December 2020

Next Generation Data Analytics: Text Mining in Library Practice and Research

Muhammad Arshad
National Library and Resource Centre, Islamabad, Pakistan, marshadnlrc@gmail.com

Amjid Khan
Allama Iqbal Open University, Islamabad, Pakistan, amjid.khan@aiou.edu.pk

Pervaiz Ahmed
Allama Iqbal Open University, Islamabad,Pakistan, pervaiz@aiou.edu.pk

Nadia Abbas Shah
Polyclinic Hospital, Islamabad, Pakistan, nadiaabbas550@gmail.com

Next Generation Data Analytics: Text Mining in Library Practice and Research

Muhammad Arshad
National Library and Resource Centre, Islamabad, Pakistan
Email: marshadnlrc@gmail.com

Amjid Khan
Allama Iqbal Open University, Islamabad, Pakistan
Email: amjid.khan@aiou.edu.pk (corresponding author)

Pervaiz Ahmad
Allama Iqbal Open University, Islamabad, Pakistan
Email: pervaiz@aiou.edu.pk

Nadia Abbas Shah
Polyclinic Hospital, Islamabad, Pakistan
Email: nadiaabbas550@gmail.com

Abstract

Text mining is the process and technique used for searching, retrieving and extracting high quality, useful and purposeful information from the ocean of unstructured and unclassified data and information in the form of text written in natural language. It is also referred to as text engineering, text data mining or text analytics. Text mining is an emerging field of research. The purpose of this paper is to review text mining in general and with special context to library and information science. Text mining involves artificial intelligence. Hence, this system is very efficient in its work as compared to human capabilities in terms of time constraints and the counts of frequency which involves accuracy. This field has many prospects to offer and in the context of the library and information science. This technique can be very helpful in managing ever-growing and proliferating information in every field of knowledge. Moreover, this technique can also be used for research in the field of LIS. Researchers and librarians may find it useful to put efforts in this field for being better adapted to future. This review is useful for the motivation of LIS researchers to involve in text mining aimed at contributing new knowledge to the improvement of library profession and services. Since text mining is an emerging field in research, less literature is available particularly related to library and information science.

Keywords: Data analytics, Text mining, Data extraction, Data mining, Knowledge management, Library practice, Library research.

1. Introduction

The majority of data which we encounter daily is in form of unstructured text, i.e. in form of books, emails, newspapers and web pages. The term unstructured means that this data lack a structured format, as the text has in a spreadsheet or a relational database. Text mining deals with such type of data convert it into a machine-readable form or in other words, a structured format and then draw knowledge from it. This technique uses artificial intelligence technologies such as machine learning and natural language processing. Furthermore, this technique has its roots in several other disciplines too. Text mining technique helps manage the ever-growing world of knowledge. Librarians have a dual role in this scenario, one is to use this technique to improve library services as an efficient information retrieval system, classification and number of other applications. Librarians can strengthen their role by facilitating researchers who are willing to endeavor the text mining task. Moreover, Text mining is a developing field of research in Library and Information Science (LIS) all over the world and in Pakistan too.

2. What is Text Mining?

Text mining is still going through a self-definition phase, so it lacks a general approach to its techniques, applications and prospects. That's why different researchers define this term from different perspectives. Miner et al. (2012) explain "text mining and text analytics are broad umbrella terms describing a range of technologies for analyzing and processing semi-structured and unstructured text data. The unifying theme behind each of these technologies is the need to 'turn text into numbers' so powerful algorithms can be applied to large document databases" (p. 30). According to Feldman and Sanger (2007) "text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools" (p. 1). Jo (2019) defines text mining as "the process of extracting the implicit knowledge from textual data" (p. 3). Similarly, Kwartler (2017) mentions "text mining is the process of distilling actionable insights from the text" (p. 1). Liddy (2000) defines "text mining is the process of analyzing naturally occurring text to discover and capture semantic information for insertion and storage in what I'll call a Knowledge Organization Structure (KOS) with the ultimate goal of enabling knowledge discovery via either textual or visual access for use in a wide range of significant applications" (p. 13).

