VisualUrText: A Text Analytics Tool for Unstructured Textual Data

Zuraini Zainol¹, Mohd T.H. Jaymes² and Puteri N.E. Nohuddin³

¹Computer Science Department, Faculty of Science and Defence Technology, Universiti Pertahanan Nasional Malaysia, Kem Sungai Besi, Kuala Lumpur, Malaysia
²,³Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Malaysia

zuraini@upnm.edu.my, tarhamizwan123@outlook.com, puteri.ivi@ukm.edu.my

Abstract. The growing amount of unstructured text over Internet is tremendous. Text repositories come from Web 2.0, business intelligence and social networking applications. It is also believed that 80-90% of future growth data is available in the form of unstructured text databases that may potentially contain interesting patterns and trends. Text Mining is a well known technique for discovering interesting patterns and trends which are non-trivial knowledge from massive unstructured text data. Text Mining covers multidisciplinary fields involving information retrieval (IR), text analysis, natural language processing (NLP), data mining, machine learning statistics and computational linguistics. This paper discusses the development of text analytics tool that is proficient in extracting, processing, analyzing the unstructured text data and visualizing cleaned text data into multiple forms such as Document Term Matrix (DTM), Frequency Graph, Network Analysis Graph, Word Cloud and Dendogram. This tool, VisualUrText, is developed to assist students and researchers for extracting interesting patterns and trends in document analyses.

1. Introduction

In the era of big data, government, corporates and business and even society are contributing to produce and use substantial amount of data. Data is often referred to facts, number or text, which can be processed by computer [1]. Data is limitless and presented everywhere in the universe. In computer system, data is referred to input symbols or signals that are stored and processed by the computer to produce information. Recently, the advancement of computing technology and devices has led to the rapid growth of data sources such as social networking applications, images, online news, email messages, medical records, etc. This huge amount of data is believed to have interesting patterns and valuable hidden knowledge within this data. This situation has attracted a large number of researchers to study on data mining techniques to extract interesting information and knowledge which can assist the stakeholders in decision making and prediction particularly in business, risk management, and medical research, etc. Data mining is an analytics process that is designed to explore large amount of data in search of consistent patterns or systematic relationships between data features and then validate the findings by applying the detected patterns to new subsets of data [2]. Technically, data mining is a process of finding correlations or patterns among dozens of fields in a large number of data. Data mining covers wide areas of research in text mining, web mining, image mining, medical mining, graph mining, etc.

The increment of text data has also increased its interest in our daily life. According to [3], 80-90% of future world’s data growth is expected from the unstructured text databases or document databases.
In other words, most of available information is kept as text databases, which contain a high commercial potential value in discovering and predicting interesting patterns and trends. When it comes unstructured text, the recurrent text processing always involves text analysis and content analysis. By using data mining tools for example RapidMiner, Weka, Sentinel Visualizer, R, Orange and many more, users are able to analyze data from different dimensions, categorize, and summarize the text features and relationships can be identified. Technically, data mining is a process of finding correlations or patterns among dozens of fields in a large number of data [4]. The aim of this paper is to develop a text analytics tool that is capable of extracting, processing, analyzing the unstructured text data and visualizing the clean text data into multiple forms such as Document Term Matrix, Frequency Graph, Network Analysis Graph, Word Cloud and Dendogram. This tool called VisualUrText is developed to assist the students and researchers in knowledge discovery of analyzing documents. The rest of this paper is organized as follows. Section 2 explores Knowledge Discovery in Databases and Text Mining. Section 3 discusses the framework for visualizing unstructured text data. Section 4 discusses the experiment and results. Finally, we conclude this paper with future work in Section 5.

2. Related Work and Background

In this section, we discuss several related topics on Knowledge Discovery in Databases (KDD), text mining and some reviews on text analyzer tools.

2.1. Knowledge Discovery in Databases

Knowledge discovery in databases (KDD) or sometimes also referred as data mining is a process of discovering interesting patterns (non-trivial, implicit, previously unknown and potentially useful) or knowledge from a large of databases [5]. The KDD involves several processes such as data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. As well as the prominence on gathering data growing around the world, there is a requirement for another era of strategies, different techniques and algorithms to help analysts, decision makers, researchers and managers in squeezing out the valuable patterns from increasing volumes of data. These tools and techniques are the subject of the appearing field of knowledge discovery in databases.

