A Text Data -Mining Approach For Technology Innovation

Sindhuja. K ,
Assistant Professor ,
MGR Janaki College Of Arts & Science For Women Raja Annamalai Puram, Chennai 600 028
Phone: 9790820286

Abstract: In today’s digital economy, knowledge is regarded as an asset and the implementation of knowledge. Technology revolution has been facilitating millions of people by generating tremendous data via ever-increased use of a variety of digital devices and especially remote sensors that generate continuous streams of digital data, resulting in what has been called as “big data”. Data mining is a popular technological innovation that converts piles of data into useful knowledge that can help the data owners and the users make informed choices to take smart actions for their own benefit. Data mining is also known as Text mining or knowledge discovery from textual database, refers to the process of extracting interesting and non –trivial patterns or knowledge from text documents. The purpose of this paper is to study about the technological innovation regards the mining of databases or texts. In conclusion, it highlights the upcoming challenges of text mining and the opportunities it offers.

Keywords: Data mining, Text –mining, Text refining, big data management,

1. INTRODUCTION

Data mining seeks to extract useful information from various forms of data, but it usually emphasizes analyses of numerical data (e.g., linking your credit card purchases to your demographics). Tech Mining especially exploits text data sources of various sorts -- keying on *structured text* sources (e.g., abstract records from managed databases that separate information into fields such as “author,” “publication date,” and “keywords”), with moderate facility on semi-structured text (tagged) sources, and limited utility with unstructured texts. It complements efforts to exploit “big data,” but those generally deal with unstructured text and other data forms. Tech mining might be said to key on “large data.”

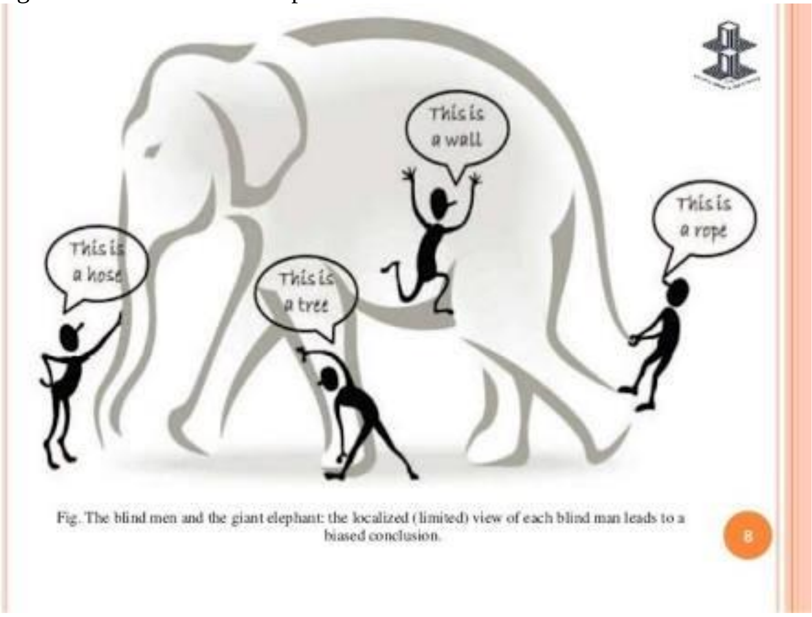

Fig. The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

As another example, on October 4, 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within two hours (Twitter Blog 2012).

Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about Medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 (Michel F. 2012). Assuming the size of each photo is 2 megabytes (MB), this resulted in 3.6 terabytes (TB) storage every single day. As “a picture is worth a thousand words”, the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters etc., only if we have the power to harness the enormous amount of data.

It demonstrates the rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a “tolerable elapsed time”. The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions (Rajaraman and Ullman, 2011). In many situations, the knowledge extraction process has to be very efficient and close to real-time because storing all observed data is nearly infeasible.

2. DEFINITION

Text and data mining involves the deployment of a set of continuously evolving research techniques which have become available as a result of widely distributed access to massive, networked computing power and exponentially increasing digital data sets, enabling almost anyone who has the right level of skills and access to assemble vast quantities of data, whether as text, numbers, images or in any other form, and to explore that data in search of new insights and knowledge.

TDM is important to researchers of all kinds. A historian with the necessary skills and an accessible digital archive can check the frequency with which a particular set of terms was used in the first half of the 19th century, compared with the second half. Analysis of vast quantities of video is crucial to research in meteorology and police forensics. A researcher in political economy can analyse the incidence and meaning of the word ‘digital’ in the work of the EU. Retailers can combine their knowledge of shoppers’ spending patterns with analysis of their leisure time and health. A medical researcher into Alzheimer’s disease may cross-examine unprecedented quantities of neurological and lifestyle data from patient records and investigations in many territories. Genetic studies and astronomy are among the areas of science which have already benefited significantly from these still very new and developing techniques. In short, TDM-based research plays a role in almost every area of human life, from banking, government and newspaper publishing to advanced manufacturing and advertising.