3. History of Text Mining

Miner et al. (2012) reveal that there are at least three reasons due to which one should know the history of text mining. One is to see the developmental paths of text mining techniques, second is to see how text mining techniques can be expanded and improved in the future, thirdly to learn from the past and avoid repeating the same mistakes. Text mining has its roots in three processes, information retrieval, extraction and summarization. Miner et al. argue that as we can't understand the history of computers without fully understanding the work of Charles Babbage on difference engine, in the same way, we can't understand the text mining process without understanding its roots. Modern text mining has its form due to other technology and applications, which developments are due to the increasing number of textual information in the world.

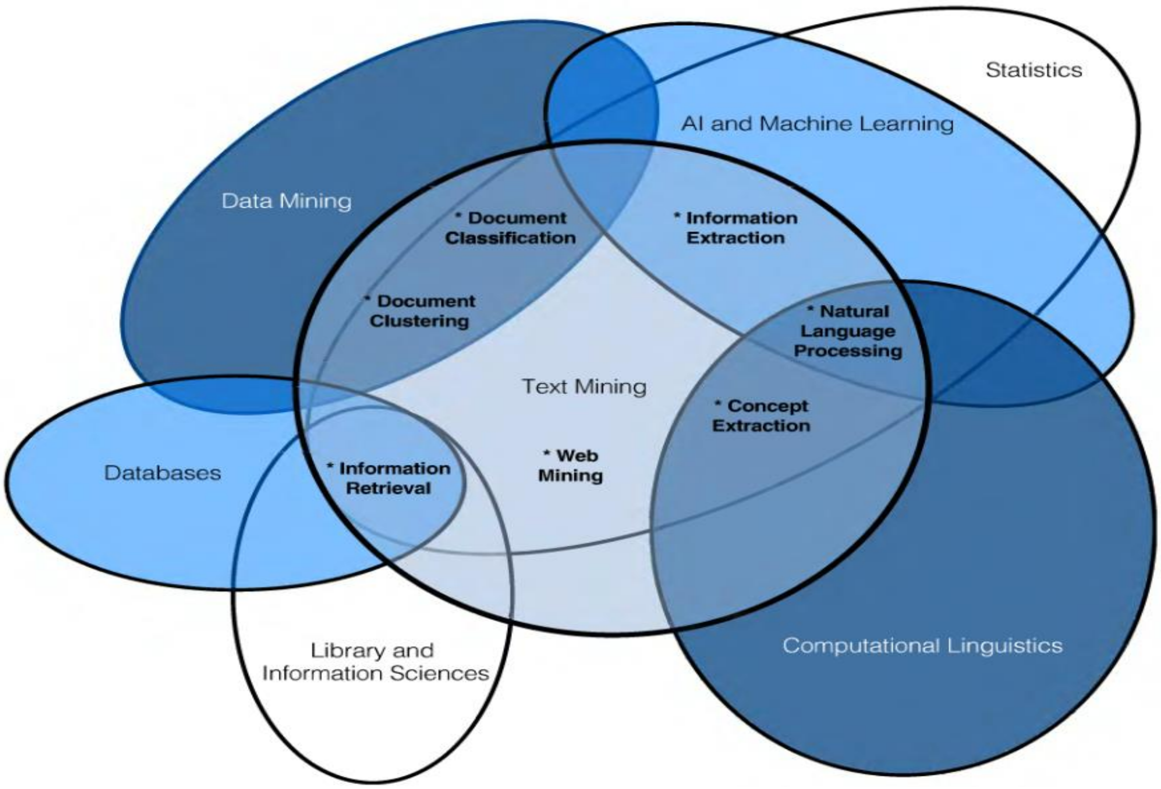

Figure 1. Text mining and related disciplines (adopted from Miner et al., 2012, p. 31)

According to Miner et al. (2012), there are three fields in which the methods to access textual information developed: Library science, information science and natural language processing.

- ***Library science***: Library catalogue is the earliest example of text summarization. The first catalogue was developed by Thomas Hyde in 1864 for the Bodleian Library at the University of Oxford.
- ***Information science***: Before the advent of the computer, the information has to be catalogued and indexed in form of catalogue card and the library users have to go through the laborious task of finding the required information.
- ***Natural language processing:*** This term is often used as a synonym for computational linguistics. The major development in natural language processing is the machine learning technology to create parsing algorithms, which split words into tokens; and stemming algorithms, which reduce words to stems. Clustering is an automated process which clusters the documents on the base of similarities.

