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.

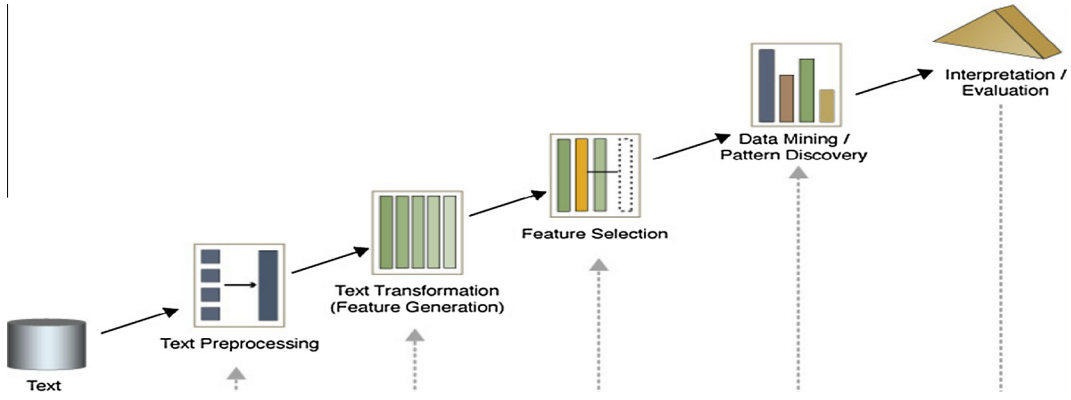

Fig. 1. Text mining process.

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

Table 1
Proposed evaluation criteria.

Criteria		Weight (%)		Connectivity based clustering models		Centroid models		K-mean clustering		Subspace models		Distribution model	
				Available	Score	Available	Score	Available	Score	Available	Score	Available	Score
General	Usability	16	4	2	0.08	1	0.04	3	0.12	2	0.08	2	0.08
	Comprehensiveness		3	1	0.03	1	0.03	2	0.06	1	0.03	3	0.09
	Flexibility		4	2	0.08	2	0.08	2	0.08	1	0.04	1	0.04
	Complexity		3	2	0.06	1	0.03	2	0.06	2	0.06	2	0.06
	GUI		2	2	0.04	1	0.02	2	0.04	1	0.02	1	0.02
Specific	*Achievement criteria*												
	Business goals satisfaction	9	4	2	0.08	2	0.08	2	0.08	1	0.04	3	0.12
	Specific Goals Achievement		5	3	0.15	1	0.05	2	0.10	1	0.05	2	0.10
	Support criteria												
	Support KPI	12	3	1	0.03	1	0.03	1	0.03	3	0.09	1	0.03
	Support MLI		2	1	0.02	1	0.02	1	0.02	1	0.02	1	0.02
	Support compliance		4	2	0.08	2	0.08	2	0.08	2	0.08	2	0.08
	Support to Individual Contents		3	1	0.03	2	0.06	1	0.03	1	0.03	3	0.09
	Data analysis criteria												
	Extraction	33	7	2	0.14	2	0.14	2	0.14	2	0.14	2	0.14
	Aggregation		5	2	0.10	2	0.10	3	0.15	1	0.05	2	0.10
	Clustering		5	3	0.15	3	0.15	3	0.15	2	0.10	2	0.10
	Indexing		6	2	0.12	1	0.06	1	0.06	1	0.06	1	0.06
	Accumulation		4	2	0.08	2	0.08	2	0.08	2	0.08	2	0.08
	Summarization		6	1	0.06	1	0.06	1	0.06	1	0.06	1	0.06
	Miscellaneous												
	Identification	30	5	2	0.10	2	0.10	1	0.05	2	0.10	2	0.10
	Effectiveness of Content		5	1	0.05	2	0.10	2	0.10	1	0.05	2	0.10
	Text Interface		4	1	0.04	2	0.08	1	0.04	2	0.08	1	0.04
	Passive		2	1	0.02	1	0.02	1	0.02	1	0.02	1	0.02
	Contraction		5	1	0.05	2	0.10	2	0.10	2	0.10	2	0.10
	Abbreviation		5	1	0.05	1	0.05	1	0.05	1	0.05	1	0.05
	Achievement of Text Learning		4	2	0.08	2	0.08	2	0.08	1	0.04	1	0.04
Total		100		1.72		1.63		1.78		1.46		1.72	

