Web-Based Text Mining of Hotel Customer Comments Using SAS® Text Miner and Megaputer Polyanalyst®

Richard S. Segall
Arkansas State University, Department of Computer & Information Technology,
College of Business, Jonesboro, AR 72467-0130, rsegall@astate.edu

Qingyu Zhang
Arkansas State University, Department of Computer & Information Technology,
College of Business, Jonesboro, AR 72467-0130, qzhang@astate.edu

Mei Cao
University of Wisconsin – Superior, Department of Business and Economics,
Superior, WI 54880, mcao1@uwsuper.edu

ABSTRACT

This paper presents text mining using SAS® Text Miner and Megaputer PolyAnalyst® specifically applied for hotel customer survey data, and its data management. The paper reviews current literature of text mining, and discusses features of these two text mining software packages in analyzing unstructured qualitative data in the following key steps: data preparation, data analysis, and result reporting. Some screen shots are presented for web-based hotel customer survey data as well as comparisons and conclusions.

Keywords: Web-based data, Algorithms, Web data management, Text mining, SAS® Text Miner, Megaputer PolyAnalyst®, Unstructured Data, Hotel Customer Surveys

INTRODUCTION

The increasing use of textual knowledge applications and the growing availability of online textual sources caused a boost in text mining research. This paper reviews some current literature of text mining and presents comparisons and summaries of the two selected software of SAS® Text Miner and Megaputer PolyAnalyst® for text mining.

This paper discusses features offered by these two-selected software in the following key steps in analyzing unstructured qualitative data: software composition for text mining, data preparation, data analysis and result reporting. This paper also provides insight with the data management example of application to an actual web-based dataset for hotel customer survey data.

The novelty of this research is the illustration of the usefulness of text mining to the application to web-based hotel customer comments and the comparison of two popular software packages for applications to text mining. This paper also provides some visual insights into text mining. This paper compares integrated text mining software of SAS® Text Miner with a data mining software with text mining capabilities of Megaputer PolyAnalyst®.

LITERATURE REVIEW

Text mining (TM) is the discovery of new and useful information by automatically extracting information from textual document repositories. Text mining is to mine the patterns from natural language rather than from structured database of facts. It is a process that employs a set of algorithms for converting unstructured text into structured data conveying the insightful information.

Text mining process includes text preprocessing, feature generation and selection, pattern extraction, to analyzing results (see Figure 1 from Liang [5]). Saravanan et al. [9] discuss how to automatically clean data by discovering classes of similar items that can be grouped into prescribed domains. Hersh [4] evaluates different text-mining systems for information retrieval. Turmo et al. [14] describe and compare different approaches to adaptive information extraction from textual documents and different machine language techniques. Amir et al. [1] describe a new tool called maximal associations that allows the discovering of interesting associations often lost by regular association rules.

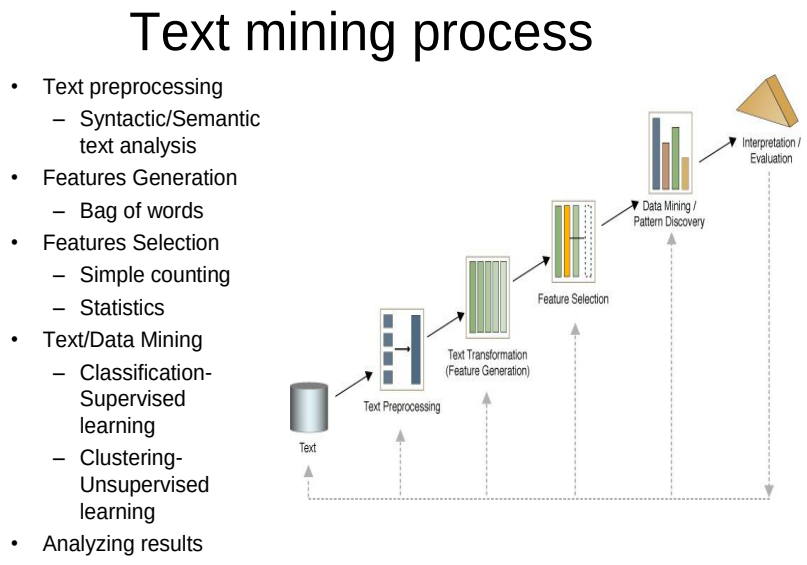

Figure 1: Text Mining Process (Liang [5])

Srinivasan [13] develops an algorithm to generate interesting hypotheses from a set of text collections using Medline database. This is a fruitful path to ranking new terms representing novel relationships and making scientific discoveries by text mining. Romero and Ventura [7] survey text mining applications in the educational setting. Yoon et al. [17] propose a road mapping methodology of text mining to assist decision-making (i.e., extract key information from documents such as product manuals and patent documents by text mining). Van Geiist et al. [15] use data mining and boosting algorithms to create a support system for predicting end prices on eBay.