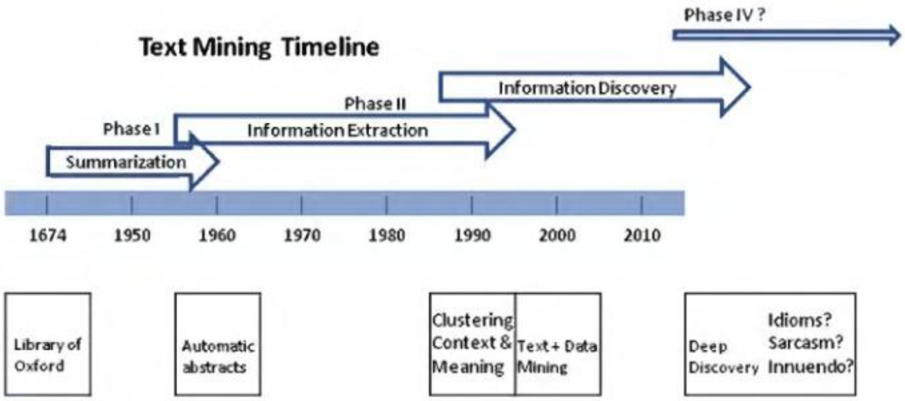

Figure 2. Phases in the development of text mining (adopted from Miner et al., 2012, p. 12)

4. Objectives of Text Mining

Experts (e.g. Miner et al., 2012; Ying, 2012) divide text mining into following seven practice areas:

- ***Storage and information retrieval***: This include the storage and retrieval of information from documents, like in search engines and keyword search.
- ***Document clustering***: Document clustering technique is used to group and categorize terms, snippets, paragraphs and documents.
- ***Document classification***: Classification methods based on models of labelled examples are used to group and categorize terms, snippets, paragraphs, and documents.
- ***Web mining***: It is done by using the internet with a special focus on the scale and interconnectedness of the web.
- ***Information extraction***: This is used to convert unstructured and semi-structured text into structured text to identify and extract relevant facts and relationships.
- ***Natural language processing***: It is the low-level computer processing to interact with human language.
- ***Concept extraction***: It is the process of grouping words and phrases into semantically related groups.

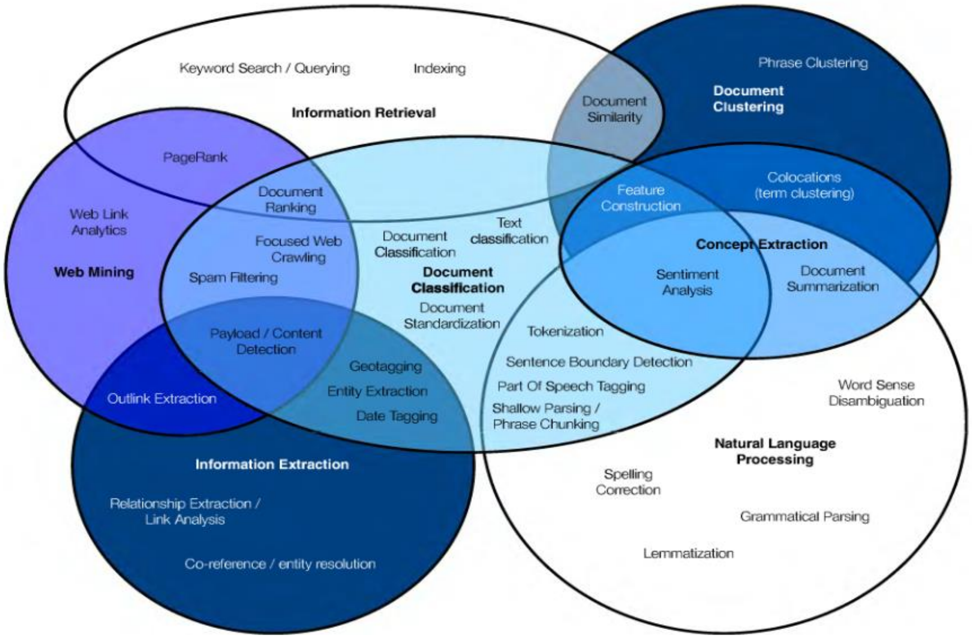

Figure 3. Seven practice areas of text mining (adopted from Miner et al., 2012, p. 40)

5. Methodology of Text Mining

Miner et al. (2012) define methodology as “a documented and somewhat standardized process for executing and managing complex projects that include many interrelated tasks by the use of a variety of methods, tools and techniques” (p. 73). They further explain that text mining is a relatively new technique and unstandardized while data mining is relatively mature. There is no commonly accepted methodology for text mining. They introduced the following phases of this methodology.

