Computers in Human Behavior xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Selection criteria for text mining approaches

Hussein Hashimi *, Alaaeldin Hafez, Hassan Mathkour

College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Text mining approaches
Classification
Clustering
Selection criteria

ABSTRACT

Text mining techniques include categorization of text, summarization, topic detection, concept extraction, search and retrieval, document clustering, etc. Each of these techniques can be used in finding some non-trivial information from a collection of documents. Text mining can also be employed to detect a document's main topic/theme which is useful in creating taxonomy from the document collection. Areas of applications for text mining include publishing, media, telecommunications, marketing, research, healthcare, medicine, etc. Text mining has also been applied on many applications on the World Wide Web for developing recommendation systems. We propose here a set of criteria to evaluate the effectiveness of text mining techniques in an attempt to facilitate the selection of appropriate technique.

1. Introduction

Knowledge about data or text mining from important and relatively larger database has been recognized by numerous scholars and researchers. Data mining or knowledge discovery, works well on data stored in a structured manner. Often, the data that has not been well structured yet still contains a lot of hidden information. Text mining entails automatically analyzing a corpus of text documents and discovering previously hidden information. The result might be another piece of text or any visual representation. We start by extracting the useful information from text like facts and events and eventually perform some data mining tasks to gain new knowledge. Text mining generally includes categorization of information or text, clustering the text, extraction of entity or concept, development and formulation of general taxonomies.

Text mining deals with unstructured or textual information for the extraction of meaningful information and knowledge from huge amount of text. They are required for the efficient analysis and exploration of information available in text form. Text mining is required to convert the text into data which then pass through other data mining techniques for analysis. Most of the times, data that we gather from different sources is so large that we cannot read it and analyze it manually so we need text mining techniques to deal with such data. Identifying and separating out any specific type of information from the given text requires text mining techniques or methods. These methods also help in clustering the data into different groups on the basis of specific requirements. In the field of education, text mining techniques helps to explore and analyze data coming from new discoveries and researches that are made on daily basis in large amount. Text mining methods are also required whenever we need to validate extensive data by analyzing it with some special criteria. Text mining includes statistical, linguistic and machine learning techniques that are needed for studying and examining textual information required for further data analysis, research and investigation.

From the available literature and applications, text mining is used heavily in different domains such as

- Web document based text clustering (Ahmad & Khanum, 2010; Bhushan, Pushkar, Shivaji, & Nikhil, 2014; Navaneethakumar & Chandrasekar, 2012).
- Information retrieval (Rath, Jena, Nayak, & Bisoyee, 2011; Senellart & Blondel, 2008; Vashishta & Jain, 2011).
- Knowledge transfer and integration (Achtert et al., 2006; Kriegel, Kröger, & Zimek, 2009; Silwattananusarn & Tuamsuk, 2012).
- Topic tracking (Krause, Leskovec, & Guestrin, 2006; Patel & Sharma, 2014).
- Summarization, categorization, clustering, and concept linkage (Caropreso, Matwin, & Sebastiani, 2009; Kriegel et al., 2009; Lehmam, 2010; Lincy Liptha, Raja, & Tholkappia Arasu, 2010; Navaneethakumar & Chandrasekar, 2012; Patel & Sharma, 2014; Senellart & Blondel, 2008).
- Information visualization and question answering (Burley, 2010; Don et al., 2007).
- Emotional contents of texts in online social networks (Dhawan, Singh, & Khanchi, 2004, 2014; Shelke, 2014).

* Corresponding author.
E-mail addresses: ha1426@yahoo.com (H. Hashimi), ahafez2001@yahoo.com (A. Hafez), binmathkour@yahoo.com (H. Mathkour).

http://dx.doi.org/10.1016/j.chb.2014.10.062
0747-5632/

- Data collection, database schemas, data processing (Don et al., 2007; Kiyavitskaya, Zeni, Mich, Cordy, & Mylopoulos, 2006; Tan & Lambrix, 2009; Zhai, Velivelli, & Yu, 2004).
- . . .. etc.

There is a need of fast, automatic and intelligent computational power that can deal with huge data, extract required information, and help us to predict future aspects in small amount of time e.g. in business, education, security systems, etc. Text mining has many advantages:

- Help extract useful information from bulk of data in short time and efficiently.
- Assist in predicting future aspects based on provided observations and statistics.
- Help to create and build patterns from the provided data which tells us about increasing or decreasing trends, e.g. in business and economy.
- Text mining software's also helps in security agencies by monitoring and analysis of textual data gathered from internet sources blogs, etc.

Another advantage of text mining techniques is their use in biomedical databases, where these techniques improve the search from literature. Text mining methods advances the analysis, storage and availability of information on different websites and search engines to make the process of searching more efficient and more accurate. It also deals with lexical analysis and pattern recognition and helps to study word frequency distribution. The text mining process has the basic stages depicted in Fig. 1.