Text mining uses recall and precision to measure the effectiveness of different information extraction techniques, allowing quantitative comparisons to be made. Scherf et al. [11] discuss the applications of text mining in literature search to improve accuracy and relevance. Yang and

Lee [17] develop an approach to automatically generate category themes and reveal the hierarchical structure. Fan et al. [3] describe a method using genetic programming to discover new ranking functions in the information-seeking task for better precision and recall. Seewald et al. [12] describe an application for relevance assessment for multi-document summarization. To characterize certain document collections by a list of pertinent terms, they have proposed a term utility function, which allows a user to define parameters for continuous trade-off between precision and recall.

Zhang and Segall [19] discussed using SAS® Enterprise Miner™ and Megaputer PolyAnalyst® for data mining needs for forecasting but did not address the application to text mining. PC AI [6] is a Buyer's Guide that includes that for Text Mining and specially lists and describes both SAS® Enterprise Miner™ and Megaputer PolyAnalyst®.

Robb [7] discussed both SAS® Text Miner and Megaputer PolyAnalyst® in an article about text mining tools applied to unstructured data. Robb [7] claims that unstructured data, "most of it in the form of text files, typically accounts for 85% of an organization's knowledge stores, but it's not always easy to find, access, analyze or use."

Crowsey et al. [2] conducted study using five text mining software tools with four undergraduate students at the University of Virginia who were had some experience in data mining but little experience in text mining. The text mining software used in study by Crowsey et al. [2] included a discussion of Megaputer PolyAnalyst® and SAS® Enterprise Miner but did not include SAS® Text Miner or a working copy of Megaputer PolyAnalyst®. The evaluation by Crowsey et al. [2] concluded using their available resources that the two software products of SAS® and SPSS had qualities that made them more desirable than others.

TEXT MINING SOFTWARE

As a comparison of the features for the two selected text mining software, Table 1 below is constructed where essential functions are indicated as being either present or absent with regard to data preparation, data analysis, and results reporting. As Table 1 shows that both software has similar extensive text mining capabilities except that there is no categorization analysis in SAS® Text Miner and no automatic text cleaning in Megaputer PolyAnalyst®.

Table 1: Text Mining Software

Software Features		SAS® Text Miner	Megaputer PolyAnalyst®
Software Composition for Text Mining	Separate Text Mining Software	x	
	Data mining software with text mining capabilities		x
Data Preparation	Text parsing and extraction	x	x
	Automatic Text Cleaning	x	
Data Analysis	Categorization		x
	Concept Linking	x	x
	Text Clustering	x	x
	Dimension reduction techniques	x	x
Results Reporting	Interactive Results Window	x	x
	Support for multiple languages	x	x

RESULTS

SAS® Text Miner

SAS® Text Miner is actually an “add-on” to SAS® Enterprise Miner™ with the inclusion of an extra icon in the ”Explore” section of the tool bar. SAS® Text Miner performs simple statistical analysis, exploratory analysis of textual data, clustering, and predictive modeling of textual data. SAS® Text Miner uses the “drag-and-drop” principle by dragging the selected icon in the tool set to dropping it into the workspace.

