Text Analysis Frame Work for Mining Knowledge from Text Data

Sadhana J. Kamatkar[a]

[a]*Central Computing Facility, University Environment, 109, First Floor, Fort Campus, Mumbai, 400 032, INDIA. Email:* [a]*sjkamatkar@mu.ac.in;* [b]*sadhanajk@gmail.com*

Abstract: With huge growth in textual data Text Mining, Text Analysis have gained research focus. This paper addresses the differences between Data Mining and Text Mining. It also discusses the Text Mining Applications and Relationship between text mining information retrieval and text mining. We propose the new approach adopted to analyze the text data consisting of text, barcode and images. We demonstrate the application of new approach in real life situation adopted in University of Mumbai.

Keywords: Text Mining, Data Mining, Knowledge Discovery, Text Analysis.

1. INTRODUCTION

Text mining is a relatively new area of computer science, and its use has grown as the unstructured data available continues to increase exponentially in both relevance and quantity. Text Analysis applications scan a set of documents written in a natural language. These applications model the document set for predictive classification purposes or populate a data base or search index with the information extracted. Text Analytics is the process of converting unstructured text data into meaningful data for analysis, to measure customer opinions, product reviews, feedback, to provide search facility, sentimental analysis (opinion mining) and entity modeling to support fact based decision making. Text analysis uses many linguistic, statistical, and machine learning techniques. Text Analytics involves information retrieval from unstructured data and the process of structuring the input text to derive patters and trends and evaluating and interpreting the output data. It also involves lexical analysis, categorization, clustering, pattern recognition, tagging, annotation, information extraction, link and association analysis, visualization, and predictive analytics. Text Analytics determines key words, topics, category, semantics, tags from the millions of text data available in an organization in different files and formats. The term Text Analytics is roughly synonymous with text mining.

2. LITERATURE REVIEW

The information retrieval (IR) process of data mining requires extraction of valid patterns and relationships in very large data sets automatically [6] [7]. It is usually portrayed like a “voyage into the unknown”, and hence requires the use of techniques from AI and statistics, such as machine learning, pattern recognition, classification, and visualization [8] [9]. Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analysing text to extract information that is useful for particular purposes[41]. The phrase “text mining” is generally used to denote any system that analyses large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information [42].’ It is estimated that 80% of the world’s

on-line content is based on text' [43]. Text mining 'performs various searching functions, linguistic analysis and categorizations'. Search engines focus on text search, especially directed at 'text-based web content' [43]. Data mining is a step in the process of knowledge discovery from data (KDD). KDD concerns the acquisition of new, important, valid and useful knowledge from data. Berson and Smith [44] maintain that: 'In the case of large databases sometimes users are asking the impossible: "Tell me something I didn't know but would like to know."

Data mining is a proactive process that automatically searches data for new relationships and anomalies on which to base business decisions in order to gain competitive advantage [45]. Although data mining might always require some interaction between the investigator and the data-mining tool, it may be considered as an automatic process because 'data-mining tools automatically search the data for anomalies and possible relationships, thereby identifying problems that have not yet been identified by the end user', while mere data analysis 'relies on the end users to define the problem, select the data, and initiate the appropriate data analyses to generate the information that helps model and solve problems they uncovered' [45]. Text analysis *is about deriving high-quality structured data from unstructured text.* Another name for text analytics is *text mining*. A good reason for using text analytics might be to extract additional data about customers from unstructured data sources to enrich customer master data, to produce new customer insight or to determine sentiment about products and services. [46]. This paper aims to address the gap in literature by proposing an intelligent Text Analysis framework for mining knowledge from Text data.

3. COMPARISON BETWEEN DATA MINING AND TEXT MINING

Text Mining and Data Mining are becoming increasingly widespread as companies try to tackle their unstructured information, or big data, for business value. While the goal is often the same—exploiting information for knowledge discovery—these techniques vary significantly when it comes to data complexity, deployment time and application. Here we will take a deeper look at how they are applied in real-world projects.