Each publication could include several criteria. We have used the scale of 2–7 to determine the weights of the selected criteria. For example, if we are only interested in 12 criteria, and the percentage of those publications that are emphasizing on each of the chosen criteria are 25%, 24%, 22%, 20%, 18%, 12%, 9%, 7%, 5%, 4% and 2%. Then the assigned weights of those 12 criteria should be 7, 7, 6, 6, 5, 5, 4, 4, 3, 3, 2, and 2, respectively.

We also include another measure called availability that represents the percentage of occurrences of a specific criterion in its related publications. The availability percentage is mapped to the range 1–3 values. The score of each criterion is calculated by multiplying the criterion's weight by the availability of occurrences of that criterion in that specific method.

Text mining criteria are classified into two main categories, the general criteria category and the specific criteria category. Such classification is based on the surveyed publications (Kotsiantis, 2007). Through the literature review and previous studies, we have identified the general or specific criteria that satisfy business goals. In the general criteria category, which could be considered as the non-functional requirements, we include the weights of Usability, Comprehensive, Flexibility, Complexity and GUI. Specific criteria could be considered as the functional requirements that include

- Business goals satisfaction: The success an organization hopes to achieve during its time in operation; the mission statements
- Specific Goals Achievement: The special goals that are achieve the organization aspirations
- Support KPI: The performance measurements of text mining
- Support MLI: The functions that are used to achieve text mining purposes through interface to deal with different needs for text mining
- Support compliance: The implementation of security measures for IT aspects which offers skills help to reduce the risks and errors rate in organization
- Support to Individual Contents: Easiness in managing and customizing contents that are extracted by text mining techniques
- Extraction: The useful information in unstructured and semi-structured forms of data
- Aggregation: The combination of data (possibly from a variety of sources) to facilitate analysis and make individual-level data deducible
- Clustering: The automatic document organization, information filtering, fast retrieval, and topic extraction
- Indexing: Map objects to a – dimensional space where distances between objects are roughly preserved
- Accumulation: Gather objects of value from un-structured and semi structured data
- Identification: Work with text mining techniques that are based on identification to determine the useful objects or terms that can be extracted from unstructured and semi-structured data
- Effectiveness of Content: Construction of useful extracted information from unstructured data
- Text Interface: Access and graphical exploration of data

(*continued on next page*)

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

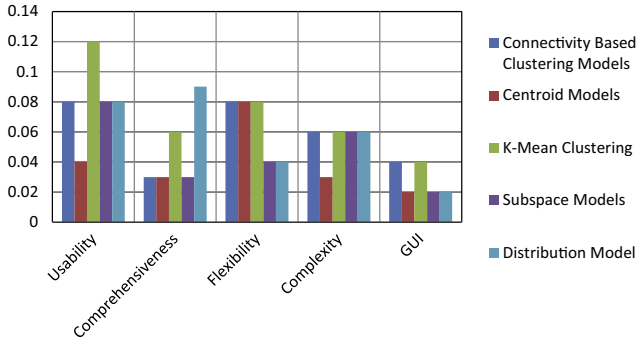

Fig. 2. General criteria for clustering techniques.