![Figure 1. Overview of steps constituting the KDD process adopted from [6]](image)

Figure 1 shows a general KDD process which consists of seven main essential steps which starts from raw data repository to new knowledge. The steps are described as follows [7]:

- Data Cleaning: This phase deals with missing values, noisy data and inconsistent data. For example, remove advertisement from web pages, detect and correct inaccurate records, normalize text converted from binary formats, deal with tables, etc.
- Data Integration: The phase involves with combining or merging data from multiple data stores.
- Data Selection: This is an important phase as it will select the relevant data to the analysis.
- Data Transformation: This phase exchanges the required data into a compatible format using data mining methods such as smoothing, aggregation, generalization, normalization and discretization, etc.
- Data Mining: This phase will apply intelligent methods to extract useful patterns and knowledge.
- Pattern Evaluation: Interesting patterns data represent the identified knowledge.
- Knowledge Representation: The discovered knowledge is then displayed to users in a visual form. Generally, visualization techniques will be applied to assist users in understanding and interpret information.

Although these are the main steps in any KDD process, some of the steps could be executed collectively during the actual process. For example, data cleaning, data transformation, data selection and data transformation can be combined together. Basically, the first four steps are listed above can be considered as data preparation for data mining process whereas the last three steps can be seen as post-processing step [8].

2.2. Text Mining [1]

Text Mining (TM) or text data mining refers to an analytic process, which is designed to explore the unstructured text documents in search of useful information and knowledge hidden from a large amount of text resources [9]. In business world, this useful information may reveal interesting patterns or semantic relationships and trends in large volumes of unstructured data. In TM, patterns are extracted from natural language text rather than databases which can lead to new business opportunities and improvements in processes. Previous studies in the field of knowledge data discovery have estimated that 80% of company available information is in the form of text documents [3]. Based on this fact, it is believed that TM has a higher potential value compared to data mining. TM is an extension of data mining, which involves multiple disciplines areas such as information retrieval (IR), Statistics, Web Mining, Computational Linguistics and Natural Language Processing (NLP) [10-11]. TM can also be described as intelligent text analysis, text data mining and knowledge discovery in text (KDT). TM is different compared to web search or directory search site. In the web search, people are usually looking for what has already been known or written by someone. In comparison, the aim of the TM is to discover something that is nobody knows or not yet written down. It mines information and knowledge from a mountain of texts [12]. TM has been widely applied in many real world applications and is reported to have been applied in large-scale problems in both data science and business [13]. For example:

- prevent cybercrime – using TM intelligence and anticrime applications would make the internet crime prevention easier. For example, [14] discovered social relations between criminals from unstructured text using text mining method. [15] proposed a method to analyze criminal networks in a set documents that is able to discover both direct and indirect relationships among members in a crime community.
- Sentiment analysis – text analytics has been utilized in analyzing large volumes of unstructured data such as extracting opinions, emotions, sentiments and their relations with products. For instance, [16] proposed a technique to group the forums into several clusters for online communities.
- biomedical applications – TM can be applied as knowledge based search engine doe medical text. For example, PubGene is an application of TM in biomedicine.
- Documents clustering and frequent patterns – TM is also capable of extracting frequent patterns and interesting relationships in set of documents. For example, [12] introduced a keyword processing technique to cluster hadith chapters as references for Islamic scholars. [9] investigated the frequent patterns and interesting relationships in hadith chapters to that can benefits Muslim scholars in their daily Islamic life and practice using the technique of Association Rule Mining. [17-18] proposed a framework to discover significant text patterns...
exist in military peacekeeping documents.

3. The Framework For Visualizing Unstructured Text Data

Figure 2 illustrates the framework for Visualizing Unstructured Text Data. It consists of two main phases: (i) Unstructured Text Data Pre-processing and (ii) Analyzing and Visualizing Selection of Keyword.

![Figure 2. A Framework for Visualizing Unstructured Text Data](image)

The first component consists of the pre-processing task while the second component is mainly focused on analyzing and visualizing the selection keywords/terms. The input materials being processed in this framework was a collection of text data. The details of each module are discussed in the following sub-sections.