First we will see their definitions: **Data mining is a process based on algorithms to analyze and extract useful information from data**. It can be used to automatically discover hidden patterns and relationships in data, and to predict outcomes from large data sets.

Text mining is the set of processes required to turn unstructured text documents or resources into valuable structured information. This requires both sophisticated linguistic and statistical techniques able to analyze unstructured text formats and techniques that combine each document with actionable metadata, which can be considered a sort of anchor in structuring this type of data. Once content has been annotated, it can automatically be classified, routed, summarized, visualized through link mapping and, most importantly, it becomes easier to search.

While the end goals are quite similar—use information to fuel decision making, reduce costs and increase revenue for business activities like issues detection, analysis and correction, or knowledge discovery, forecasting and strategic planning—we need to look closely at text mining vs. data mining to understand how they are different.

A. Data Mining vs Text Mining: Unstructured Versus Structured Data

(i) Data mining systems essentially analyze figures that may be described as homogeneous and universal. They extract, transform and load data into a data warehouse [5]. Business analysts use data mining software applications to present analyzed data in easily understandable forms, such as graphs. Currencies, dates, names, might have to be managed, but they are easy to link to data and do not

require any deep understanding of their context.

(ii) Text mining tools have to face major technical challenges such as heterogeneous document formats (text documents, emails, social media posts, verbatim text, etc.), as well as multilingual texts and abbreviations and slang typical of SMS language.

B. Data Mining vs Text Mining: Deployment Time

(i) Data mining is focused on data-dependent activities such as accounting, purchasing, supply chain, CRM, etc. The required data is easy to access and homogeneous. Once algorithms are defined, the solution can be quickly deployed.

(ii) The complexity of the data processed make text mining projects longer to deploy. Text mining counts several intermediary linguistic stages of analysis before it can enrich content (language guessing, tokenization, segmentation, morpho-syntactic analysis, disambiguation, cross references, etc.). Next, relevant terms extraction and metadata association steps tackle structuring the unstructured content to nurture domain-specific applications [41]. Moreover, projects may involve some heterogeneous languages, formats or domains. Finally, few companies have their own taxonomy. However, this is mandatory for starting a text mining project and it can take a few months to be developed.

C. Data Mining Vs Text Mining: Technology Perception

(i) Data mining has been considered a proven, robust and industrial technology for many decades.

(ii) Text mining was historically thought of as complex, domain-specific, language-specific, sensitive, experimental, etc. In other words, text mining was not understood well enough to have management support and therefore, was never valued as a 'must-have'. However, with the advent of digitalization, the rise of social networks and increased connectivity, companies are now more concerned about their online reputation and are looking for ways to increase loyalty with customers in a world of increasing choice. As a result, sentiment analysis (opinion mining) is the new focus of text mining. Companies have realized that information is a strategic asset made of text and that text mining is no longer a luxury, but a necessity.

Thus while text and data mining are now considered complementary techniques required for efficient business management, text mining tools are becoming even more important. A subset of text mining, Natural Language Processing (NLP) is all the more relevant when the customer is 100% involved and available to help define accurate and complete domain-specific taxonomies. In turn, this helps information extraction and metadata association become easier and more efficient. Natural language will never be as easy to handle, but text mining is now more mature and its association with data mining makes more sense, as 80% of information is made of text.

4. THE RELATIONSHIP BETWEEN TEXT MINING INFORMATION RETRIEVAL AND TEXT MINING

Two topics are closely related. But is it appropriate to say "when the information is text format, the IR and TM are equivalent"? Text mining (TM) is vast area as compared to information retrieval. Typical text mining tasks include document classification, document clustering, building ontology, sentiment analysis, document summarization, Information extraction etc. Whereas information retrieval (IR) typically deals with crawling, parsing and indexing document, retrieving documents [16].