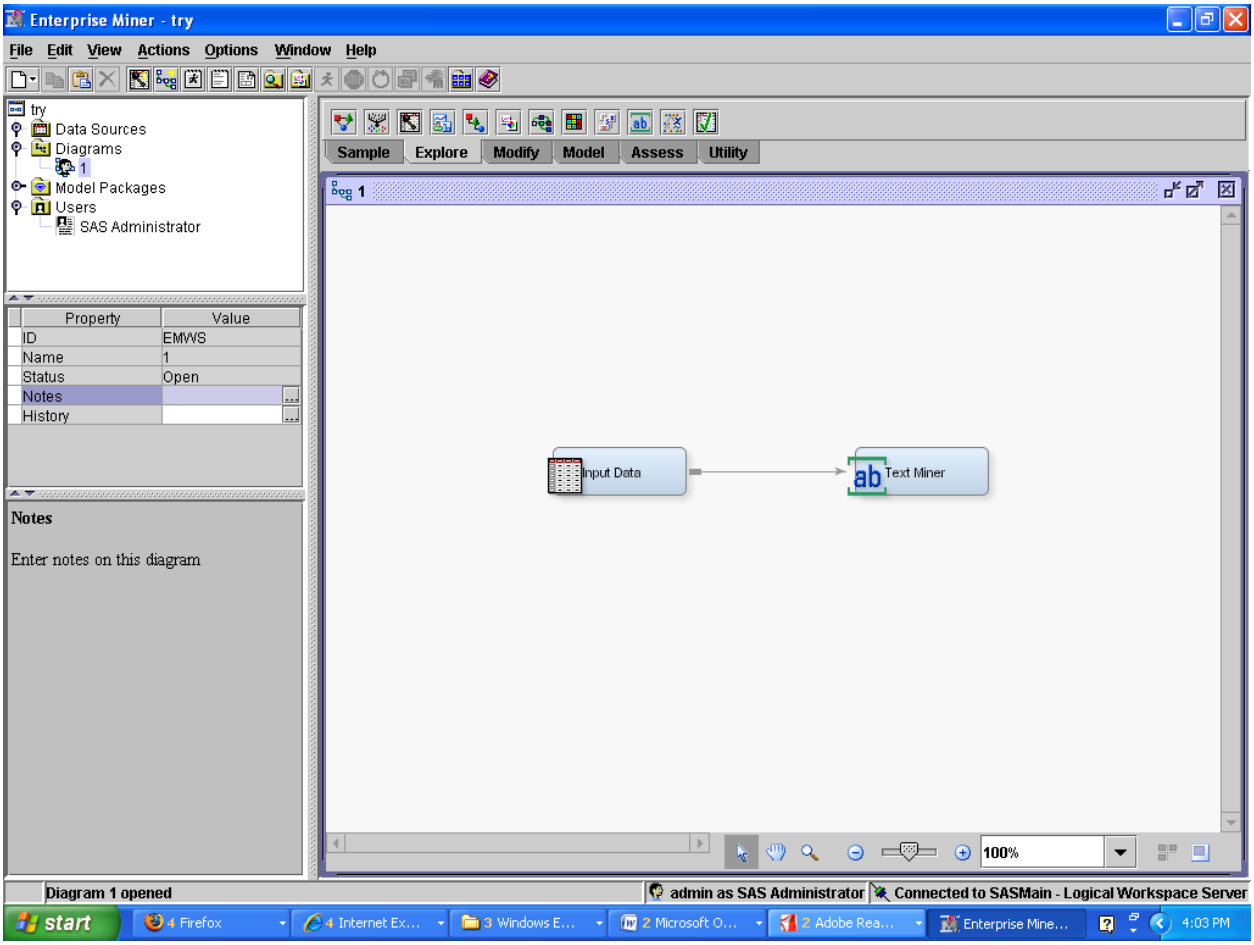

Figure 2: Workspace of SAS® Text Miner for SAS Enterprise Miner 5.3

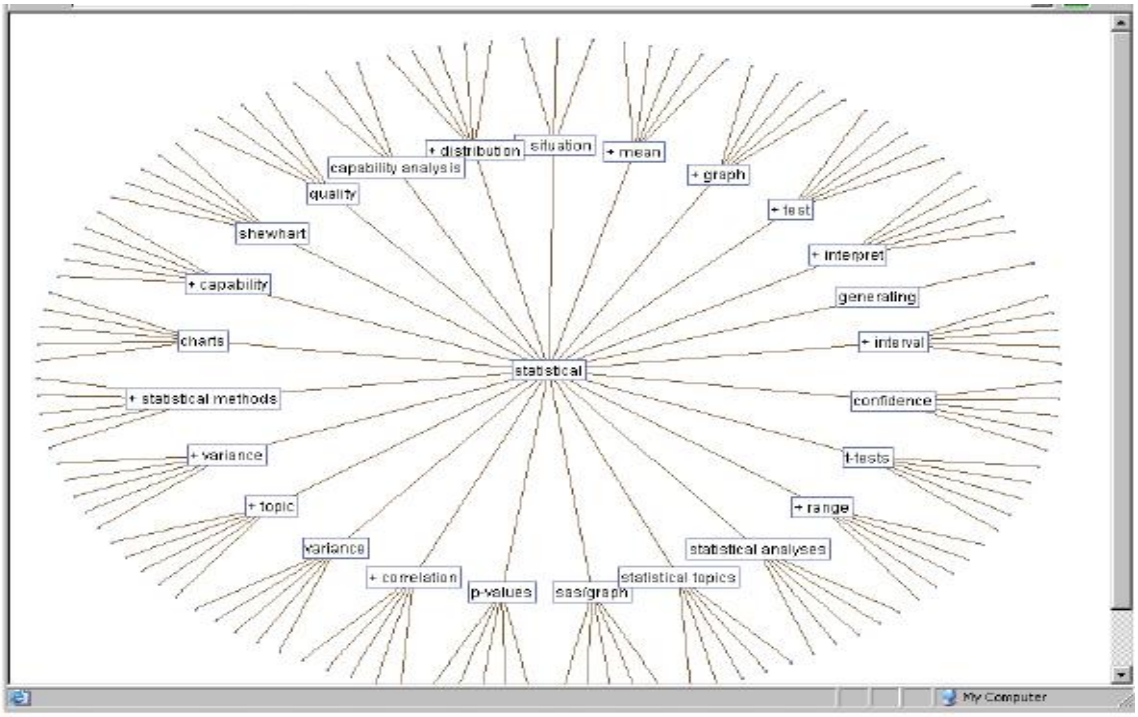

Figure 3: Concept Links in SAS® Text Miner showing first order terms (Woodfield [16])