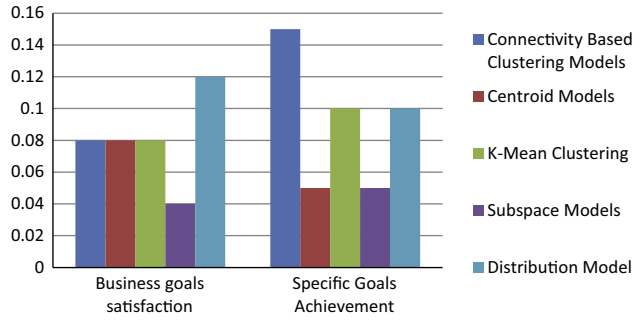

Fig. 3. Achievement criteria.

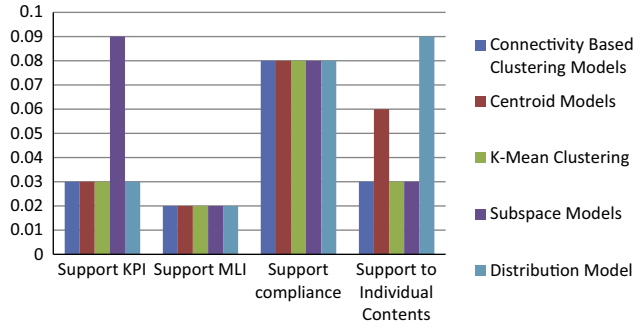

Fig. 4. Support criteria.

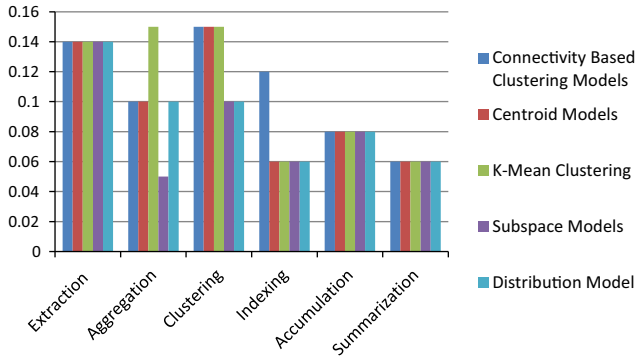

Fig. 5. Data analysis criteria.

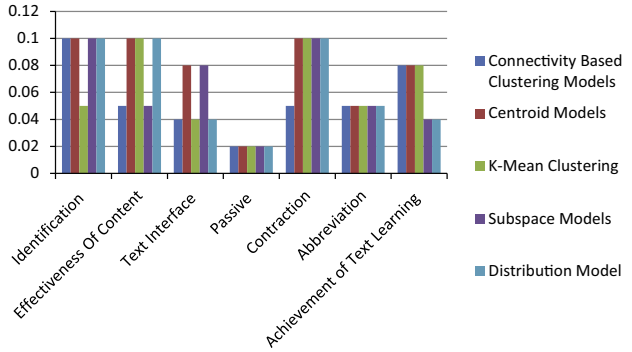

Fig. 6. Miscellaneous criteria.

• Passive	Extract meaningless patterns
• Summarization	Migrate or reduce text documents from numerous documents into a little set of paragraph or words
• Contraction	A shortened version of the words or syllable whether in spoken or written forms that created by distraction of internal letters
• Abbreviation	A group of letters taken from a phrase of a word
• Achievement of Text Learning	A form of learning through linear classifier and machine learning

Table 1 depicts the proposed set of criteria to evaluate text mining techniques. The table is divided broadly into two parts: general criteria and specific criteria along with their overall weights with break-up of each criterion that is used for text mining in the two main groups. The table shows that specific criteria have much more weights than the general criteria. General criteria have fewer options and carry an overall weight of only 16%. The options are only five in number to include usability (4%), comprehensiveness (3%), flexibility (4%), complexity (3%) and Graphic User Interface GUI (2%). Specific criteria on the other hand are further subdivided into 19 sub-criterions carrying an overall weight of 86%. The weights in specific sub-criterions vary from a weight of 2% to 7%. The heaviest of them all is Extraction (7%) followed by Indexation (6%) and Summarization (6%). 5% weight has the maximum frequency and is common to 7 unique sub-criterions from the Specifics Group, being specific goal achievement, Aggregation, Clustering, Identification, Effectiveness of Content, Contraction and Abbreviation. Four sub-criterions carry the weight of 4% each, to include, Satisfy Business Goals, Support Compliance, Accumulation, and Achievement of text Learning. Similarly weight of 3% is common to 3 sub-criterions which are, Support KPI, Support Individual Contents, Text Interface and Passive. Support MLI (Multi Link Interface) is the only sub-criterion in the list of specific criteria to carry the minimum weight of 2%.