The intuition of text mining evolves while one studies data mining in general and constraints himself to

text. Majority of the techniques learnt/applied to general data can be extended to text. Information Retrieval is more to do with search engines - concerning mostly about queries, document retrieval (though its text most of the times). This is more of hybrid topic evolved using Machine Learning (ML), Natural Language Processing (NLP) techniques quite heavily.

5. TEXT MINING APPLICATIONS

Text Mining can be used to make the large quantities of unstructured data accessible and useful, thereby generating not only value, but delivering ROI (Return On Investment) from unstructured data management. Through techniques such as categorization, entity extraction, sentiment analysis and others, text mining extracts the useful information and knowledge hidden in text content. In the business world, this translates in being able to reveal insights, patterns and trends in even large volumes of unstructured data. In fact, it's this ability to push aside all of the non-relevant material and provide answers that is leading to its rapid adoption, especially in large organizations. The following 10 text mining examples can give you an idea of how this technology is helping organizations today.

A. *Risk Management:* Insufficient risk analysis is often a leading cause of failure. This is especially true in the financial industry where adoption of Risk Management Software based on text mining technology can dramatically increase the ability to mitigate risk, enabling complete management of thousands of sources and petabytes of text documents, and providing the ability to link together information and be able to access the right information at the right time.

B. *Knowledge Management:* Not being able to find important information quickly is always a challenge when managing large volumes of text documents. Especially in Health care sector [8], organizations are challenged with a tremendous amount of information—decades of research in genomics and molecular techniques, for example, as well as volumes of clinical patient data—that could potentially be useful for their largest profit center, new product development. Here, knowledge management software based on text mining offer a clear and reliable solution for the "info-glut" problem.

C. *Cybercrime prevention:* The anonymous nature of the internet and the many communication features operated through it contribute to the increased risk of internet-based crimes. Today, text mining intelligence and anti-crime applications are making internet crime prevention easier for any enterprise and law enforcement or intelligence agencies.

D. *Customer care service:* Text mining, as well as natural language processing (NLP) are frequent applications for customer care. Today, text analytics software is frequently adopted to improve customer experience using different sources of valuable information such as surveys, trouble tickets, and customer call notes to improve the quality, effectiveness and speed in resolving problems [8]. Text analysis is used to provide a rapid, automated response to the customer, dramatically reducing their reliance on call center operators to solve problems.

E. *Fraud detection through claims investigation*: Text analytics is a tremendously effective technology in any domain where the majority of information is collected as text. Insurance companies are taking advantage of text mining technologies by combining the results of text analysis with structured data to prevent frauds and swiftly process claims.

F. *Contextual Advertising:* Digital advertising is a moderately new and growing field of application for text analytics. Here, companies such as Admantx have made text mining the core engine for contextual retargeting with great success. Compared to the traditional cookie-based approach, contextual advertising provides better accuracy, completely preserves the user's privacy.

G. *Business intelligence:* This process is used by large companies to uphold and support decision making. Here, text mining really makes the difference [8], enabling the analyst to quickly jump at the answer even when analyzing petabytes of internal and open source data. Applications such as the Cogito Intelligence Platform (link to CIP) are able to monitor thousands of sources and analyze large data volumes to extract from them only the relevant content.

H. *Content enrichment:* Though it is true that working with text content still requires a bit of human effort [21], text analytics techniques make a significant difference when it comes to being able to more effectively manage large volumes of information. Text mining techniques enrich content, providing a scalable layer to tag, organize and summarize the available content that makes it suitable for a variety of purposes.

I. *Spam filtering:* E-mail is an effective, fast and reasonably cheap way to communicate, but it comes with a dark side: spam. Today, spam is a major issue for internet service providers, increasing their costs for service management and hardware, software updating; for users, spam is an entry point for viruses and impacts productivity. Text mining techniques can be implemented to improve the effectiveness of statistical-based filtering methods.

