Merits and characteristics of text mining

Chong Ho Yu, Ph.D. (2009)

Arizona State University

http://creative-wisdom.com

The objective of this article is to discuss the merits and characteristics of text mining. As a survey-based researcher in political science, Nie was impressed by how much effort survey researchers devote to developing open-ended questions and transcribing the responses, but was also disappointed by how little use is made of the data in the end (SPSS Inc, 2006). One of the reasons why open-ended responses are underused is that analyzing such data is time-consuming and human coders must be well-trained in order to yield accurate classifications. With advances in computing technology, automated text analysis can alleviate much of the preceding problem. However, while this technology has been gaining popularity across various industries and fruitful results have been benefited corporations for over a decade, its use in educational research is still rare. More specifically, most activities regarding text mining applications can be found in the biomedical field (e.g. identification of protein genes), government and national security (e.g. detection of terrorist activities), financial (e.g. sentiment analysis, trend analysis), political science (e.g. analysis of election campaign and voter profile), and business (e.g. analysis of customer queries/satisfaction, improvement of hospitality human resource management, spam mail filtering) (Cohen & Hersh, 2005; Corti, 2006; Miller, 2005; Singh, Hu, & Roehl, 2007). One seminal accomplishment via text mining happened at Hewlett-Packard (HP). After merging with Compaq, HP employed *SAS Text Miner* to determine what its most valuable customers were discussing and then developed new customer service strategies based on the recurring themes extracted from text mining (SAS Institute, 2008). On the other hand, applications of text mining in educational research are still sporadic, such as assessing asynchronous discussion forums (Dringus & Ellis, 2005) and research on adaptive Web-based courses (Romero & Ventura, 2007). The objective of this article is to illustrate how text mining can also be applied to educational research.

Transplanting industry applications into academic settings is not new. When data mining tools became more prevalent in industry, some educational researchers (e.g. Luan, 2002) had mapped corresponding research questions between industry and higher education. For example, customer profile analysis is equivalent to student profile analysis. Web log analysis for website enhancement is similar to Web traffic analysis for online course improvement. Retaining loyal customers is almost exactly the same as student retention research. The use of data mining techniques for exploratory data analysis and prediction, such as recursive partition trees, K-mean clustering, and neural networks, has been gaining momentum among educational researchers. In fact, the nature of the research question in the HP case mentioned above is not much different from the preceding questions. Procedurally speaking, there is also a high degree of resemblance between the two methods. Unlike hypothesis testing that has a strong confirmatory character, both data mining and text mining use certain classification algorithms to extract the *hidden patterns* from the data. The major difference is the data type. While data mining procedures are designed for structured, numeric data, text mining is primarily for unstructured or semi-structured data. Weiss, Indurkhya, Zhang, and Damerau (2005) asserted that text mining and data mining for solving pattern-recognition and prediction problems are the same type of method, no matter whether the data are structured or not. Thus, they contended that there is nothing unique about text mining that makes it different from data mining. If data mining is a suitable candidate for educational research by translating business-oriented questions into institutional research questions, there is no reason that text mining cannot be widely adopted for advancing educational research.

Text mining and undiscovered public knowledge

Text Mining is typically defined as a process of extracting useful information from document collections through the identification and exploration of interesting patterns (Feldman & Sanger, 2007). Some authors attribute the historical roots of text mining to information retrieval (IR), a branch of science that aims to search for information in documents developed in the 1950s and 1960s (Weiss et al. 2005); however, IR is just a small part of text mining. Specifically, common applications of IR are found in library systems and search engines. In the former the objective is to provide users with access to books, journals, and other documents, whereas in the latter, Web search engines, such as Google and Yahoo, return Web pages according to the key phrases entered by users. First, using pre-determined labels to retrieve information is in sharp contrast to the exploratory character of text mining. Text mining does not assume a pre-established taxonomy. Rather, the discovery engine examines the corpus each time a query is made, and therefore it is potentially capable of discovering new relationships and network nodes that were not known before (Haravu & Neelameghan, 2003). More importantly, returning a list of search results in IR, which may or may not be relevant to the subject matter under study, does not make a substantive contribution to new knowledge discovery, which is a major goal of text mining. For example, it is estimated that Google, arguably the most powerful search engine, could return only 30% relevant documents out of all results. In other words, every time an inquirer conducts a search, more than 70% of the documents in the output are irrelevant (Rzhetsky, Seringhaus, Gerstein, 2008).