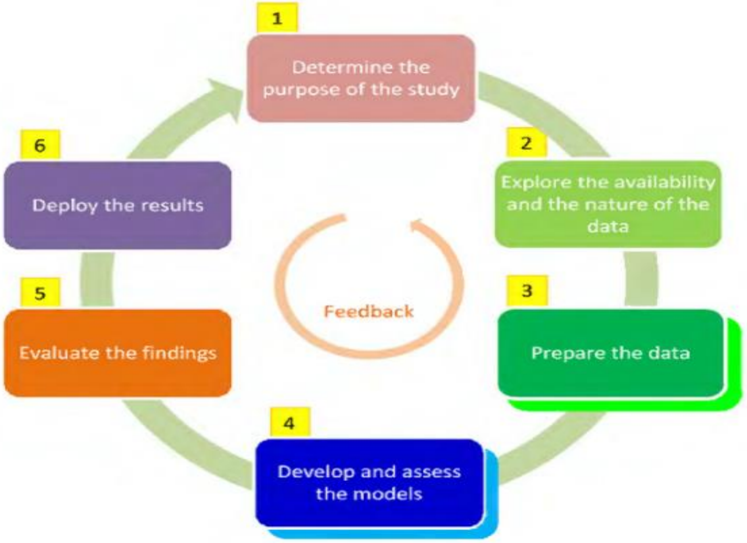
1
Determine the purpose of the study
2
Explore the availability and the nature of the data
3
Prepare the data
4
Develop and assess the models
5
Evaluate the findings
6
Deploy the results
Feedback

Figure 4. The methodology of text mining (adopted from Miner et al., 2012, p. 75)

- ***Determine the purpose of the study***: Text mining starts with determining the purpose of the study like any other project. It is necessary to understand the case and define the aims. This can be achieved with the thorough study of the problem due to which the study initiated and it can involve discussing the issue with domain experts for deep understanding.
- ***Explore the availability and nature of the data***: After formulating the objectives, the next step is to check the availability of data for the process. Some of the tasks involved can be the identification of the sources to get textual data, assessment of data accessibility and its usability, collection and then exploring its richness, the final assessment about the quality and quantity of the data. After having positive outcomes, the next process is to collect the data and integrate it for further use.
- ***Prepare the data and develop and assess the models***: This phase involves three basic processes; establishment of the corpus, pre-processing the data and knowledge extraction.