J. *Social media data analysis**: Now a days,*** social media is one of the most inexhaustible sources of unstructured data that organizations have taken notice. Social media is increasingly being recognized as a valuable source of market and customer intelligence and companies are using it to analyze or predict customer needs and understand the perception of their brand. In both needs Text analytics can address both by analyzing large volumes of unstructured data, extracting opinions, emotions and sentiment and their relations with brands and products.

Though the above discussed applications are available in Market, there is No application available which will suit to University requirements. Hence we developed the new Text Analysis frame work for mining knowledge from text data typically the situation involved in many Universities. This will be useful for Universities having large number of students enrolled in Examination.

6. NEW TEXT ANALYSIS FRAME WORK

A. The Need for New Approach

As many applications make use of barcode and Optical Mark Recognition (OMR) systems to aid in automated information processing. Typically, we find their use in many university settings, where examination question papers and answer booklets adopt such systems and the answer booklets are captured by the system as images. One of the main reasons for this reform towards the use of OMR is firstly to protect the identity of the student writing the exam and the seat number of the student, before the answer booklet goes for evaluation to the examiner. Secondly, and more importantly, OMR systems are expected to aid in automate the process of answering the variety of queries raised that relate to the students' exam results after evaluation. Such queries on the answer booklets require intelligent information retrieval and Text Analysis system.

One of the universities we considered for proposing our approach was University of Mumbai. This university conducts the examinations twice a year, and they are referred as First Half i.e. examinations conducted in April-May and Second Half i.e. examinations conducted in October-November. The total number of students enrolled in 2016 were 1,61,173 (One Sixty One Hundred Thousand and One Hundred and Seventy Three). Thus the volume of data is very high. For example, the B.Com. Examination of 2016, had around 80,000 candidates who appeared for their final examinations. B.Com. has 64 subjects and from these subjects, student can opt for 7 subjects. Thus there are ($80000 \times 7 = 560{,}000$ answer booklets) i.e. half million and sixty thousand records

(answer books) for this one exam alone. Each answer booklet has 2 parts (560000 × 2 = 1,120,000), and hence there results in One Million Twenty Thousand images for this single exam alone.

In the barcode system, the first page of each answer booklet is divided into two parts – the first part contains the student information like seat number, subject code, center code, etc. along with unique bar code. The second part contains examiner information such as marks awarded for each question, total marks scored, subject code, signature of examiner, bundle number, answer book number etc. along with unique bar code. This two parts are scanned and the images are stored in the database. After declaration of examination results, students may apply to make a query of their marks obtained, such as, verification of marks, revaluation of marks, view their answer booklet, etc.

Due to huge number of answer booklets collected during every exam, and considering security concerns as well as cost, manpower, and other resources involved in scanning all the pages of the entire answer booklet, the management has decided to only scan the first page that provides two parts, namely student information and examination information. Figure 1 gives a sample image capturing the first page of the exam booklet. All the answer booklets are stored physically in a storage location, and in order to answer any of the user queries, the answer booklet is manually retrieved from the physical storage location. To achieve this, information relevant to the query had to be filtered and extracted from this huge data of images manually for identifying the correct storage location code for physical retrieval of the booklet. With growing number of student population and different subjects and degrees being offered, it is not practical to accomplish this manually for a timely answering of these queries. In addition to the problem of timely response to the queries, there are several drawbacks of the existing system. We adopted a systematic feedback and review of their manual and IT systems to reveal the drawbacks [32] [33]. So to extract the relevant information quickly from this text data was a real challenge.

