DESIDOC Bulletin of Information Technology, Vol. 25, No.4, July 2005, pp. 31-38

Role of Text Mining in Information Extraction and Information Management

M. Natarajan

Abstract

This article deals with the definitions of text mining and the need for it. The components of text mining are explained as information retrieval, information processing and information integration. The powerful technique that helps to identify the relevant S&T literature is enumerated with five primary conditions for information extraction. The three stages of Knowledge Discovery in Data and Data Mining (KDD) process is given with the applications of text mining. Applications of text mining vary from Information retrieval, bioinformatics, patent analysis, sorting gene expression, mining hospital records and multilingual approach to cross-lingual text retrieval (CLTR), analysis of e-mails, etc. Also explained the current status like new methods for text analysis in biology and medicine, the software viz. TextAnalyst meant for semantic analysis, navigation and search of unstructured texts with the 'Linking Information for Novel Discovery and Insight' (LINDI) project at Berkeley for helping the researchers. The internationally leading consortium for text mining viz. National Centre for Text Mining at UMIST with four UK partner institutions has been given. The limitations and benefits of text mining are given and concluded that library and information science professionals should implement the same for providing the relevant information from the vast amount of information.

1. INTRODUCTION

Web is a universal repository of knowledge such that almost any kind of knowledge can be found there. The users pay little to access the web. There is no central editorial board in the web, so anyone can contribute his resource to the web. These properties stimulate the users to retrieve information from the web and add information to the web, making the web grow in an incredible speed. Information finding is thus a serious problem since most users find it hard to obtain the information, using current information retrieval strategies.

Two kinds of strategies are now adopted by the web communities namely searching and browsing. The performance of a searching system is the direct result of its indexing strategy and ranking mechanism.

Data may potentially contain a great wealth of knowledge. However, analyzing huge amounts of textual data requires a tremendous amount of work in reading all of the text and organizing the content. Thus, the increase in accessible textual data has caused an information flood in spite of hope of becoming knowledgeable about various topics. A text mining technology called Text Analysis and Knowledge MIning (TAKMI) has been developed to acquire useful knowledge from large amounts of textual data such as internal reports, various technical documents, and messages from various individuals and so on. This article deals with text mining and applications to information management.

2. DEFINITIONAL ANALYSIS

Text mining (TM) also known as intelligent text analysis, text data mining or knowledge-discovery in text (KDT), refers generally to the process of extracting, interesting and non-trivial information and knowledge from unstructured text. TM is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics.

TM is the process of analyzing naturally occurring text for the purpose of discovering and capturing semantic information for insertion and storage in 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.

TM is defined as the computational process of extracting useful information from massive amounts of digital data by mapping low-level data into richer, more abstract forms and by detecting meaningful patterns implicitly present in the data.

TM is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information, together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. TM falls into an area called information extraction.

Information extraction (IE) software identifies and removes relevant information from texts, pulling information from a variety of sources and aggregates it, to create a single view. IE translates content into a homogeneous form through technologies like eXtensible Mark-up Language (XML). The goal of IE software is to transform texts composed of everyday language into a structured, database format. In this way, heterogeneous documents are summarized and presented in a uniform manner.

Librarians do data mining almost implicitly—to them, information falls into patterns, groups, clusters and hierarchies. Good text mining depends on the quality of the knowledge base on which it operates. If relationships, concepts, chronological information, and entities have already been extracted, then the text mining process can take advantage of this information and seek patterns within it.

3. COMPONENTS OF TEXT MINING

TM tools make it possible to discover new knowledge in the form of trends, anomalies, relationships and patterns that span multiple documents and large document collections. It can add valuable content analysis and decision support tools to existing intranets. TM for Science and Technology is the extraction of information from technical literature by the three major components viz. Information Retrieval, Information Processing and Information Integration.

High quality information retrieval produces both comprehensive and highly relevant records. Information processing is the extraction of patterns from the retrieved records by the three components viz. Bibliometrics, Computational Linguistics and Clustering. For multi-field structured records, with some free-text fields (such as paper Abstracts), bibliometrics is the extraction of the technical discipline infrastructure (authors, journals, organizations) as represented by the core records. Computational linguistics is the computer-based extraction of technical themes and their relationships. It is complex for technical literature analysis, because the technical phraseology appears as a foreign language to the computer. Clustering is the grouping of common technical themes and could be executed as phrase pattern groupings or actual document groupings. Information integration is the synergistic combination of the information processing computer output with the reading of the retrieved relevant records. The information processing output serves as a framework for the analysis and the insights from reading the records enhance the skeleton structure to provide a logical integrated product.