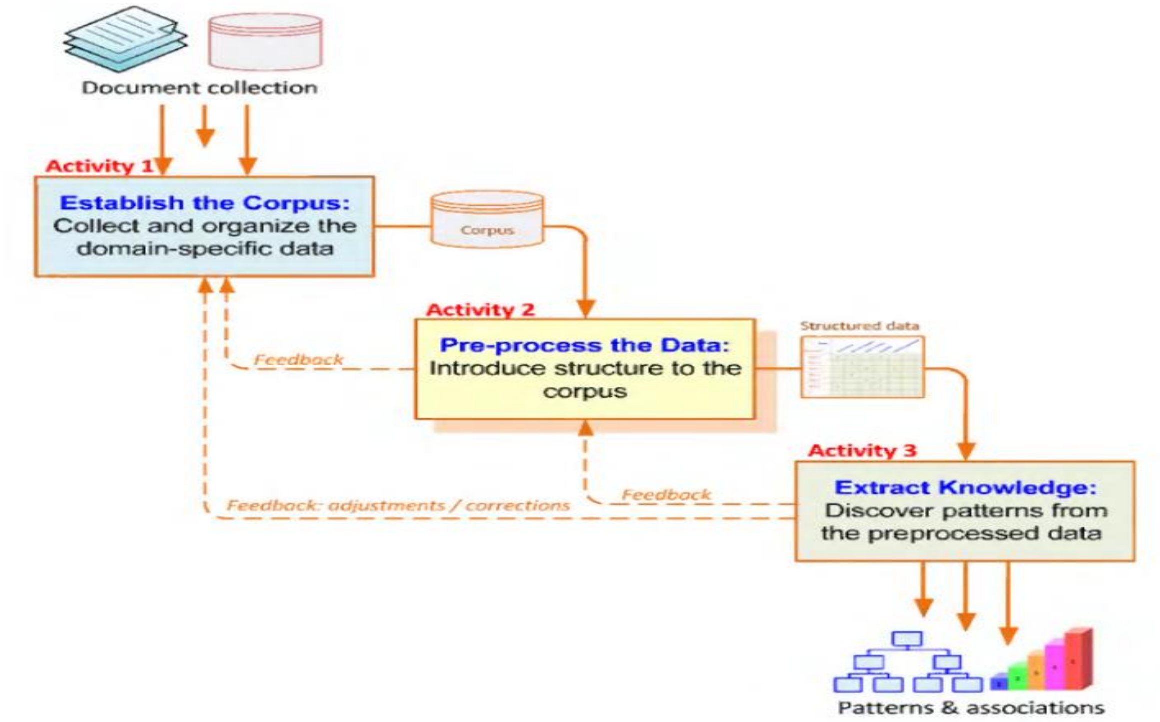

Figure 5. Processing of data (adopted from Miner et al., 2012, p. 79)

- ***Evaluate the results***: After developing the models and assessing for accuracy and quality, it is necessary to verify and validate all the activities involved. It is the process of repeating the whole process and checking its validity, only then we can move to the deployment of results. This stage helps to eliminate the chance of error that can lead to the faulty decision-making process.
- ***Deploy the results:*** Once the process passes the assessment phase, it is ready for deployment which means to be applied. This process can be as simple as writing and report of findings which help in decision making or as complex as developing a new intelligence system.

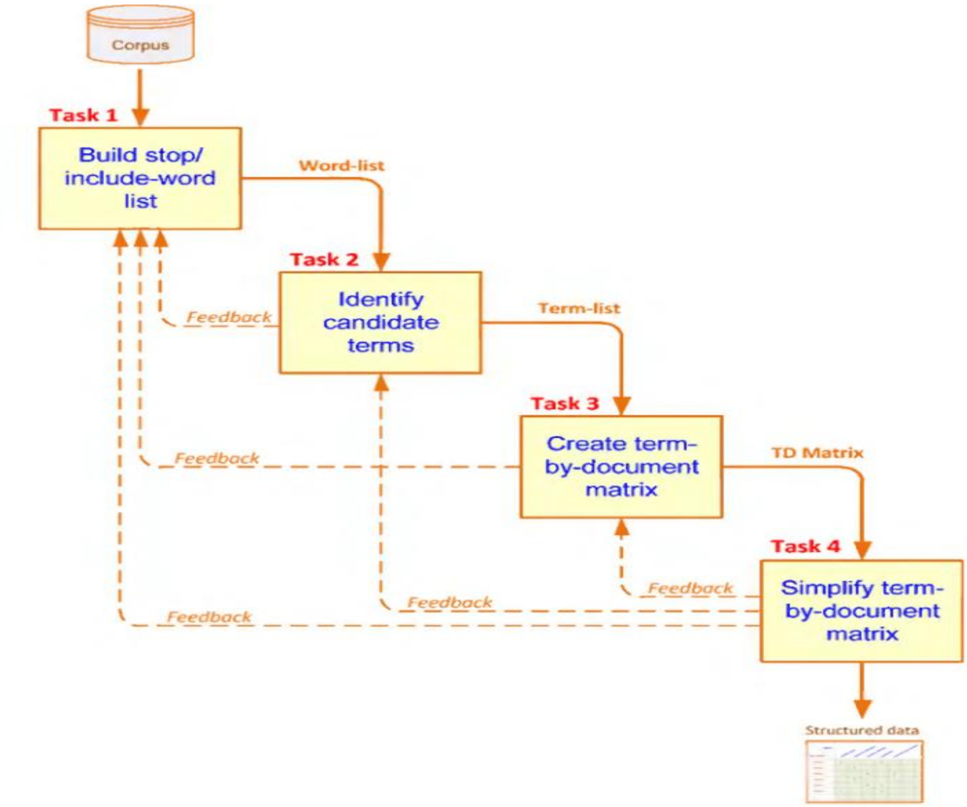

Figure 6. Preprocessing of data (adopted from Miner et al., 2012, p. 81)

6. Data Visualization Techniques

Kwartler (2017) gives word clouds' visualization technique for presenting data. Word cloud is another way of data visualization. In this method, the frequency is shown by the size of the word, the most frequent words appear in big sizes. Another approach is to use colors or groupings.