3.1. Unstructured Text Data Pre-processing Phase

For investigation purposes, this experiment will apply a sample of six text documents (files) which are selected randomly from articles on data mining. Text pre-processing phase is the most important step in text analysis as it removes errors and inconsistencies from data in order to improve the quality of data. In this phase, the extracted documents need to be processed in order to facilitate the text analysis. Real world data is often incomplete and inconsistent. The pre-processing data can improve the accuracy and effectiveness of text mining [19]. In this study, the text pre-processing method consists of the following two basic steps: removing stop words and reducing words into their grammatically root or stem through stemming. Stop words is a list of words that normally does not carry any meaning and information such as ‘a’, ‘about’, ‘after’, ‘all’, ‘before’, ‘such’, ‘the’, ‘some’, ‘she’, etc. The most frequently used lists of stop word are available at www.ranks.nl/stopwords. These words need to be removed because they are most probably irrelevant for this study. The second step of pre-processing is stemming. A large document may consist of many words that often have similar common root. Stemming is a process that chops off the end of words without considering linguistics features of the words. In other words, stemming is a process of getting the basic word of a particular sentence. For instance, the words ‘used’, ‘using’, and ‘uses’ would all be stemmed to the same word that is ‘use’. In this study, we apply one of the most popular algorithms for stemming English words, which is known as Porter algorithm. For example, Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. The example of steps of pre-processing is presented in Table 1:

| Table 1. Steps of Preprocessing Unstructured Text Data |
### Function

| Removing stop words whitespace | data mining extraction hidden predictive information large databases powerful new technology great potential help companies focus most important information data warehouses |
|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Stemming                       | data mine extract hid predict inform larg databas power new technolog great potenti help compani focu most import inform data warehouse                                                                 |

3.2. **Analyzing and Visualizing Selection of Keyword**

In this phase, all cleaned text data is then ready for selecting keywords and visualizing them in multiple forms such as DTM, frequency graph, network analysis graph and dendogram. In keyword selection, important keywords are selected and their presence in each file is counted. DTM is a conceptual representation that represents the relationship between m keywords and n files. In this study, each row represents for a file whereas each column represents a keyword. Each cell contains values that show the number of occurrences of the keyword in the file. For instance, if the keyword appears in a file then the matrix entry for corresponding to that particular row and column is set as 1. If the keyword appears twice then it will be recorded as two in that particular matrix entry. Many data mining techniques can be carried out on DTM such as clustering, classification, and association analysis. The next step is to illustrate the top frequent terms/keywords visually using a bar plot. The number of minimum frequency can be set up by users by inserting the number of frequency at the bottom of this application tool. After building the DTM, the importance of terms/keywords with a word cloud can be easily produced. A word cloud, is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or colour.

4. **Experimental Result And Analysis**

This section discusses the experimental result for the proposed visual unstructured text data (VisualUrText). In this experiment, six text files were selected and uploaded into the VisualUrText tool for cleaning before they can be applied into DTM. As mentioned earlier in Section 3, the text pre-processing involves two main steps such as removing stop words and stemming.

![Figure 3. The screenshot for Document-Term Matrix (DTM)](image)

In remove stop words phase, stop words such as articles (‘a’, ‘an’, ‘the’, etc.), conjunctions (‘and’, ‘but’, etc.), qualifiers (‘yet’, ‘however’, etc.) are removed whereas the stemming phase will reduce the a word into the root form such as the keyword ‘mining’ can be reduced into ‘mine’. The next step is to generate a DTM. DTM is applied to represent the relationship between keywords and files, where each column refers to a keyword and each row represents files that were used in this experiment. The number of occurrences of each keyword in each file is presented in matrix to that particular row and column. As shown in Figure 3, the DTM is composed of 340 keywords and 6 files. The extract from DTM presents the keyword ‘data’ has the highest frequencies that occurs 55 times in 6 files while the keyword ‘extract’ presence once in file number 1 but not present in any of other 5 files. The complete list of keywords occurrence in six files is presented in Figure 3.
The next step is to find the frequent term with frequency is not less than six. In order to present the most frequent words visually, a bar plot is applied in this study. From the TDM, we derive the frequency of terms. As seen in Figure 4, the two most frequent keywords are ‘data’ and ‘mine’ are clearly shown in the bar plot where the minimum number of frequency is set as six in this experiment.