4. WHY TEXT MINING?

Text mining has attracted increasing interest and has been actively applied in knowledge management. Finding useful facts or 'nuggets' of knowledge in databases of text is the essence of text mining, an analysis process that attempts to uncover hidden patterns in unstructured text data. TM is currently being used in knowledge discovery and business intelligence applications ranging from human resource management to market intelligence to research and development. Its techniques are also being used to extend conventional information retrieval systems with features that create a more interactive and contextually aware search experience. Powerful information technology techniques now exist to identify the relevant S&T literatures and extract the required information. Techniques help to:

1. Substantially enhance the retrieval of useful information from global S&T databases;
2. Identify the technology infrastructure (authors, journals, organizations) of a technical domain;
3. Identify experts for innovation-enhancing technical workshops and review panels;
4. Develop site visitation strategies for assessment of prolific organizations globally;
5. Generate technical taxonomies (classification schemes) with human-based and computer-based clustering methods;
6. Estimate global levels of emphasis in targeted technical areas;
7. Provide roadmaps for tracking myriad research impacts across time and applications areas.

TM illuminates the trans-citation thematic relationships, and provides insights of knowledge diffusion to other intra-discipline research, advanced intra-discipline development and extra-discipline research and development. The addition of TM to citation bibliometrics makes feasible the large-scale multi-generation citation studies that are necessary to display the full impacts of research.

5. REQUIREMENTS FOR HIGH QUALITY TEXT MINING

For comprehensive access to the global S&T literature, and maximum extraction of useful information from this literature, five primary conditions are required.

1. A large fraction of the S&T conducted globally must be documented - information comprehensiveness.
2. The documentation describing each S&T project must have sufficient information content to satisfy the analysis requirements - information quality.
3. A large fraction of these documents must be retrieved for analysis - information retrieval.
4. Techniques and protocols must be available for extracting useful information from the retrieved documents - information extraction.
5. Technical domain and information technology experts must be closely involved with every step of the information retrieval and extraction processes - technical expertise.

Two types of technical expertise are required for a credible TM study viz. text mining technology expertise and technical (and related) domain expertise.

6. TEXT MINING PROCESS

The text mining process consists of three stages, namely data preparation, data processing and data analysis.

(a) Text preparation

The selection, cleansing and preprocessing of text. In this stage selection of sites or sources for TM would occur, usually under the guidance of a human expert or a well-trained software agent and early text preprocessing, such as sentence/paragraph identification and part-of-speech tagging, would take place.

(b) Text processing

The use of a data-mining algorithm to process the prepared data, compressing and transforming it to identify latent nuggets of information. At this stage, a fully featured Natural Language Processing (NLP) system would determine canonical and variant identities of entities (people, companies, organizations, etc.) identify conceptual relations between entities, and even instantiate particular frames of interest. Slot-filling of participants, dates and outcomes, as well as tables of extracted entities and relations, provides meaningful features for standard algorithms and techniques such as decision trees, neural networks, case-based learning, association rules or genetic algorithms.

(c) Text analysis

The evaluation of the output to see if knowledge was discovered and to determine its importance. Having run the algorithms, the mined text/data is submitted to various techniques that will enable direct usage of the mined information, either by a Link Discovery tool or by visualization in a tool that will enable human analysts to complete the analysis begun by the TM technology.

These three stages must be accomplished in a thoughtful manner, with appropriate attention paid to the goals of the particular TM task, the limitations of the data/text being mined and the strengths and weaknesses of the particular algorithm selected for the task. If these conditions are met, experience has shown that both confirming and disconfirming information will be discovered. Some quite unexpected ahas will result – which is the goal of text mining, data mining and all KDD.