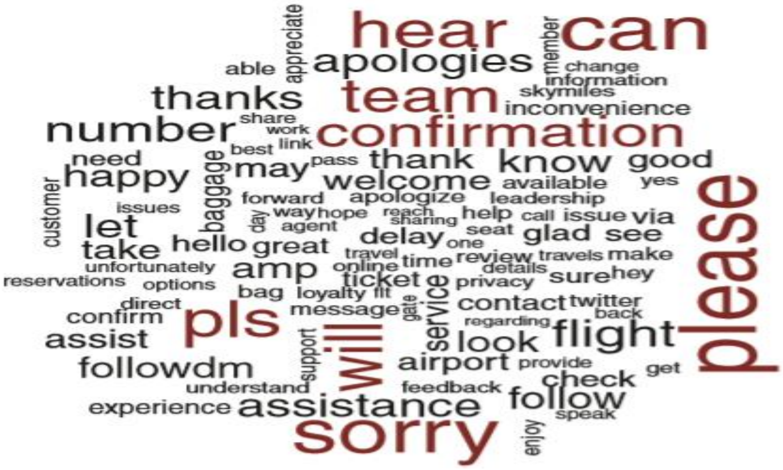

Figure 7. Tag plot (adopted from Kwartler, 2017, p. 76)

7. Applications of Text Mining

There are five basic types of analytical text mining applications that are applied to text analysis issues (Talib, Hanif, Ayesha, & Fatima, 2016).

- ***Extracting meaning*****:** This involves the extraction of meaning from unstructured data. This includes the understanding of core themes and relevant messages without actually reading the text.
- ***Automatic text categorization*****:** This is an excellent way of downstream processing by automatically classifying text.
- ***Improving predictive accuracy*****:** In predictive modelling and unsupervised modelling, it is an efficient method to achieve accuracy by combing unstructured data with structured numeric information.
- ***Identifying specific document*****:** This is an important task in information retrieval to efficiently extract similar or relevant documents on a particular topic.
- ***Extracting specific information*****:** Extracting specific information like names, geographic regions etc. is an efficient method to provide data to decision-makers.

8. Constraints and Enablers in Text Mining

Miner et al. (2012) indicate various constraints to the text mining process; privacy/access issues, software limitations, hardware limitations, linguistic challenges. While text mining has several enablers too, which are natural language processing methods, software tools, domain expertise, and fast computers.

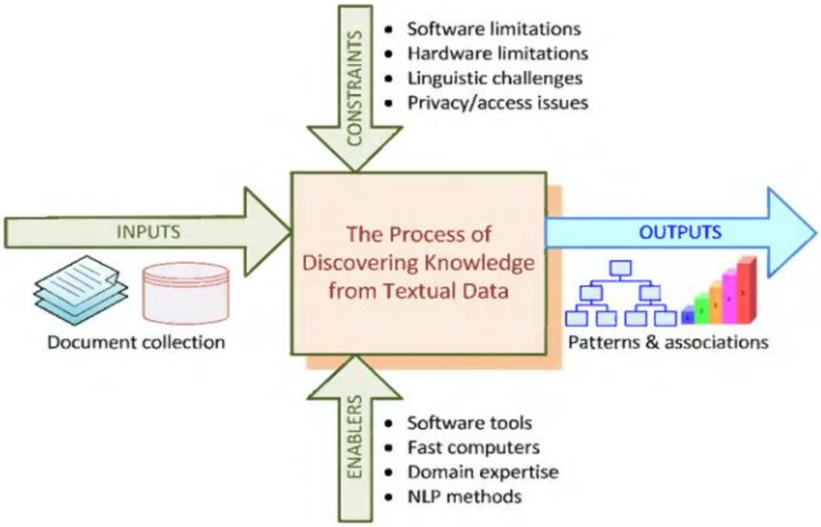

Figure 8. Constraints and enablers of text mining (adopted from Miner et al., 2012, p. 78)

9. Methods and Models Used in Text Mining

Traditionally there are so many techniques developed to solve the problem of text mining but the relevant information retrieval according to user's requirement. According to the information retrieval methods/techniques, basically there are following four methods used in text mining (Gaikwad, Chaugule, & Patil, 2014).