In Fig. 2, we have shown the significance of the General criteria on the studied Clustering models. It is clear that the distribution model is the most comprehensive one, while the connectivity based algorithm has a good advantage in usability, flexibility, complexity and GUI over the other models.

In Figs. 3–6, we have shown the significance of four different groups of specific criteria, namely, the achievement criteria group, the support criteria group, the data analysis criteria group, and the miscellaneous criteria group, respectively.

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

4. Discussion and conclusion

In Text mining, the use of machine learning and data mining approaches has developed different tools and techniques that have been well studied and examined in the literature. Text Mining has been applied on wide areas of research including eLearning, social networking, bio informatics, pattern matching, user experience, intelligent tutoring systems, etc.

Most text mining techniques are based on different approaches such as clustering, classification, relationship Mining and Pattern Matching (Kotsiantis, 2007). Those approaches have been used in locating, identifying and extracting relevant information and data from unstructured and unorganized textual resources. Mining techniques have been described along with different algorithms and classifications to provide a framework and design. Seven different approaches have been identified including classification, clustering, regression analysis, association rule learning, anomaly detection techniques, summarization and other supervised learning approaches. All those approaches have individual importance in designing and implementing effective data warehouses that would be used for different purposes. Mainly data warehouses are used by scholars, researchers, development centers, etc.

A proposed criterion has been introduced for comparison between different text mining techniques. The criterion is based on two main criteria categories:

- *General:* General Criteria on Usability, Comprehensiveness and Flexibility. Analysis weights have been assigned to each criterion.
- *Specific:* Those criteria divided into different sub-criteria, such as, Graphic User Interface, goals of research, satisfaction level, and KPI and Support Compliance.

Acknowledgement

This work was supported by College of Computer and Information Sciences, King Saud University. The authors are grateful for this support.

References

Achtert, E., Böhm, C., Kriegel, H. P., Kröger, P., Müller-Gorman, I., Zimek. Finding hierarchies of subspace clusters. *LNCS: Knowledge discovery in databases: PKDD*. Lecture Notes in Computer Science (Vol. 4213, pp. 446–453).

Ahmad, R., & Khanum, A. (2010). Document topic generation in text mining by using cluster analysis with EROCK. *International Journal of Computer Science & Security (IJCSS), 4*(2).

AlSumait, L., & Domeniconi, C. (2007). Text clustering with local semantic kernels.

Bhushan, J., Pushkar, W., Shivaji, K., & Nikhil, K. (2014). Searching research papers using clustering and text mining. *International Journal of Emerging Technology and Advanced Engineering, 4*(4).

Burley, D. (2010). Information visualization as a knowledge integration tool. *International Journal of Knowledge Management Practice, 11*(4).

Caropreso, M. F., Matwin, S., & Sebastiani, F. (2009). Statistical phrases in automated text categorization.

Dhawan, S., Singh, K., & Khanchi, V. (2004). A framework for polarity classification and emotion mining from text. *International Journal of Engineering and Computer Science, 3*(8).

Dhawan, S., Singh, K., & Khanchi, V. (2014). Emotion mining techniques in social networking sites. *International Journal of Information & Computation Technology, 4*(12), 1145–1153.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., et al. (2007). Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *Proceedings of the 16th ACM conference on information and knowledge management (CIKM)*, November (pp. 213–222).