7. APPLICATIONS OF TEXT MINING

Text mining finds applications in many diverse areas of wide interest such as drug discovery, predictive toxicology, protein interaction, competitive intelligence, protection of the citizen, identification of new product possibilities, detection of links between lifestyle and states of health, and many more. Some of the applications are:

Beginning with the release of IBM's Intelligent Text Miner in 1998, the bar for information retrieval (IR) has been raised and new IR products are now expected to have at least a clustering capability that group texts according to similarity of content, if not providing full mining capabilities. TM tries to apply these same techniques to unstructured text databases. It relies heavily on technology from the sciences of NLP, computational linguistics (CL) and machine learning to automatically collect statistics and infer structure and meaning in otherwise unstructured text. The usual approach involves identifying and extracting key features from the text that can be used as the data and dimensions for analysis. This process is known as extraction, which is a crucial step in text mining.

Companies looking to do duo-mining need to be wary of several things. First, some text mining technologies need large amounts of text to analyze - several page memos, for example - while call logs are sometimes just snippets in comparison. Second, "stemming," a popular technique in text analysis in which various forms of a word are distilled into one word - "pay," "paid," "will pay," "won't pay" = "pay" - may need to be turned off. Stemming would prevent the enterprise from understanding the customer's intent. Therefore, companies need to ensure that the technology they're using is tuned to the problem at hand. In addition, some companies' solutions are more toolkit-oriented while others are more application-oriented (Intelligent Results).

Another application is in bioinformatics, where details of experimental results can be automatically extracted from a large corpus of text and then processed computationally. For example it has been quoted that a support vector machine (SVM) with appropriate training can extract details of protein-protein interaction from the literature with greater than 90 percent accuracy. Some bioinformaticians have termed the body of literature the textome, which derives its name from the same naming convention, which gave us the genome, however this term is far

from universal. One of the largest text mining applications that exist is probably the classified ECHELON surveillance system.

For patent analysis, Byungun Yoon and Yongtae Park used text mining as a data processing and information-extracting tool. Since the original patent documents are expressed in text (natural language) format, it is necessary to transform raw data into structured data. Then, the process of keyword extraction is applied to identify keywords and to measure similarity between patents.

Tanabe Scherf, Smith Lee, Hunter and Weinstein developed a system named MedMiner to help them to sort out the thousands of gene expression correlations resulting from microarray experiments to separate "interesting biological stories" from mere epiphenomena and statistical coincidences. The first module gathers the relevant texts by querying PubMed (MEDLINE) and GeneCards (an Israeli gene information database) on the expressed genes. The second module filters the retrieved texts by user-specifiable relevance criteria based on classical proximity or term frequency scores. The third module is a "carefully designed user interface" to facilitate access to the most likely-to-be-interesting documents. MedMiner is not a true text mining system, but rather a search and display enhancement to PubMed. It is designed to deal with highly technical information by assisting expert users in their traditional IR tasks rather than attempting to automate them completely.

Electronic Data Systems (EDS) began using PolyAnalyst, which can mine intelligence from structured and unstructured data. It is based on an implementation of the WordNet semantic dictionary, which assigns words to subject categories and provides related words. Talkington uses PolyAnalyst for analyzing the open-ended questions but still uses traditional Business Intelligence (BI) software for the multiple-choice questions and combines the information from the consolidated reports.

There are separate tools that specialize in analysing either databases or text files, but there are also techniques that allow the two to be correlated. Patricia B. Cerrito, mines hospital records to discover ways to improve patient outcomes. She uses SAS Text Miner on text files, such as patient charts. But she also pulls in flat-file snapshots of billing and pharmaceutical databases and analyzes those as text, rather than as database entries. By examining thousands of patient outcomes with Text Miner, she has found useful information - that prescribing certain medications can prolong hospital stays for patients. Integrating text-based information with structured data enriches the predictive modeling capabilities and provides new stores of insightful and valuable information for driving business and research initiatives forward.

The application by Rowena Chau and Chung-Hsing Yeh is the multilingual text mining approach to cross-lingual text retrieval (CLTR). It is focused on exploiting the knowledge discovery capability of TM over multilingual text. This is a logical approach due to the complementary nature of these two areas. Both CLTR and multilingual text mining analyze multilingual textual data employing techniques from information retrieval, natural language processing and machine learning. In terms of the functions they perform, CLTR facilitates multilingual information access while multilingual text mining enables knowledge discovery from multilingual texts.