- ***Term-Based Method (TBM):*** This method is based on the terms to analyze the entire set of documents.
- ***Phrase-Based Method (PBM):*** This method is based on the phrases which are less clear to more meaningful and distinctive to analyze document.

- ***Concept-Based Method (CBM):*** This method is based on the sentence and document level to analyze the concepts in a document.
- ***Pattern Taxonomy Method (PTM):*** This method is based on the patterns to analyze the document using data mining techniques such as "association rule mining, frequent item set mining, sequential pattern mining, and closed pattern mining" (p. 2).

10. Text Mining Techniques

There are five basic techniques used in text mining system as follows (Gaikwad, Chaugule, & Patil, 2014).

- ***Information Extraction***: This technique uses unstructured text to identify key phrases and relationships within text comprising "tokenization, identification of named entities, sentence segmentation, and part-of-speech assignment" (p. 2).
- ***Categorization:*** Categorization is a supervised learning method because it is based on input-output examples to classify new documents. The typical text categorization process consists of pre-processing, indexing, dimensionally reduction, and classification.
- ***Clustering****:* This technique is used to find similar content in a group of several documents called clusters.
- ***Visualization****:* This technique uses text flags (density colors) to discover document category or relevant information and it can be applied to individual or group of documents.
- ***Summarization****:* This technique is used to reduce lengthy documents to a reduced or summarized form while retaining the key points and common connotation.

11. Text Mining in Library Practice

The major applications of text mining in library services are in classification, keyword extraction, named entity recognition, topic modelling and clustering which can make libraries work efficiently for their information management, information retrieval and decision making. Zhang and Gu (2011) explain that as a high percentage of the world's knowledge is in unstructured form, so it is likely to use technology as text mining to extract information from these sources efficiently. Their work is to do named entity recognition to determine proper names, their variations and classes. Cong (2017) suggests that due to the continuous growth of book price and less available funds to purchase books, it is necessary to spend fund on buying needed books. For this purpose, Chinese word segmentation and Chinese library classification are used to find out readers' reading preferences news, management and finance discipline's borrowing data. Results showed that readers prefer specific books of subcategories and themes.

Al-Daihani and Abrahams (2016) collected dataset from the Twitter accounts of ten academic libraries. The single word which has the most frequency was 'open', the bigram word which showed the highest frequency was 'special collections' and the trigram word was 'save the data'. The most common category was 'resources'. The findings of this study highlight the importance of analyzing the aggregate data of academic libraries at social accounts to help in decision making and improve planning for services to patrons.

Okerson (2013) suggests two main areas of activity for librarians to meet text mining needs in research institutions. One is developing a licensed language, many librarians are active in developing license language and permission for text mining, which is necessary for the process. The second opportunity is supporting researchers. Librarians can help researchers to guide where text mining can be more effective. She further explains that knowing our communities, make them understand that what data mining can do for them and providing them with the tools is surely a librarian's job.

Similarly, Higgins (2014) points out some reasons for librarians to be involved in text mining techniques. For example, to support the latest forms of research being employed in humanities. Text mining has the potential to draw librarians and researchers closer to their mutual interests. Librarians and humanities scholars are alike in concerning with the value and implications of textual data. Text mining initiatives can support in enriching the digital collections of the library. Text mining can help in generating highly descriptive metadata which is the main area of concern for librarians. Text mining procedures help in gathering specific named entities. Librarians have a chance to generate descriptions for data with broader thematic content. Topic modelling can be used for both collocating and distinction of resources.

Anderson and Craiglow (2017) explain the stages of text mining in academic librarianship. It is a detail view of how librarians can help facilitate text mining research by assisting faculty members. Librarians can help in identifying sources, licensing data, extracting data, data munging, devising models, curation and preservation, and publication.