Gharehchopogh, F. S., & Abbasi Khalifehlou, Z. (2011). Study on information extraction methods from text mining and natural language processing perspectives. *AWER Procedia information technology & computer science, 2nd world conference on information technology*.

Gupta, V. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1) (August).

Howland, P., & Park, H. (2007). Cluster-preserving dimension reduction methods for document classification.

Kiyavitskaya, N., Zeni, N., Mich, L., Cordy, J., & Mylopoulos, J. (2006). Text mining through semi automatic semantic annotation. In *Proceedings of Practical Aspects of Knowledge Management (PAKM'06)*, Vol. 4333 of LNCS, (pp. 143–154).

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*, 249–268.

Krause, A., Leskovec, J., & Guestrin, C. (2006). Data association for topic intensity tracking. In *International conference on machine learning (ICML)* (pp. 497–504).

Kriegel, H. P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *Transactions on Knowledge Discovery from Data (New York, NY: ACM), 3*(1), 1–58.

Lehmam, A. (2010). Essential summarizer: Innovative automatic text summarization software in twenty languages – ACM digital library. *Published in proceeding RIAO'10 adaptively, personalization and fusion of heterogeneous information*, CID Paris, France.

Lincy Liptha, R., Raja, K., & Tholkappia Arasu, G. (2010). Text clustering using concept-based mining model. *International Journal of Electronics and Computer Science Engineering* (ISSN: 2277-1956).

Navaneethakumar, V. M., & Chandrasekar, C. (2012). A consistent web documents based text clustering using concept based mining model. *IJCSI International Journal of Computer Science Issues, 9*(4) (No. 1).

Patel, R., & Sharma, G. (2014). A survey on text mining techniques. *International Journal of Engineering and Computer Science, 3*(5).

Rath, S. K., Jena, M. K., Nayak, T., & Bisoyee, B. (2011). Data mining: A healthy tool for your information retrieval and text mining. *International Journal of Computer Science and Information Technologies, 2*.

Scott, S., & Matwin, S. (2011). Feature engineering for text classification.

Senellart, P., & Blondel, V. D. (2008). *Automatic discovery of similar words, survey of text mining II: Clustering, classification and retrieval*. Springer-Verlag (pp. 25–44).

Shelke, N. M. (2014). Approaches of emotion detection from text. *International Journal of Computer Science and Information Technology Research, 2*(2), 123–128.

Silwattananusarn, T., & Tuamsuk, K. (2012). Data mining and its applications for knowledge management: A literature review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process, 2*(5).

Tan, H., & Lambrix, P. (2009). Selecting an ontology for biomedical text mining. *Human language technology conference*. BioNLP Workshop Association for Computational Linguistics.

Vashishta, S., & Jain, Y. K. (2011). Efficient retrieval of text for biomedical domain using data mining algorithm. *International Journal of Advanced Computer Science and Applications, 2*(4).

Velmurugan, T., & Santhanam, T. (2010). Performance evaluation of K-means and fuzzy C-means clustering algorithms for statistical distributions of input data points. *European Journal of Scientific Research, 46*(3), 320–330 (ISSN 1450-216X).

Wang, X., Wang, J. T. L., Lin, K. I., Shasha, D., Shapiro, B. A., & Zhang, K. (1999). An index structure for data mining and clustering.

Wu, S. T., Li, Y., & Xu, Y. (2006). Deploying approaches for pattern refinement in text mining. In *Proceedings of the sixth international conference on data mining*.

Yassine, M., & Hajj, H. (2010). A framework for emotion mining from text in online social networks. In *IEEE international conference on data mining workshops* (pp. 1136–1142). Sydney, NSW: IEEE publications.

Zhai, C., Velivelli, A., & Yu, B. (2004). A cross collection mixture model for comparative text mining. *Proceedings of ACM KDD 2004 (KDD'04)* (pp. 743–748).

Zhou, Y., Zhang, Y., Vonortas, N., & Williams, J. (2012). A text mining model for strategic alliance discovery.

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