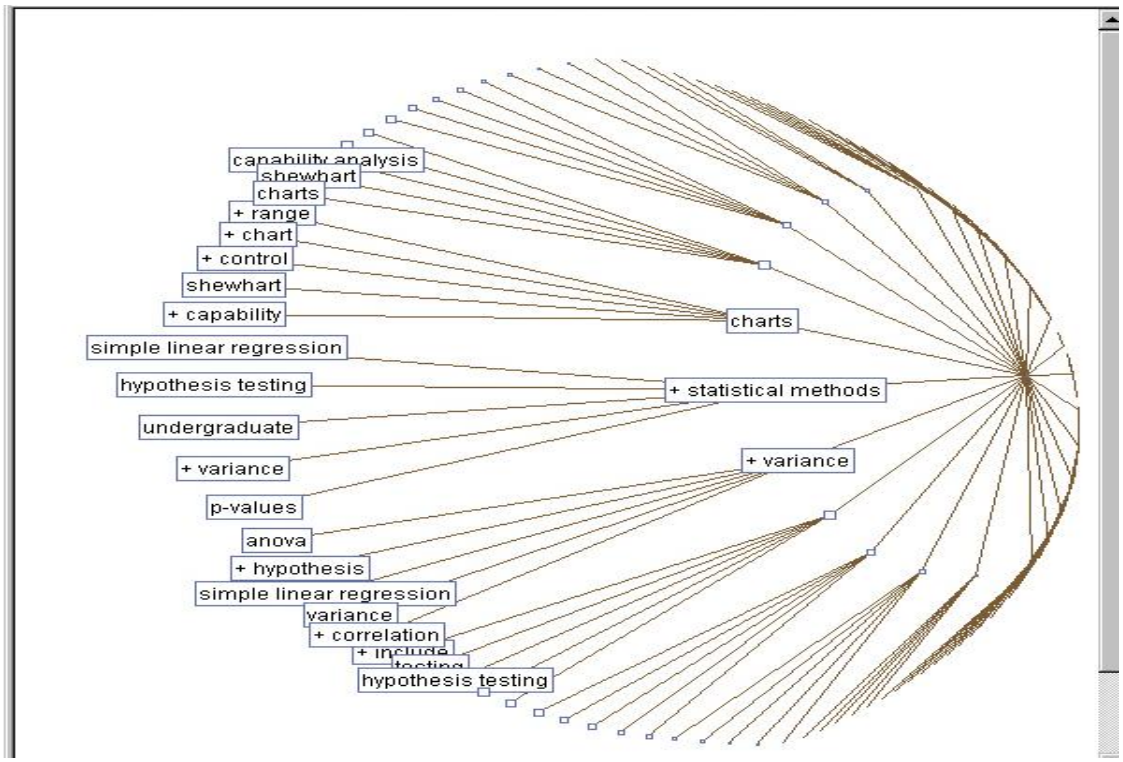

Figure 4: Concept Links in SAS® Text Miner showing both first and second order terms (Woodfield [16])

