the graph-based ranking algorithm does not require the use of training sample as opposed to Machine Learning approaches (Xie *et al*., 2010). The algorithm also yields independency capability, whereby it can be possibly adapted in many domains or subjects. However, despite its independency, the output of Graph-based Ranking algorithm is not able to achieve the same precision as the output generated from Machine Learning approaches, such as GenEx extraction system (Litvak *et al*., 2011).

KEA: KEA extraction system (Frank and Medelyan, 2009) is developed based on Naïve Bayesian learning algorithm, which works by building model from a list of candidate phrases extracted from the training samples. The candidate phrases contain four features, which are TFxIDF, first occurrence, length of phrase and node degree (Frank and Medelyan, 2009). Recent improvement on KEA allows the output of the KEA extraction system to be improved greatly through the use of a thesaurus-based automatic keyphrase extraction algorithm. However, this system suffers from domain specific keyphrase extraction (Wang *et al*., 2008) because the domain of thesaurus used for the keyphrase extraction has to be comparable to the domain of the inputs or documents.

GenEx: GenEx extraction system combines the parameterized heuristic keyphrase extraction rules with the application of Genetic Algorithm (GA) (Wang *et al*., 2008). GenEx uses GA in order to readjust the 12 parameters used in candidate filtering process (17). The 12 parameters used in GenEx are additional features used to pre-process the documents such as stemming, first occurrence and so forth. Although GenEx is proven to have the best extraction accuracy as compared to others, GenEx requires a complex computational method for the training phases and consumes a lot of time for training (Litvak *et al*., 2011).

Text rank: In contrast to KEA and GenEx systems that are based on supervised machine learning approach, TextRank extraction system is developed based on unsupervised approach with a simple, syntactic graph-based representation in extracting the keywords (Xie *et al*., 2010). TextRank works similarly with the concept of word co-occurrences, whereby two terms or words in a sentence are considered as co-occurring with each other (Matsuo and Ishizuka, 2003). The co-occurrences are then mapped into a graph, in which the terms and co-occur terms are labeled as vertexes and edges respectively (Wei *et al*., 2008). This means the importance of a word or vertex is determined from the global information on the graph recursively as the graph represents the co-occurrence relationship between the words (Zhou *et al*., 2009).

Next, by updating the important value of a word based on the value of its co-occurring words, the system will yield a rank of word importance (Zhao *et al*., 2010). Because TextRank heavily relies on co-occurrence relation between words in the document, the approach is simple in the sense that it is language independent and requires almost no language-specific linguistics processing (Litvak *et al*., 2011). Nonetheless, TextRank is not without drawbacks. The disadvantage lies in the process, where high frequency words are more likely to appear in the outputs of TextRank as compared to less frequency keywords (Zhao *et al*., 2010).

DISCUSSION AND CONCLUSION

Although the researches in text mining has yield considerable support for a lot of applications such as the search engines, marketing or forecasting, text mining application in educational domain is very limited. In this study, an overview of text mining techniques in general along with its applications has been presented, in particular algorithms and systems related to Information Extraction (IE). By focusing on text mining techniques, this study advocates the concept of presentation mining instead of presentation mapping as in Kasinathan *et al*. (2013). Three IE systems have also been presented, which are TextRank (Xie *et al*., 2010), KEA (Frank and Medelyan, 2009) and GenEx (Litvak *et al*., 2011).

In the future, this research will proceed to apply the Word Co-occurrence statistical algorithm TextRank (Matsuo and Ishizuka, 2003) in developing the presentation mining system, together with the implementation of Graph-based Ranking algorithm (Xie *et al*., 2010) to support the quality of the output produced by the presentation mining system. Word Co-occurrence is sufficient for the keyword extraction purposes because it is domain independent. Furthermore, Word Co-occurrence algorithm does not require any training samples or pre-established thesaurus for keywords extraction. It is hoped that through the presentation mining system, students are able to extract the relevant information from the slides without the need of scanning through the entire information. This, in turn, will help or support the diversity of learning styles among the digital natives in present years.

REFERENCES

Chiang, C.C., J. Talburt, N. Wu, E. Pierce, C. Heien *et al*., 2008. A case study in partial parsing unstructured text. Proceeding of 5th International Conference on Information Technology: New Generations (ITNG, 2008), pp: 447-452.

Fan, W., L. Wallace, S. Rich and Z. Zhang, 2005. Tapping into the power of text mining. Proceeding of Commun. ACM, 49(2): 76-82.

Frank, E. and O. Medelyan, 2009. KEA: Keyphrase Extraction Algorithm. (Online) University of Waikato (5.0). Retrieved form: HYPERLINK http://www.nzdl.org/Kea/, http://www.nzdl.org/Kea/. (Accessed on: August 20, 2012)

Grobelnik, M., D. Mladenic and M. Jermol, 2002. Exploiting text mining in publishing and education. Proceeding of the ICML-2002 Workshop on Data Mining Lessons Learned, Sydney, Australia.

Gupta, V. and G.S. Lehal, 2009. A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell., 1(1): 60-76.

Hsu, C.L. and Y.F. Chang, 2012. Qualitative text mining in student's service learning diary. Proceeding of 3rd International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA), pp: 350-354.

Huang, H. and H. Wang, 2010. Keyphrases extraction research based on structure of document. Proceeding of 2nd International Conference on Education Technology and Computer (ICETC, 2010). Shanghai.

Kasinathan, V., A. Mustapha and M.F.C.A. Rani, 2013. Structure-based algorithm for presentation mapping in graphical knowledge display. Int. J. Inform. Educ. Technol., 3(2): 196-200.