Sakurai and Suyama tested the TM for e-mail, which is mainly composed of textual data with usually a lot of grammatical errors; because few people take great care and concentrate when composing e-mails. Their system uses subjects and body of the e-mails and able to divide into text classes, which provide a viewpoint for analysis of e-mails. Also, the system extracts key concepts from the emails and presents statistical information, such as the distribution of key concepts and the number of e-mails included in each class. Moreover, the system lists emails relating to the number. Users are able to read only the e-mails classified into text classes corresponding to their interests. Users are able to understand the trend of the

key concepts for all e-mails by referring to the information.

8. TEXT MINING - CURRENT STATUS

Owing to the obvious applications for TM, many researchers have developed text-analysis methods tailored to the specific needs of biology and medicine. TM can be roughly divided into two disciplines viz. 1) information retrieval, which identifies relevant documents; and 2) information extraction, which determines relationships between relevant terms from the identified documents. TM embraces new methods for semantic analysis of texts, information retrieval and management. Making correct decisions often requires analyzing large volumes of textual information. Researchers, analysts, magazine editors, venture capitalists, lawyers, help desk specialists and even students are faced by various text analysis tasks.

Huge piles of information accumulate in numerous text repositories held at news agencies, libraries, corporations, individual PCs, and the Web. The amount of stored information proliferates at a disastrous rate, and the human eyes and brain are increasingly unable to meet the challenges of this growth. Mankind is searching for intelligent electronic assistants to help with text analysis projects. One needs to:

- Distill the meaning of a text in a concise form
- View accurate summaries before plunging into full documents
- Efficiently navigate through large textbases
- Perform natural language information retrieval

TextAnalyst, a unique software tool for semantic analysis, navigation and search of unstructured texts, can successfully tackle these and many other tasks. It helps to quickly summarize, efficiently navigate and cluster documents in textbase and performs semantic information retrieval or focus text exploration around a certain subject. A synergy of unique linguistic and neural network technologies implemented in TextAnalyst ensures high speed and accuracy in the analysis of unstructured texts.

Linking Information for Novel Discovery and Insight (LINDI) project at Berkeley is to investigate, how researchers can use large text collections in the discovery of new important information and to build software systems to help support this process. The main tools for discovering new information are of two types: support for issuing sequences of queries and related operations across text collections, and tightly coupled statistical and visualization tools for the examination of associations among concepts that co-occur within the retrieved documents. Both sets of tools make use of attributes associated specifically with text collections and their metadata. The LINDI interface provides a facility for users to build and so reuse sequences of query operations via a drag-and-drop interface. These allow the user to repeat the same sequence of actions for different queries.

There is growing interest in text mining as a tool to tap data that exists in unstructured text data formats. On the software vendor front, the few small software companies that specialize in text analysis and mining have remained relatively prosperous despite the recent economic downturn (see ClearForest [http://www.clearforest.com/], Inxight [http://www.inxight.com/], Temis [http://www.temis-group.com/], and Megaputer [http://www.megaputer.com]. Other more established companies are also getting into the act and forming the consortium for text mining. The National Centre for Text Mining at University of Manchester Institute of Science and Technology (UMIST) runs an internationally leading consortium. The consortium has four UK partner institutions: UMIST, the Victoria University of Manchester, the University of Liverpool, and the University of Salford. A major challenge for the Centre is to handle efficiently and robustly very large volumes of text and the intermediate data produced while processing. The consortium represents expertise in all the component areas of text mining, with an impressive array of

international partners. The Centre will play a leading role both nationally and internationally in developing the research agenda in text mining, promulgating associated best practice, and developing service provision. All those involved look forward to the challenges and opportunities that lies ahead.

TM is based on the marriage of two traditionally isolated technologies: natural language processing (NLP) and information management (IM). Selective use of NLP gives us the ability to identify words, phrases, relationships and even emotions in text. IM gives us the ability to filter, categorize, group and discover characteristics in messages. Together, these capabilities make it possible to fashion systems that turn text into data and messages into information. With respect to business decision making, TM has the ability to:

- Analyse vast amounts of unstructured free text
- Determine relationships among messages
- Group a large number of texts according to similarity of content
- Analyse documents to enable businesses to query both objective and — more important— subjective customer data. A sales manager can ask the system, "How many complaints today mentioned delayed delivery of our new printer?" Such analysis can reveal significant trends, enabling management to act preemptively and promptly
- Automatically alert executives to changes.