3. TECHNOLOGICAL CHALLENGES

Traditional publishers have raised concerns about the technologies employed in TDM and their ability adequately to service this activity without damage to their normal day to day operations. They argue that customers who have paid to read would experience a significant slowing down of the service available to them and this could result in publishers breaching their contract. Reed Elsevier, for example, believes that 20 researchers crawling their site would significantly reduce its functionality for other users.

Thomson Reuters supports this view, arguing that their system is not configured for third party TDM programmes crawling their systems which is likely to seriously impair if not crash their platforms.48 The Royal Society of Chemistry claims that, should the volume of TDM requests rise substantially, it would have to introduce additional server capacity, bandwidth and monitoring to deliver an online ‘on demand’ text mining service.

4. PROBLEMS AND FUTURE DIRECTIONS

Despite the great potential and the mushrooming of text mining products, there are technical issues to be overcome before text mining becomes a main stream technology.

Intermediate Form

Intermediate forms with varying degrees of complexity are suitable for different mining purposes. For a fine-grain domain-specific knowledge discovery task, it is necessary to perform semantic analysis to derive a sufficiently rich representation to capture the relationship between the objects or concepts described in the documents. However, semantic analysis methods are computationally expensive and often operate in the order of a few words per second. It remains a challenge to see how semantic analysis can be made much more efficient and scalable for very large text corpora.

Multilingual Text Refining

Whereas data mining is largely language independent, text mining involves a significant language component. Multilingual text mining is the area we expect to see a lot of activities in the next few years due to the substantial competitive advantages and the huge commercial potential that one can obtain through mining in languages other than English. Languages that are of particular interests include European languages and Asian languages, in particular Japanese and Chinese. As each language has a different syntactic structure and requires specialized semantic interpretation, a systematic approach for bringing in language modeling is inevitable and will form an essential part of multilingual text mining.

Domain Knowledge Integration

Current text mining systems do not make use of domain knowledge. We expect it to be an integral component of the future text mining tools. Domain knowledge is useful in orientating and focusing attention so as to improve the text parsing efficiency and to help to derive a more compact representation. Domain knowledge also plays a major role in knowledge distillation tasks. In a classification or predictive modeling task, for example, domain knowledge helps to improve learning/mining efficiency as well as the quality of the learned model (or mined knowledge) (Tan, 1997). It is also interesting to explore how a user's knowledge can be used to initialize a system's knowledge structure and make the discovered knowledge more interpretable.

Personalized Autonomous Mining

Another important dimension of research is to make text mining tools more user friendly. Current text mining products/applications are designed for trained knowledge specialists. Future text mining tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives. There have been some efforts in developing systems that interpret natural language queries and perform appropriate mining operations automatically. Text mining tools could also embedded in intelligent personal assistants (Tan & Teo, 1998). Under the agent paradigm, a personal miner would learn a user's profile, conduct text mining operations automatically, and forward information without requiring an explicit request from the user.

5. CONCLUSIONS

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are (1) huge with **h**eterogeneous and diverse data sources, (2) **a**utonomous with distributed and decentralized control, and (3) **c**omplex and **e**volving in data and knowledge associations. Such combined characteristics suggest that Big Data requires a "big mind" to consolidate data for maximum values (Jacobs 2009).

We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real-time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

REFERENCES

[1] .Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 603-630

[2] Alam et al. 2012, Md. Hijbul Alam, JongWoo Ha, SangKeun Lee, Novel approaches to crawling important pages early, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 707-734

[3] .Aral S. and Walker D. 2012, Identifying influential and susceptible members of social networks, *Science*, vol.337, pp.337-341.

[4] Machanavajjhala and Reiter 2012, Ashwin Machanavajjhala, Jerome P. Reiter: Big privacy: protecting confidentiality in big data. *ACM Crossroads*, 19(1): 20-23, 2012.

[5] Banerjee and Agarwal 2012, Soumya Banerjee, Nitin Agarwal, Analyzing collective behavior from blogs using swarm intelligence, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 523-547

[6] Birney E. 2012, The making of ENCODE: Lessons for big-data projects, *Nature*, vol.489, pp.49-51.

[7] Bollen et al. 2011, J. Bollen, H. Mao, and X. Zeng, Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, 2(1):1-8, 2011.

[8] .Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering,Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.

[9] Navathe, Shamkant B., and Elmasri Ramez, (2000), "*Data Warehousing And Data Mining*", in "*Fundamentals of Database Systems*", Pearson Education pvt Inc, Singapore, 841-872.

[10] Tan, A.-H. (1999), "Text Mining: The state of the art and the challenges", in *Proceedings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*,Beijing, April, 1999

[11] Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship toinformation access. Presentation notes for UW/MS workshop on data mining, July 1997.