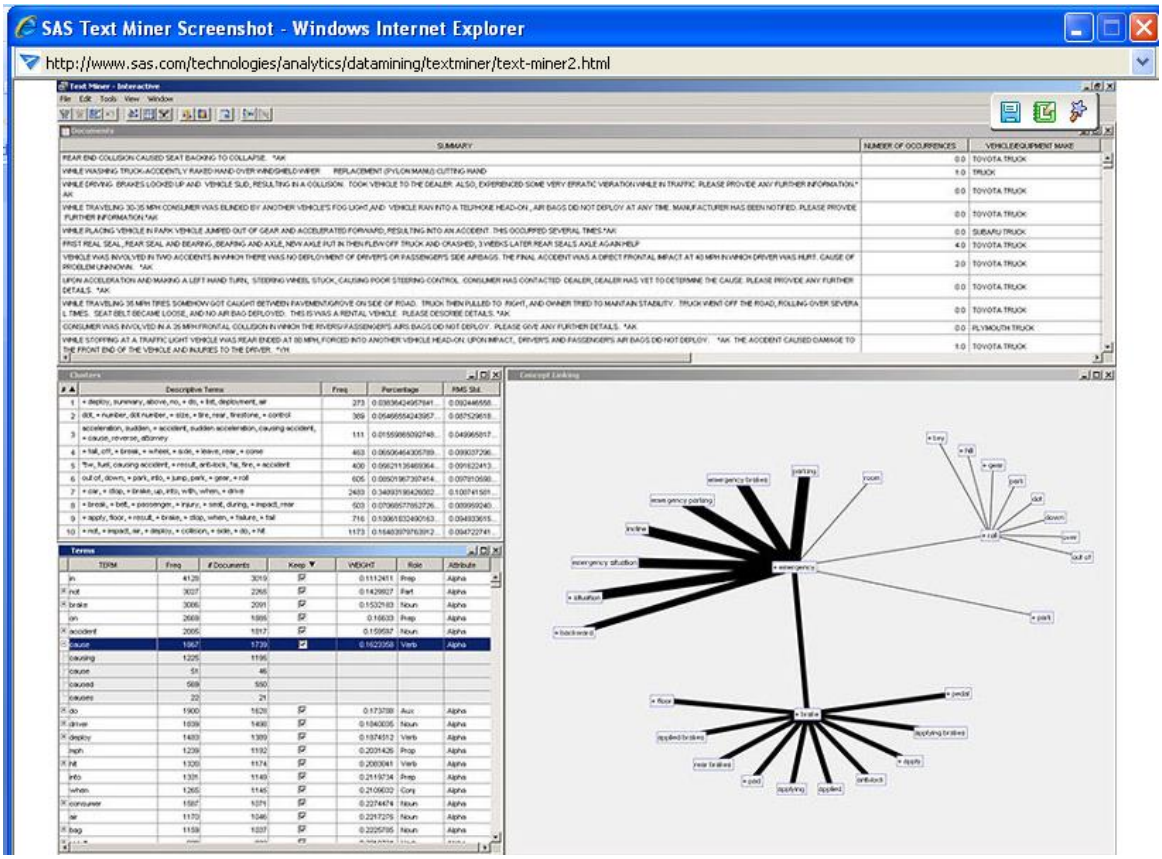

Figure 5: Screen shot of SAS® Text Miner showing linkage and frequency of terms (SAS [10])

The workspace of SAS® Text Miner was constructed with a data icon of selected data, shown in Figure 2. Figure 3 shows first order link analysis results of using SAS® Text Miner, which includes the text for each document and the respective weight for each term and also the concept linking figure, and Figure 4 shown the corresponding for both first and second-order terms (Woodfield [16]). Figure 5 from (SAS [10]) shows linkage and also frequency of terms by the thickness of the connecting lines.

Web-based text mining of hotel customer comments using SAS® Text Miner would be analogous to that as visualized by Figures 3, 4 & 5 and that as can be compared to those of Figures 6, 7 & 8 of Megaputer PolyAnalyst® 6.0 below.

Megaputer PolyAnalyst®

The new release of PolyAnalyst® 6.0 includes text mining and specifically new features for Text OLAP (on-line analytical processing) and taxonomy-based categorization which is useful for dealing with large collections of unstructured documents such as tracking the number of known issues in product repair notes and customer support letters.

The hotel customer database is as that of actual hotel customer written comments that was provided to us by Metaputer Intelligence Incorporated in a web-based conference call with

Sergei Ananyan in summer 2007 as supported by Arkansas State University Summer Research Grant for research proposal written by Zhang and Segall (2006).