Kongkachandra, R. and K. Chamnongthai, 2008. Abductive reasoning for keyword recovering in semantic-based keyword extraction. Proceeding of 5th International Conference on Information Technology: New Generations (ITNG, 2008). Las Vegas, NV.

Li, S., X. Lv, Q. Zhou and S. Shi, 2010. Study on key technology of topic tracking based on VSM. Proceeding of IEEE International Conference on Information and Automation (ICIA, 2010), pp: 2419-2423.

Litvak, V., S.A. Ramsey, A.G. Rust, D.E. Zak, K.A. Kennedy, A.E. Lampano, M. Nykter, I. Shmulevich and A. Aderem, 2009. Function of C/EBPdelta in a regulatory circuit that discriminates between transient and persistent TLR4-induced signals. Nat. Immunol., 10: 437-443.

Litvak, M., M. Last, H. Aizenman, I. Gobits and A. Kandel, 2011. DegExt: A language-independent graph-based keyphrase extractor. In: Mugellini, E., P.S. Szczepaniak, M.C. Pettenati and M. Sokhn (Eds.), Proceeding of the 7th Atlantic Web Intelligence Conference, AWIC 2011. Fribourg, Switzerland.

Matsuo, Y. and M. Ishizuka, 2003. Keyword extraction from a single document using word co-occurrence statistical information. Proceeding of the 16th International Florida Artificial Intelligence Research Society Conference. St. Augustine, 2003. The AAAI Press.

Milward, D. and J. Thomas, 2000. From information retrieval to information extraction. Proceeding of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (RANLPIR '00), Vol. 11, Association for Computational Linguistics, Stroudsburg.

Miner, G., J. Elder IV, A. Fast, T. Hill, R. Nisbet *et al*., 2012. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. 1st Edn., Elsevier Science, Burlington.

Mishra, A. and G. Singh, 2011. Improving keyphrase extraction by using document topic information. Proceeding of IEEE International Conference on Granular Computing (GrC). Kaohsiung.

Ogrenci, A.S., 2012. Empirical results about efforts for effective teaching to Y-generation freshman students. Proceeding of International Conference on Information Technology Based Higher Education and Training (ITHET). Istanbul.

Prensky, M., 2001. Digital natives, digital immigrants part 1. Horizon, 9(5): 1-6.

Prensky, M., 2004. The emerging online life of the digital native. Retrieved form: http://www.marcprensky.com/writing/Prensky-The_Emerging_Online_Life_of_the_Digital_Native-03.pdf. (Accessed on: Jan, 13, 2009)

Qi, Y., Y. Zhang and M. Song, 2009. Text mining for bioinformatics: State of the art review. Proceeding of the 2nd IEEE International Conference on Computer Science and Information Technology (ICCSIT, 2009). Beijing.

Qu, S., S. Wang and Y. Zou, 2008. Improvement of Text Feature Selection Method based on TFxIDF. Proceeding of the International Seminar on Future Information Technology and Management Engineering (FITME ’08). Leicestershire, United Kingdom.

Shi, T., S. Jiao, J. Hou and M. Li, 2008. Improving keyphrase extraction using wikipedia semantics. Proceeding of the 2nd International Symposium on Intelligent Information Technology Application (IITA '08). Shanghai.

Thakkar, K.S., R.V. Dharaskar and M.B. Chandak, 2010. Graph-based algorithms for text summarization. Proceeding of the 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET, 2010). Goa.

Tufte, E., 2003. PowerPoint is evil. Wired Magazine, September.

Wang, X.L., D.J. Mu and J. Fang, 2008. Improved automatic keyphrase extraction by using semantic information. Proceeding of the International Conference on Intelligent Computation Technology and Automation (ICICTA, 2008). Hunan.

Wang, X., J. Cao, Y. Liu, S. Gao and X. Deng, 2012. Text clustering based on the improved TFIDF by the iterative algorithm. Proceeding of the IEEE Symposium on Electrical and Electronics Engineering (EEESYM, 2012). Kuala Lumpur, pp: 140-143.

Wei, F., Y. He, W. Li and Q. Lu, 2008. A query-sensitive graph-based sentence ranking algorithm for query-oriented multi-document summarization. Proceeding of 2008 International Symposiums on Information Processing (ISIP, 2008). Moscow.

Xie, F., X. Wu and X. Hu, 2010. Keyphrase extraction based on semantic relatedness. Proceeding of 9th IEEE International Conference on Cognitive Informatics (ICCI, 2010). Beijing.

Yong-Qing, W., L. Pei-Yu and Z. Zhen-Fang, 2008. A feature selection method based on improved TFIDF. Proceeding of 3rd International Conference on Pervasive Computing and Applications (ICPCA, 2008). Alexandria, pp: 94-97.

Zhang, W., T. Yoshida and T. Xinjin, 2008. TFIDF, LSI and multi-word in information retrieval and text categorization. Proceeding of IEEE International Conference on Systems, Man and Cybernetics (SMC, 2008). Singapore, pp: 108-113.

Zhang, X., Z. Guo and B. Li, 2009. An effective algorithm of news topic tracking. Proceeding of WRI Global Congress on Intelligent Systems (GCIS ’09). Xiamen, pp: 510-513.

Zhao, L., L. Yang and X. Ma, 2010. Using tag to help keyword extraction. Proceeding of International Conference on Computer and Information Application (ICCIA, 2010). Tianjin, pp: 95-98.

Zhou, B., P. Luo, Y. Xiong and W. Liu, 2009. Wikipedia-graph based key concept extraction towards news analysis. Proceeding of IEEE Conference on Commerce and Enterprise Computing (CEC '09). Vienna, pp: 121-128.