2. Related work

Text mining involves all activities in discovery of information and other pertinent data from a variety of textual sources. However, the extracted data have been always of little value in its raw formats. In many instances, people confuse Text Mining with the regular web search. As much as both result in acquisition of data, a large gap exists on the input. In a common web search, users are dedicated toward acquiring specific data, which may be mostly, entails looking for known and/or specified data (Achtert et al., 2006).

Navaneethakumar and Chandrasekar (2012) have studied a consistent web document based text clustering. A comparison has been conducted between new mining methods for web documents and existing clustering process. Performance of the proposed web document clustering method has been analyzed with the concept based mining models using a different set of datasets with F-Measure and Entropy measures (Kriegel et al., 2009). A model has been proposed to improve clustering efficiency.

Gupta (2009) worked on the application domain where text mining can be used in information retrieval, topic tracking, summarization, categorization, clustering, concept linkage, information visualization and question answering. Yassine and Hajj (2010) have focused on extracting emotional contents of texts in online social networks. A new framework has been proposed for characterizing emotional interactions in social networks. This proposed framework includes a model for data collection, database schemas, data processing and data mining steps. In Gharehchopogh and Abbasi Khalifehlou (2011), have addressed issues related to unstructured data. Their work is built on natural language processing and artificial intelligence techniques to extract actionable information from unstructured data.

Different studies have focused on algorithm performances and on deploying new pattern techniques to improve the efficiency of pattern based methods (Wu, Li, & Xu, 2006). A better clustering quality has been achieved by extracting the semantic structure of sentences in documents (Lincy Liptha et al., 2010; Wang et al.,1999). In case of normal and uniform distributions K-Means algorithm is better than that of FCM (Scott & Matwin, 2011). Text Categorization dimensionality of feature space is an important parameter (Velmurugan & Santhanam, 2010). Text mining and natural language processing techniques have the capability of understanding the semantics of web texts (Gharehchopogh & Abbasi Khalifehlou, 2011). The notion that dimension reduction should only be performed on pre-processing stage of any document classification (Caropreso et al., 2009; Howland & Park, 2007). Computer-written thesauri have several advantages such as ease to build and maintain (Senellart & Blondel, 2008). Semantic distances have resulted in more robust and stable subspace clustering (AlSumait & Domeniconi, 2007). A new perspective for studying friendship relations and emotions' expression in online social networks where it deals with the specific nature of these sites and the nature of the language used (Yassine & Hajj, 2010). There is a noticeable opportunity of bringing text mining and knowledge discovery techniques into the field of economics and public policy where the research will foster the awareness of cross-disciplinary research and enrich collaboration between social science and computer science paradigms (Zhou, Zhang, Vonortas, & Williams, 2012).

3. The proposed selection criteria

In this work, we propose a selection technique that is based on determining weighting of text mining criteria based on the number of those papers whom emphasized on each specific criterion. We have calculated criteria's weights after surveying more than 130 research papers in different text mining techniques publications.

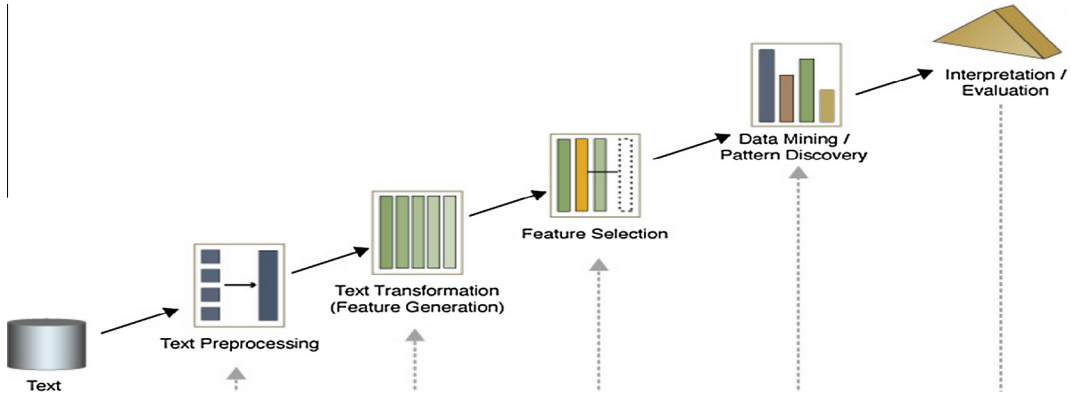

Fig. 1. Text mining process.

Please cite this article in press as: Hashimi, H., et al. Selection criteria for text mining approaches. *Computers in Human Behavior* (2015), http://dx.doi.org/10.1016/j.chb.2014.10.062