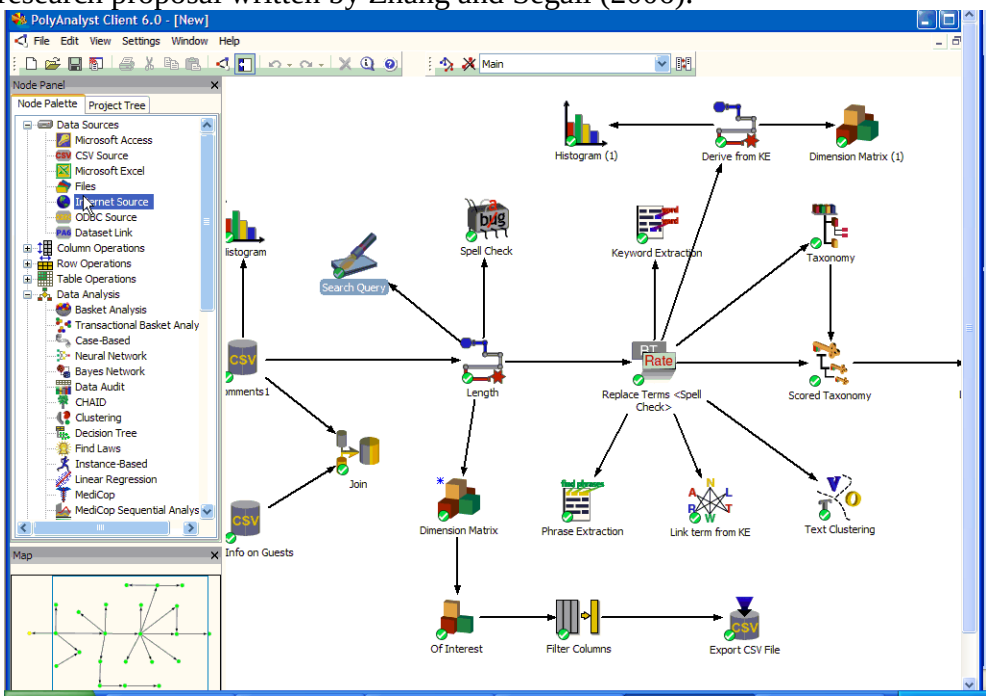

Figure 6: Workspace of Megaputer PolyAnalyst®

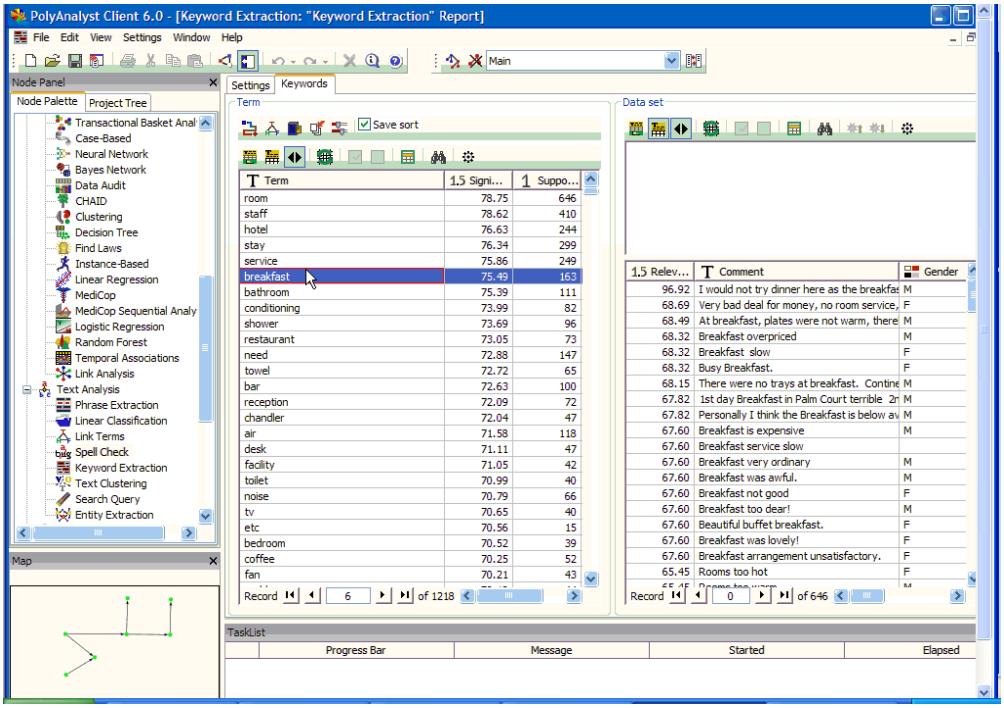

Figure 7: Keyword Extraction Report in Megaputer PolyAnalyst®

PolyAnalyst® provides simple means for creating, importing, and managing taxonomies, and carries out automated categorization of text records against existing taxonomies, for examples, applications to executives, customer support specialists, and analyst so that executives are able to

make better business decisions upon viewing a concise report on the distribution of tracked issues.

This paper provides several figures of actual screen shots of Megaputer PolyAnalyst® 6.0 for text mining of hotel customer survey data. These are Figure 6 for workspace of text mining of Megaputer PolyAnalyst®, Figure 7 for key word extraction window for the word “breakfast” from customer written comments of web-based hotel customer survey, Figure 8 for link analysis report with nodes of key words from hotel customer web submitted comments. Megaputer PolyAnalyst® can also provide screen shots with drill-down text analysis and histogram plot of text analysis (Figure 9) also of key word nodes of these hotel customer web submitted comments.