In terms of objective and procedure, text mining is closely related to searching for “undiscovered public knowledge.” “Undiscovered public knowledge,” the notion that existing knowledge bases can be mined to generate new knowledge, was advocated by Swanson (1986a, 1986b, 1987, 1988, 1989s, 1989b, 1990a, 1990b) and Swanson and Smalheiser (1997, 1999). Swanson asserted that “gems” are buried in voluminous texts available in the public domain, but their potential is hardly actualized. It is because specialists can only read a small subset of literature in their own fields and are often unaware of developments in other seemingly unrelated areas. Swanson argued that through an intensive mining process it is possible to find relevant linkages between “fragmented” information in literature. In the 1980s and 1990s, Swanson’s idea was not widely used by researchers (Spasser, 1997), but with the advance of text parsing algorithms Swanson’s process has been formulated in text mining. Although earlier text mining was commonly used in literature review, later it was extended to many archival textual data, such as memos, websites, blogs, email messages, customer call center logs, etc. Today, text mining is by no means restricted to analyzing existing data. Rather, new textual data can be collected by inserting open-ended items in surveys as well as by other means.

In addition, applications of text mining have been extended to both unstructured and semi-structured data. While documents that do not conform to certain formats, such as articles scattering across various journals written for different topics in different styles (e.g. APA, Chicago, MLA…etc), are regarded as unstructured data, essay-type survey responses are considered *semi-structured* data because participants responded to the same survey item under the same question format (Morse, 1997). Nonetheless, there is no sharp demarcation between unstructured and semi-structured data. The meanings of semi-structured data and unstructured data can vary from author to author. One of the criteria for their distinction is whether portions of the data have associated meta-data. According to Ukelson (2007), a purchase order sent by fax has no explicit meta-data about the nature of the “data” and thus it takes a human to extract the relevant data items from the document. On the other hand, a text message packaged in XML has meta-data and therefore it is analyzable by software. This differentiation is debatable because the latest text mining technology is capable of handling textual data despite the absence of XML-type meta-data. Thus, some authors go beyond XML and relax the definition of meta-data. In this view, virtually all electronic documents, including email messages, HTML Web pages, PDF files, and word processing files with heavy document templating or style-sheet constraints, have some form of meta-data and thus can be classified as semi-structured data (Feldman & Sanger, 2007). By this definition the demarcation is meaningless because today it hardly finds any plain text electronic document. Some authors have suggested that written texts, which have a rich amount of syntactical and semantic structure, should be considered semi-structured or “weakly structured” (Feldman & Sanger, 2007; Natarajan, Berrar, Hack, Dubitzky, 2005). Again, this conceptualization of data type may not be helpful because all languages have certain syntactical and semantic structures, and following this reasoning one can always find structure in anything. It is the conviction of the authors that whether the written texts are in response to a common question under a common format should be the criterion for distinguishing semi-structured data from unstructured data. In short, text mining should be defined as computer-assisted knowledge discovery by analyzing both unstructured and semi-structured textual data from existing public knowledge bases and newly collected data sets using algorithms based on natural language processing (NPL). NPL is beyond the scope of this paper. Interested readers could consult Jurafsky and Martin (2000), Gelbukh (2007), Kao and Poteet (2007), Manning and Schütze (1999).

A typical text mining process is composed of the following three steps:

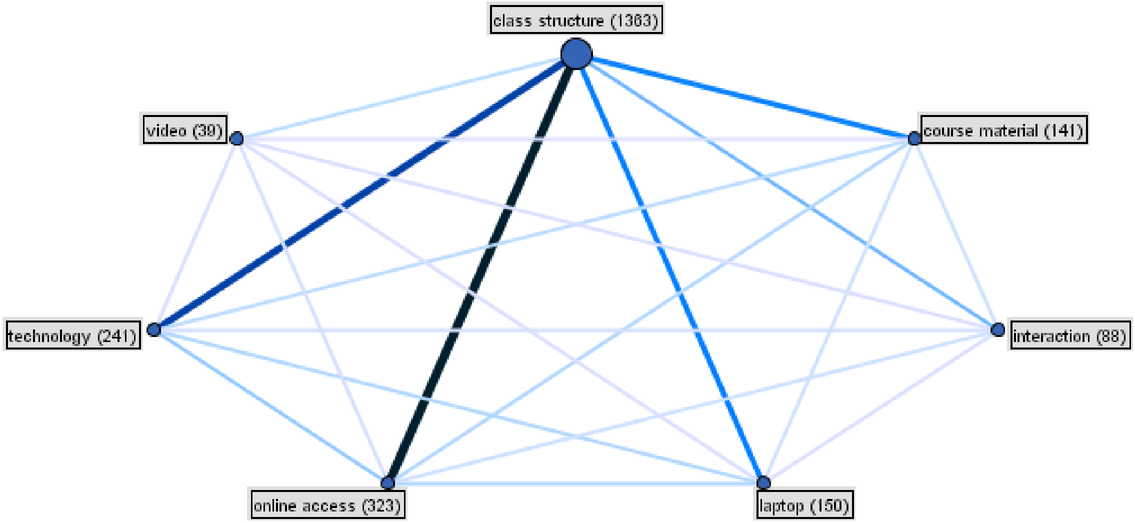
1. *Extraction*: As the name implies, this is the process of extracting information to serve as the building blocks for categorization from the textual data. A typical extraction process returns three types of results, namely, terms, types, and patterns. Terms are words or phrases that carry important connotations. During the extraction process, many frequently recurring but trivial words, such as “a,” “an,” “the,” “is,” “am,” “are,” “however,” “although,” “but,” etc., are ignored. Types are semantic groupings of terms. For example, consider the “teacher” type groups terms: “professor,” “faculty,” “instructor,” and “trainer.” Patterns are composed of a combination of terms and built-in types (SPSS Inc., 2006).

2. *Categorization*: In text mining, categorization refers to the process of grouping related concepts, themes, or common threads. This process is data-driven and iterative. Specifically, some concepts are proposed based upon the initial findings of the text patterns. Afterwards, all subsequent responses are scanned to check whether any text falls into the existing categories. New categories may be created and existing ones may be updated during this process. But categorization is not mutually exclusive; the same passage can be assigned into several categories. This overlapping of memberships enables the next step, namely, concept linking (SPSS Inc., 2006).

3. *Concept linking*: In conventional statistical research, it is common for researchers to formulate a new hypothesis by visually inspecting a correlation matrix to determine potential linkage between variables. By the same token, after grouping responses into several concepts, the text miner can examine the relationships among these themes using concept maps. Figure 1 is a typical concept map yielded from *SPSS Text Analysis* (SPSS Inc., 2006).[1]

[1] In 2009 SPSS renamed its product line to *Predictive and Analytical Software* (PASW). SPSS’s text mining software module is now called *PASW Text Analytics*.

Figure 1. Example of concept map.



The classic examples of hypothesis generation via concept linking are a series of studies conducted by Swanson (1986). Based on the idea of concept linking, Swanson carefully scrutinized the medical literature and identified relationships between some apparently unrelated events, namely, consumption of fish oils, reduction in blood viscosity, and Raynaud’s disease. His hypothesis that there was a connection between the consumption of fish oils and the effects of Raynaud’s syndrome was eventually validated by experimental studies (DiGiacomo., Kremer, & Shah, 1989). Using the same methodology, the links between stress, migraines, and magnesium were also postulated and verified (Swanson, 1988, 1989a; Thomas, Thomas, & Tomb, 1992).

Since then, numerous studies set the goal of mimicking Swanson’s process of hypothesis generation (e.g. Banerjee, Hu, & Yoo, 2005; Bekhuis, 2006; Weeber, et al., 2001, 2003). It is important to note that although concept linkage can logically lead to hypothesis generation, the latter is not an indispensable component of text mining. Bluntly, qualitative researchers may simply want to report the relationships between common threads in a descriptive fashion rather than going further to hypothesis testing.

Conclusion

Unlike imposing the researcher's preconceived ideas on the phenomenon under study by using force-option survey responses or pre-determined hypotheses, in text mining the researchers let the themes emerge freely from the subjects. Text mining for discovering common threads fits naturally into the sequential exploratory design in the mixed-method paradigm (Creswell & Plano Clark, 2007). According to Creswell and Plano Clark, when the research problem is ill-defined and the variables are unknown, the researcher could employ qualitative methods to understand the problem rather than imposing a theory on the phenomenon. This design is particularly useful when the researcher needs to develop an instrument to collect data, but is not sure about what constructs should be measured and what questions should be asked.

Acknowledgement

Special thanks to Mr. Charles Kaprolet for his professional editing of this paper.

References

Banerjee, P., Hu, X., & Yoo, I., (2005). Discovering the wealth of public knowledge: An approach to early threat detection. *AIS SIGSEMIS Bulletin, 2*, (3&4), 77-85.

Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries, 3*(2). Retrieved May 31, 2008 from http://www.bio-diglib.com/content/3/1/2

Cohen, A., & Hersh, W. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6,* 57-71.