Thus, customer relationship management systems enhanced with TM will produce different outcomes.

9. LIMITATIONS/BENEFITS OF TEXT MINING

The fundamental limitations of TM are first, that we will not be able to write programs that fully interpret text for a very long time, and second, that the information one needs is often not recorded in textual form. TM can substantially improve the comprehensiveness and relevance of records retrieved from databases. It can help to overcome the problems created by limited coverage of information in current databases. It can provide strategies to find quickly text articles or database annotations referring to a specific entity, and can extract relationships between different entities.

TM has also been used or proposed for discovery and innovation from disjoint and disparate literatures. This application has the potential to serve as a cornerstone for credible technology forecasting, and help predict the technology directions of global military and commercial adversaries. Potential uses of such a TM capability include those essential to strategic or competitive intelligence. Scientists benefit from the limited capabilities of existing TM systems today. TM uses recall and precision (borrowed from the information retrieval research community) to measure the effectiveness of different information extraction techniques, allowing quantitative comparisons to be made.

CONCLUSIONS

Enormous efforts have been made for automatically discovering novel information from texts. The popularity of text mining techniques is reflected in number of publications and can be found in established forums such as ACM SIG on Knowledge Discovery in Data and Data Mining (KDD) and IEEE International Conference on Data Mining (ICDM). TM has the potential to make a significant impact on genome-based clinical research because free-text is currently the most widely used format for the storage and communication of information, both in a research and clinical environment. However, transformation of the extracted information into a structured representation of information is essential. Further, development of both ontologies and knowledge representations and their integration into NLP, in addition to improved accuracy of text mining methods will be required to enable broader acceptance and use of TM in the biomedical community. Furthermore, developments in visualization, curation and annotation of extracted information are required to integrate such tools into the day-to-day work of the genome researcher.

Advances in computer hardware speed and improvements in the effectiveness of text analysis software are making TM applications more realistic for commercial use than ever before. On the other hand, government intelligence agencies have long been interested in TM and associated technologies. Growth of the internet and the events of September 11, 2001, have increased interest in TM as a tool to track potential terrorist threats by analysing e-mail messages, online chat rooms, and other sources. Library professionals should take interest in finding the TM techniques and supply the requested relevant information to the users.

REFERENCES

1. Gieger, Christian; *& et al*. The future of text mining in genome-based clinical research. *Biosilico,* 2003, **1**(3), 97–102.
2. Hearst, M. What is text mining? Oct 17 2003.
3. Nasukawa, T. & Nagano, T. Text analysis and knowledge mining system. *IBM Systems Journals on Knowledge Management,* 2001, **40**(4).
4. Yang, Hsin-Chang & Lee, Chung-Hong. A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications,* 2004, **27**, 645–63.
5. Yoon, B. & Park Y. A text mining-based patent network: Analytical tool for high-technology trend. *Journal of High Technology Management Research,* 2004, **15**, 37–50.
6. http://bailando.sims.berkeley.edu/lindi.html
7. http://mappa.mundi.net/trip-m/hearst/
8. www.asis.org/Bulletin/Oct-00/liddy.htm
9. www.computerworld.com/printthis/2004
10. www.cs.waikato.ac.nz/~nzdl/textmining/
11. www.db2mag.com/db_area/archives/1998/q3/98fsoto.shtml
12. www.dmreview.com/editorial/dmreview/
13. www.dtic.mil/dtic/kostoff/index.html
14. www.intranetstoday.com/Articles
15. www.jisc.ac.uk/index.cfm?name=press_release_text_mining_290404
16. www.sas.com/technologies/analytics/datamining/
17. www.scils.rutgers.edu/~msharp/text_mining.html
18. www.sims.berkeley.edu/~hearst/text-mining.htm
19. www.unisysworld.com/monthly/2000/07/mining.shtml

Contributors: **Sh. M. Natarajan**, Scientist, NISCAIR (CSIR), 14, Satsang Vihar Marg, New Delhi – 110 067; e-mail: m_natarajan@hotmail.com