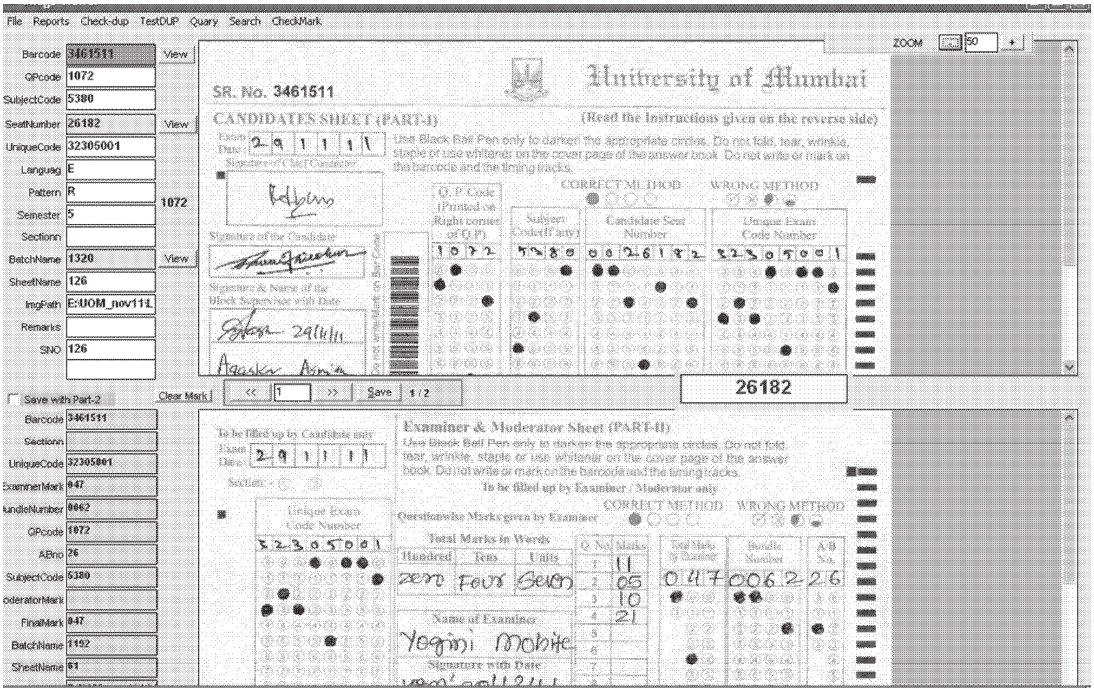

Figure 1: Sample scanned exam booklet with student and exam information

B. Proposed Intelligent Text Analysis Frame Work

A real-life situation described earlier is a typical example where IT plays an important role in automating the

processes for improving customer service and operations [34] [35]. Our proposed system consists of the following components:

(i) *Uploading of scanned Images:* This component uploads the images of all answer booklets consisting of First part (Student information) and Second part (Exam information). The images of First part have examination code, subject code, Bar code and other information of Student. While images of Second part have question wise and total marks given by Examiner or in some cases marks are also given by Moderator; examination code, subject code, bar code, bundle number, & answer book number. It also consist of processes to turn unstructured documents into structure information.

(ii) *Linking of images:* Images of First part and Second part are linked using the Barcode. Also, they are linked by examination code and seat number so that retrieval can be fast. For example, a query to retrieve the image based on seat number or subject code (examination code) would be much faster with such associations established.

(iii) *Document Classification:* This component is used to allocate an appropriate physical storage location for storing the answer booklet.

(iv) *Document clustering and routing:* This component does the information clustering and routing for initiating processes pertaining to other departments such as Photocopy Services department and Revaluation department.

(v) *Summarization of data:* This component does the summarization of data based on examination code, subject code, examiner code, etc.

(vi) *Information extraction:* This component consists of Query operations to transform the query in order to improve information extraction and retrieval.

(vi) *Interactive visualization:* This component interfaces with the front-end of the system which communicates with the query module to retrieve the images. For example - from these images the information about bundle number and answer booklet number can be used to retrieve the exact location in the storage for faster physical retrieval of the answer booklet.