12. Text Mining in LIS Research

Library and information science are taking advantage of using text mining technology to enhance research in their field and to generate thoughtful insights, as most of the other fields are doing like and the emerging field of digital humanities too. Nagarkar and Kumbhar (2015) did an analytical study of research published in Information and Library science from 1999 to 2013. They analyzed the chronological growth of research literature about text mining, major countries, institutions, departments and individuals who are actively playing part in text mining research. They used Pajek and VoSviewer for this purpose.

Timakum, Kim, and Song (2020) analyzed the knowledge structure of library science by using full-text journal articles. They applied text mining techniques co-word analysis, text summarization and topic modelling. They use text mining visual techniques to map their findings and find that digital information management is the major developed area of research. In Korea, Lee, Moon, and Kim (2007) applied text mining to examine the intellectual structure of records management and archival science. In another study, Lee, H. Kim, and P. J. Kim (2010) applied domain analysis with text mining to study research trends in digital library research. In China, Ying (2012) did a study on mining bibliographic records by using a text mining software, SATI. He did cluster analysis and multidimensional scaling analysis to draw a knowledge map and strategic diagram of the results.

13. Conclusion

Text mining is an emerging field of research but it has the potential to generate insights in any area of study. This technique is exploring itself by including a vast number of applications in all fields. It offers the same prospects to libraries to enhance their services and apply this technique in the area of Library and Information research. Moreover, it demands from librarians to fulfil their additional duties of assisting other fields in their research endeavors and actively participate to make this process applies to all fields. It is advised to researchers of library and information science in general and specifically to Pakistani researchers to adopt this technique for making grounds to better decision making in this area of study at all conceptual, theoretical and practical levels.

14. References

Al-Daihani, S. H., & Abrahams, A. (2016). A text mining analysis of academic libraries' tweets. *The Journal of Academic Librarianship*, 42(2), 135-143.

Anderson, C. B., & Craiglow, H. A. (2017). Text mining in business libraries. *Journal of Business & Finance Librarianship*, 22(2), 149-165.

Cong, D. (2017). *Application of text mining in library book procurement*. Retrieved from https://www.researchgate.net/publication/314783043_Application_of_text_mining_in_library_book_procurement

Feldman, R. & Sanger, J. (2007). *The text mining hand book: Advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, 85(17), 42-45.

Higgins, D. (2014). Reading and non-reading: Text mining in critical practice. In K. J. Varnum (Ed.), *The top technologies every librarian needs to know: A LITA guide* (pp. 85-99). Chicago, IL: ALA Techsource.

Jo, T. (2019). *Text mining: Concepts, implementation and big data challenge*. Seoul: Springer International Publishing.

Kwartler, T. (2017). *Text mining in practice with R*. New Jersey: John Wiley & Sons.

Lee, J., Moon, J., & Kim, H. (2007). Examining the intellectual structure of records management and archival science in Korea with text mining. *Journal of the Korean Society for Library and Information Science*, 41(1), 345-372.

Lee, J. Y., Kim, H., & Kim, P. J. (2010). Domain analysis with text mining: Analysis of digital library research trends using profiling methods. *Journal of Information Science*, 36, 144-161.

Liddy, E. D. (2000). Text mining. *Bulletin of the American Society for Information Science*, 13-14. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bult.184

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Amsterdam: Academic Press.

Nagarkar, S., & Kumbhar, R. (2015). Text mining: An analysis of research published under the subject category 'information science and library science' in web of science database during 1999-2013. *Library Review*, 64(3), 248-262.

Okerson, A. (2013). *Text and data mining: A librarian overview*. Retrieved from https://www.fosteropenscience.eu/sites/default/files/original/3628.pdf

Talib. R., Hanif. M. K., Ayesha. S., & Fatima, F. (2016). Text mining: Techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 11(7), 414-418.

Timakum, T., Kim, G., & Song, M. (2020). A data driven analysis of the knowledge structure of library science with full text journal articles. *Journal of Librarianship and Information Science*, 52(2), 345-365.

Ying, L. Q. Y. (2012). A study on mining bibliographic records by designed software-SATI: Case study on library and information science. *Journal of Information Resource Management*, 1, 35-67.

Zhang, Y., & Gu, H. (2011). *Text mining with application to academic libraries*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-22694-6_28