Corti, L. (2006 April). *The SQUAD project*. Paper presented at National Centre for e-Social Science Workshop, Manchester, United Kingdom.

Corti, L. (2006 May). *Mine your data: Contrasting data mining approaches to numeric and textual data sources*. Paper presented at IASSIST conference, Ann Arbor, MI.

Creswell, J., & Plano Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

DiGiacomo, R.A., Kremer, J. M., & Shah, D. M., (1989). Fish-oil dietary supplementation inpatients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *American Journal of Medicine, 86*(2), 158-164.

Dringus, L., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education, 45*, 141–160.

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

Gelbukh, A. (Ed.). (2007). *Computational linguistics and intelligent text processing: 8th international conference, Mexico City, Mexico, February 18-24, 2007 proceedings*. New York: Springer.

Haravu, L. J., & Neelameghan, A. (2003). Text Mining and data Mining in Knowledge Organization and Discovery: The Making of Knowledge-Based Products. *Cataloging & Classification Quarterly, 37*, 97-113.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall.

Kao, A., & Poteet, S. (Eds). (2007). *Natural language processing and text mining*. London: Springer.

Luan, J. (2002). Data mining and its applications in higher education. In A. Serban & J. Luan (Eds.), *Knowledge management: Building a competitive advantage in higher education* (pp. 17-36). PA: Josey-Bass.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Miller, T. (2005). *Data and text mining: A business applications approach*. Upper saddle River, NJ: Pearson.

Natarajan, J., Berrar, D., Hack, C. J., Dubitzky, W. (2005). Knowledge discovery in biology and biotechnology texts: A review of techniques, evaluation strategies, and applications. *Critical Reviews in Biotechnology, 25,* 31-52.

Romero, C, & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005 *Expert Systems with Applications, 33,* 135–146.

Rzhetsky, A., Seringhaus, M., Gerstein, M. (2008). *Seeking a new biology through text mining. Cell, 134*, 9-13,

SAS Institute. (2008). *Mining more than gold: HP soaring to the next level of CRM with SAS.* Retrieved May 30, 2008 from http://www.sas.com/success/hp.html

Singh, N., Hu, C., & Roehl, W. (2007). Text mining a decade of progress in hospitality human resource management research: Identifying emerging thematic development. *Hospitality Management, 26*, 31–147.

Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *24*(4), 35-43.

Spasser, M. (1997). The enacted fate of undiscovered public knowledge. *Journal of the American Society for Information Science. 48,* 707-717.

SPSS Inc. (2006). *SPSS text analysis for surveys 2.0 user's guide*. Chicago, IL: The Author.

Swanson, D. R. (1986a). Fish-oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7-18.

Swanson, D. R. (1986b). Undiscovered public knowledge. *Library Quarterly, 56,* 103-118.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine, 31,* 526-557.

Swanson, D. R. (1989a). A second example of mutually isolated medical literatures related by implicit, unnoticed connections. *Journal of the American Society for Information Science*, *40*, 432-435.

Swanson, D. R. (1989b). Online search for logically-related noninteractive medicial literatures: A systematic trial-and-error strategy. *Journal of the American Society for Information Science, 40*, 356-358.

Swanson, D. R. (1990a). Medical literature as a potential source of new knowledge. *Bulletin of the Medicial Library Association, 78*(1), 29-37.

Swanson, D. R. (1990b). Somatomedin C and argine: Implicit connectionsbetween mutually isolated literature. *Perspectives in Biology and Medicine, 33*(2), 157-186.

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence. 91*, 183-203.

Swanson, D. R., & Smalheiser, N. R. (1999). Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. *Library Trends, 48*(1), 48-59.

Swanson, D. R., (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science, 38*, 228-233.

Thomas, J., Thomas, E., & Tomb, E., (1992). Serum and erythrocyte magnesium concentrations and migraine. *Magnesium Research, 5*(2), 127-30.

Ukelson, J. (2007). *Structured, semi-structured and unstructured data in business applications*. Retrieved May 29, 2008 from http://exeedtechnology.com/structured-semi-structured-and-unstructured-data-in-business-applications

Weeber, M, Klein, H., de Jong-van den, B., Lolkje T. W., & Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52, 548-57.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T., Aronson, A. R., & Molema,G. (2003). Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. *Journal of American Medical Information Association, 10*, 252-259.

Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. (2005). *Text mining: Predictive methods for analyzing unstructured information*. New York: Springer.