Each component consists of sub-components and processes. But each component makes the framework smart by delivering capabilities to management and users to create predictive intelligence by uncovering patterns and relationships in both the structured and unstructured data.

7. CONCLUSION

Recently, we have witnessed the exponential increase in the amount of information being produced. This paper discusses the importance of Web Mining and new frame work for Text Analysis. The framework was implemented in automating the process of a physical search of the answer booklets of student examinations in a university setting. As the huge data comprises of many formats, including barcodes, student information, exam information, and image data, it has become mandatory to perform intelligent automation and to devise the methods of retrieving the relevant information.

REFERENCES

[1] E. Turban and Aronson J.E. *Decision Support Systems and Intelligent Systems*. Sixth Ed. New Jersey; Prentice Hall., 2001.

[2] K.E. Pearlson, *Managing and Using Information Systems: A Strategic Approach*. New York, Wiley, : John Wiley and Sons, Inc., 2001

[3] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus, *Knowledge Discovery in Databases: An Overview*. AI Magazine, Fall 1992

[4] R Srikant, and R Agrawal, "Mining sequential patterns : Generalizations and performance improvements", Proc. of the 5th International Conf. on Extending Database Technology, France (March), 1996.

[5] D., Hand, H. Mannila, and P.Smyth, *Principles of Data Mining* Cambridge, Massachusetts, The MIT Press,2001.

[6] R. Bradman, and T. Anand, "The Process of Knowledge Discovery in Databases: A Human- Centered Approach", Advances in Knowledge Discovery and Data Mining, Menlo Park, CA: The AAAI Press/The MIT Press, p. 37, 1996.

[7] U., Fayyad G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview", Advances in Knowledge Discovery and Data Mining, Menlo Park, CA: The AAAI Press/The MIT Press, p.1, 1996.

[8] C. Westphal, T. Blaxton, *Data Mining Solutions - Methods and Tools for solving real world problems*, Wiley Computer Publishing, USA, 1998.

[9] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi,, *Discovering Data Mining from Concept to Implementation*, Prentic Hall PTR, Inc. USA, 1998.

[10] R.,Baeza-Yates, and B. Rubiero-Neto, *Modern information retrieval. Reading,* MA: Addison Wesley, 1999.

[11] H. Chu,. "Information representation and retrieval in the digital age". Medford, NJ: Information Today, 2003.

[12] G Salton,. *Automatic Information Organization and Retrieval*. New York: McGraw-Hill., 1968.

[13] D.Soergel, *Organizing information: Principles of database and retrieval systems.* Orlando, FL: Academic Press, 1985.

[14] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[15] X. Zhou, Y. Li, Y. Xu and R. Lau. "Relevance assessment of topic ontology". In The Fourth International Conference on Active Media Technology, Relevance Assessment of Topic Ontology, 2006.

[16] S. J Kamatkar., "Information Retrieval in Web Mining for Knowledge Discovery", Proc. of International Conference on Machine Learning and Computing (ICMLC 2011) at Singapore, 2011.

[17] P. Gawrysiak : "Information retrieval and the Internet", PWII Information Systems Institute Seminars, 1999.

[18] D. Gibson, J.Kleinberg, P.Raghavan: "Inferring Web communities from link topology", Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, 1998.

[19] S.C., Cazella, and L.O.C Alvares,. "Modeling user's opinion relevance to recommending research papers." Proceedings of 10th International Conference on User Modeling(UM'05). Edinburgh, Scotland, 2005.

[20] Li, Y., Zhong, N. "Mining Ontology for Automatically Acquiring Web User Information Needs", IEEE Trans. on Knowledge and Data Engineering, Vol. 18, No. 4, 2006, pp. 554-568, 2006.

[21] D. C. Blair, *Language and representation in information retrieval*. Amsterdam: Elsevier Science, 1990.

[22] M.S.,Silver, *Systems that Support Decision Making,* John Wiley, Chichester, 1991.