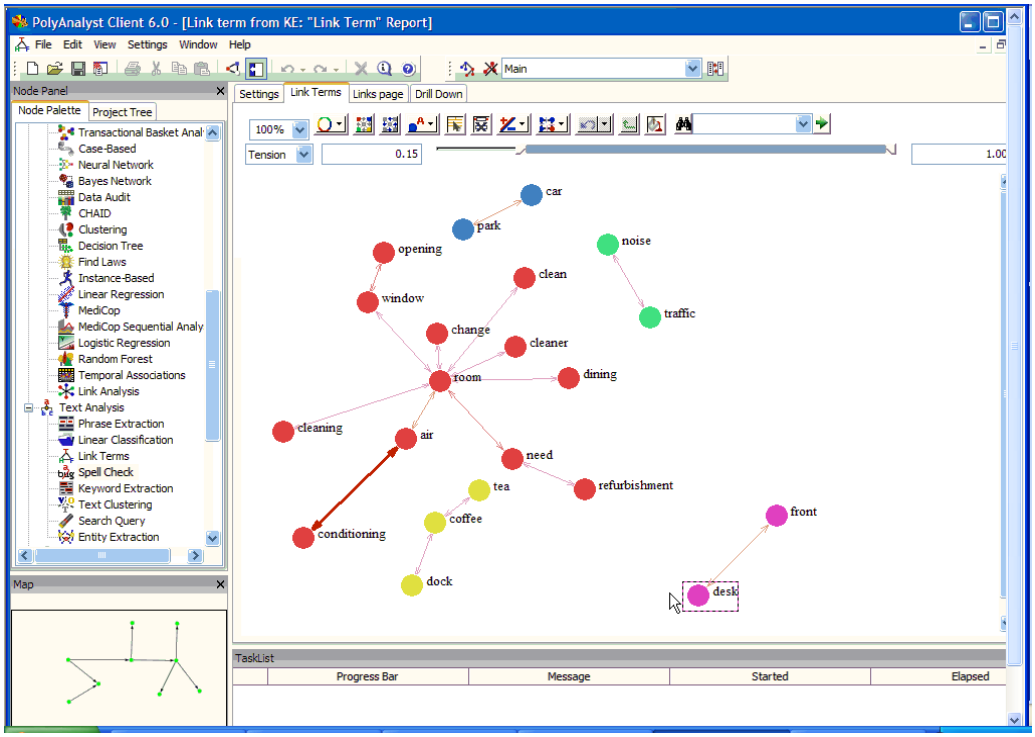

Figure 8: Link Analysis Report in Megaputer PolyAnalyst®

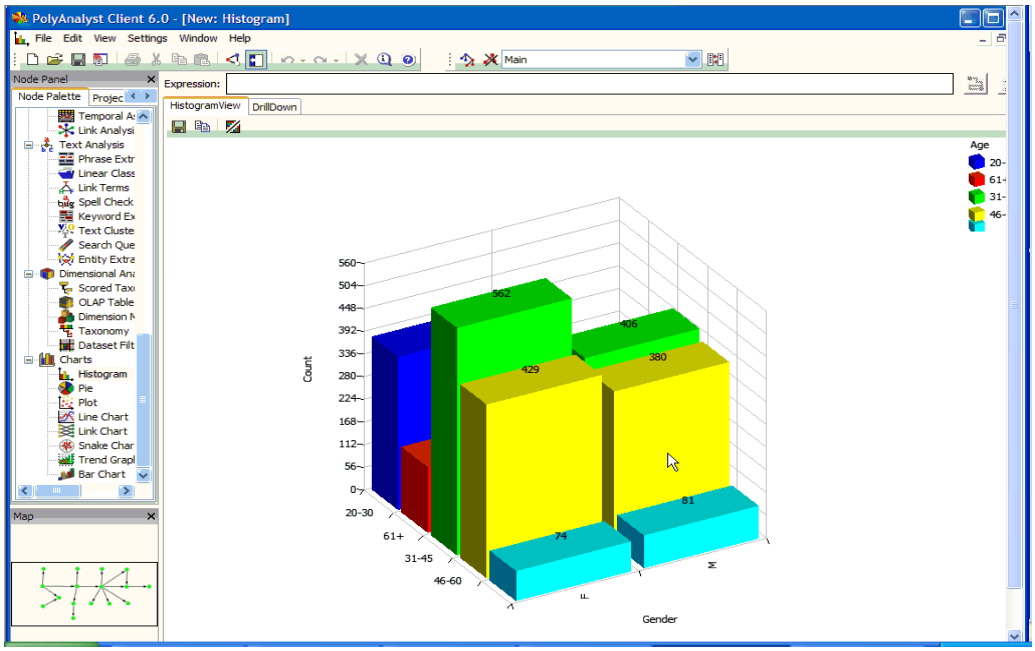

Figure 9: Histogram Plot of Text Analysis Dataset
in Megaputer PolyAnalyst®

CONCLUSIONS

This paper has reviewed the text mining literature and shown that text mining is a new and expanding area of research with web-based applications of new software available for needs that were unthinkable a decade ago. The comparison and screen shots of SAS® Text Miner and Megaputer PolyAnalyst® are helpful for organizations or individuals to adopt this technology, as specifically shown for web based hotel customer survey data. SAS® Text Miner had advantages over Megaputer PolyAnalyst® in being a integrated software specifically for text mining purposes that can be purchased separate of data mining software, however it should be noted that it comprises a drop-and-drag icon on the toolbar that is also housed within data mining software of SAS® Enterprise Miner as shown in Figure 2.

Concepts links of SAS® Text Miner as shown in Figures 3 and 4 tend to provide more depth and insight then those of link analysis reports of Megaputer PolyAnalyst® as shown in Figure 8. The key word extraction capabilities of Megaputer PolyAnalyst® such as shown in Figure 7 for the hotel customer comment database is an extremely useful feature that appears to be more simplistic and subsequently more useful than the linkage and frequency of terms of SAS® Text Miner as shown in Figure 5.