After building a DTM, the importance of keywords with a word cloud can be displayed. Figure 5 illustrates that a word cloud (or tag cloud, text cloud) that highlight the most frequently used keywords in text data. Word cloud is easy to understand and readable as it provides multiple choices of colors that symbolizing the keywords with different sizes. As presented in Figure 5, the keywords ‘data’, ‘mine’, ‘database’ are the top three most important keywords in six files, which validates that six files present information that related to data mining and databases. Some other important words are ‘business’, ‘predict’, ‘technology’ and ‘process’ which present the definition of data mining is technology that predict trends for business. What we can observe here is that the size of keywords is corresponding to the frequency of the keyword.
Figure 6 shows the network of keywords which is based on their co-occurrence in six files. The network analysis graph is plotted to show the relationships between frequent keywords. In order to make the graph more readable, colors and font sizes are set in this experiment. Figure 7 presents a dendogram (tree structured graph) that visualizes the result of hierarchical clustering analysis. It presents the arrangement of the clusters for 340 keywords in six files. From the dendrogram, we can see that the cluster analysis has eight clusters where the keyword ‘data’ is placed in the seventh group.

5. Conclusions And Future Work
This study proposes a text analytics tool that is capable of extracting, processing, analyzing the unstructured text data and visualizing the clean text data into multiple forms such as Term Document Matrix, Frequency Graph, Network Analysis Graph, Word Cloud and Dendogram. For future work, this research will extend the VisualUrText by incorporating the following features: (i) the lemmatization feature for reducing an inflectional or derivationally related word form to its base form, (ii) the enhancement of dendogram feature in term of customization of selected important keywords in documents before the visualization of result can be carried out. The setting of colours and font sizes would also assist the users in reading the graph and (iii) to upgrade the VisualUrText tool into web-based applications.
References

[1] Mourya, S. K., and Shalu Gupta. “Data mining and data warehousing”. Alpha Science International, Ltd, 2012.
[2] Punjabi, B., & Honale., S., “A Research on Big Data Analysis & Processing With Data Mining Techniques”, 2015
[3] EMC Education Services. “Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data”, John Wiley & Sons, USA, 2015
[4] “Data Mining: What is Data Mining?”. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[5] Osmar R. Zaiane. “Introduction to Data” Mining”. https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1
[6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. From Data Mining to Knowledge Discovery in Databases. AI magazine, 1996, 17(3), 37.
[7] Han, J., Kamber, M. and Pei, J., “Data Mining: Concepts and Techniques”, (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.
[8] Fivelstad, Ole Kristian. "Temporal text mining: the ttm testbench." (2007).
[9] Z. Zainol, P. N. E. Nohuddin, M. T. H. Jaymes, and S. Marzukhi, "Discovering “interesting” keyword patterns in Hadith chapter documents," in 2016 International Conference on Information and Communication Technology (ICICTM), 2016, pp. 104-108.
[10] Gupta V, Lehal G.S. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence. 2009 Jan 8;1(1):60-76
[11] Irfan R, King CK, Grages D, Ewen S, Khan SU, Madani SA, Kolodziej J, Wang L, Chen D, Rayes A, Tziritas N. “A survey on text mining in social networks”. The Knowledge Engineering Review. 2015 Mar 1;30(02):157-70.
[12] Nohuddin, P.N.E., Zainol, Z., Chao, K.F., Tarhamizwan, M., and Nordin, A. “Keyword based Clustering Technique for Collections of Hadith Chapters.” International Journal on Islamic Applications in Computer Science And Technologies – IJASAT. 2015, 4(3), pp.11-18
[13] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37.
[14] Elyezjy, N. T., & Elhaless, A. M. Investigating Crimes using Text Mining and Network Analysis. International Journal of Computer Applications, 2015,126(8).
[15] Al-Zaidy, R., Fung, B. C., Youssef, A. M., & Fortin, F. Mining criminal networks from unstructured text documents. Digital Investigation, 2012, 8(3), 147-160.
[16] Li, N., & Wu, D. D. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision support systems, 2010, 48(2), 354-368.
[17] Zainol Z., P.N.E. Nohuddin, T.A.T. Mohd and O Zakaria. Text analytics of unstructured textual data: a study on military peacekeeping document using R text mining package. International Conference on Computing and Informatics, 1–7. School of Computing, UUM. 2017
[18] Zainol Z., Marzukhi S., Nohuddin P.N.E., Noormaanshah W.M.U., Zakaria O. Document Clustering in Military Explicit Knowledge: A Study on Peacekeeping Documents. In: Badioze Zaman H. et al. (eds) Advances in Visual Informatics. IVIC 2017. Lecture Notes in Computer Science, vol 10645. Springer, Cham, 2017, 175-184
[19] Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." IEEE Data Eng. Bull. 23.4 (2000): 3-13.