[23] K.C., Laudonand J.P., Laudon, *Management Information Systems Organization and Technology*, Fourth Edition, Prentic-Hall of India, India, 1991.

[24] S.,Brin, and L. Page, "The anatomy of a large-scale hypertextual web search engine". 7th International World Wide Web Conference (WWW7). Computer Networks and ISDN Systems, 1998.

[25] S. Venkatraman, S. J. Kamatkar, "Intelligent Information Retrieval and Recommender System Framework" International Journal of Future Computer and Communication, ISSN 2010-3751, Vol 2, Issue 2, page 85, 2013.

[26] R, Agarwal, "Data Mining", Proc. International conference on Very Large Data Bases(VLDB), 1996.

[27] P. Buneman, S. Davidson, and D. Suciu, "Programming constructs for unstructured data". In Proceedings of ICDT'95, Gubbio, Italy, 1995.

[28] M. Balabanovic, S. Yoav, and Y. Yun., "An adaptive agent for automated web browsing". Journal of Visual Communication and Image Representation, 6(4), 1995.

[29] Z. Broder, S. C. Glassman, M. S. Manasse, and G Zweig., "Syntactic clustering of the web". In Proc. Of 6th International World Wide Web Conference, 1997.

[30] C. Chang and C. Hsu., "Customizable multi-engine search tool with clustering". In Proc. of 6th International World Wide Web Conference, 1997.

[31] V. Lavrenko and W.B. Croft., "Relevance-based language models", Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval ACM SIGIR, 2001.

[32] S. Venkatraman, "A Framework for ICT Security Policy Management" In Esharenana E. A. (Ed.), Frameworks for ICT Policy: Government, Social and Legal Issues, IGI Global Publishers, USA, ISBN13: 9781-6169-2012-8, pp. 1-14., 2011.

[33] S. J., Kamatkar, *Computer and Applications: A Desktop Quick Reference*, Rushwin Publisher, India, 2003.

[34] S. J., Kamatkar, "Information Technology an important tool for Management in 21st century", Proc. Of National Seminar organized by University of Mumbai, India, 2001.

[35] S. J., Kamatkar, "Role of Information Technology(IT) in Globalisation Era", Proc. Of International Conference, organized by University of Mumbai, India, 2002.

[36] G. Pandey and J. Luxenburger. " Exploiting session context for information retrieval- a comparative study". In ECIR, pages 652-657., 2008.

[37] S. Verberne, H. Van Halteren, S. Raaijmakers, D. Theijssen, and L. Boves, "Learning to Rank for Why-Question Answering. Information Retrieval", 2010.

[38] P. Lubell-Doughtie and K. Hofmann, "Learning to rank from relevance feedback for e-discovery". In ECIR., 2012.

[39] O. Chapelle and S. S. Keerthi., "Efficient algorithms for ranking with svms. Information Retrieval", 13(3):201-215, 2010.

[40] B.Xu, J. Bu C. Chen and D. Cai., "An Exploration of Improving Collaborative Recommender Systems via User-Item Subgroups". International World Wide Web Conference, 2012.

[41] Ian H. Witten, *Text Mining*, Computer Science, University of Waikato, Hamilton, New Zealand, 1999

[42] Sebastiani, F., "Machine learning in automated text categorization." ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47. 2002.

[43] Chen Hsinchun., *Knowledge management systems: a text mining perspective*. Tucson, Arizona: University of Arizona (Knowledge Computing Corporation), 2001

[44] Berson, A. and Smith, S.J., *Data warehousing, data mining, and OLAP*. New York: McGraw-Hill, 1997.

[45] Rob, P. and Coronel, *Database systems: design, implementation, and management*, 5 th ed. Boston, MA: Course Technology, 2002.

[46] Blog by Mike Ferguson, Managing Director of Intelligent Business Strategies Limited, [Online] Available: https://www.ibmbigdatahub.com/blog/what-text-analytics, 2016.