Megaputer PolyAnalyst® was useful in categorizing hotel customer written comments using text mining and its Link Analysis showed visual relationships between key terms of these comments that was more compact and colored-coded than those using SAS® Text Miner. Figure 7 shows written comments for breakfast with comments such as “breaksfast service slow”, “breakfast very ordinary, and “breakfast overpriced”. Figure 8 of Megaputer PolyAnalyst® showed links between written comments of hotel customers between “room”, “dining”, “cleaning” and “window.” Figure 8 link diagram using Megaputer PolyAnalyst® avoids the use of another

figure for second-order terms as needed using SAS® Text Miner as shown in Figure 3. An example of a second-order term in Figure 3 using Megaputer PolyAnalyst® is "room need" and link to "refurbishing." Figure 9 histogram using Megaputer PolyAnalyst® showed plot indicated the highest frequency of hotel customers who wrote written comments were male between the ages of 31 to 45, and the lowest frequency were those above age of 60 for both male and female.

Other software such as ORACLE 10g performs data mining and OLAP (Online Analytical Processing) using inputs of SQL (Structured Query Language) type commands instead of an object-oriented workspace as in both SAS® Text Miner and Megaputer PolyAnalyst® as discussed in this paper. Hence using either of the two discussed software in this paper are advantageous not only for web-based hotel customer written comments but also for other types of data.

Future trends are that text mining and web-based software will continue to grow in dimensionalities of features and available software. The applications of software for web-based text mining will be extremely diverse ranging from uses in customer survey responses as shown in this paper to drill-downs in medical records or credit or bank reports. Future directions of this research is to investigate and compare the applications of other web-based text mining software, and also the future and new developments in web-based text mining as specifically applied to hotel or other types of customer surveys.

ACKNOWLEDGEMENTS

The first two authors want to acknowledge the fact that this research was a result of previous research that was supported by a 2007 Summer Faculty Research Grant as awarded by the Arkansas State University (ASU) College of Business (CoB). The authors also acknowledge the technical support provided by the software manufactures of SAS Institute Incorporated of Cary, NC USA and Megaputer Intelligence Incorporated of Bloomington, Indiana USA without whose help this paper would not have been able to been written.

REFERENCES

1. Amir, A., Aumann, Y., Feldman, R. and Fresko, M., Maximal association rules: a tool for mining associations in text, *Journal of Intelligent Information Systems*, 25(3), (2005), 333–345.

2. Crowsey, M.J., Ramstad, A.R., Gutierrez, D.H., Paladino, G.W., and White, K.P., An Evaluation of unstructured text mining software, *IEEE Systems and Information Engineering Design Symposium*, April 27, University of Virginia, Charlottesville, VA, http://www.sys.virginia.edu/sieds07/papers/ SIEDS07_0007_FI.pdf, (2007).

3. Fan, W., Gordon, M., and Pathak, P., Genetic programming-based discovery of ranking functions for effective web search, *Journal of Management Information Systems*, 21(4), (2005), 37-56.

4. Hersh, W., Evaluation of biomedical text-mining systems: lessons learned from information retrieval, *Briefings in Bioinformatics*, 6(4), (2005), 344-356.

5. Liang, J. W., Introduction to Text and Web Mining, Seminar at North Carolina Technical University, Retrieved form http://www.database.cis.nctu.edu.tw/seminars/2003F/TWM/slides/p.ppt, (2003).

6. PC AI Buyer's Guide, *PC AI Magazine,* vol. 18, n. 3, pp. 62-74 http://www.pcai.com/Paid/Issues/PCAI-OnlineIssues/18.3_OL/New_Folder/T7L2HS/18.3_PA/PDF_18.3/PCAI_62-75-18.3-Buyers-Guide.pdf, (2004).

7. Robb, D. Text mining tools take on unstructured data, *Computerworld*, June 21, http://www.computerworld.com/printthis/2004/0,4814,93968,00.html, (2004).

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

9. Saravanan, M., Reghuraj, C., and Raman, S., Summarization and categorization of text data in high-level data cleaning for information retrieval, *Applied Artificial Intelligence*, 17, (2003), 461–474.

10. SAS, *Text Mining with SAS Text Miner* (2008), http://www.sas.com/technologies/analytics/datamining/textminer/screenshots.html

11. Scherf, M., Epple, A., and Werner, T., The next generation of literature analysis: integration of genomic analysis into text mining, *Briefings in Bioinformatics*, 6(3), (2005), 287-297.

12. Seewald, A., Holzbaur, C., and Widmer, G., Evaluation of term utility functions for very short multidocument summaries, *Applied Artificial Intelligence*, 20, (2006), 57–77.

13. Srinivasan, P., Text mining: Generating hypotheses from MEDLINE, *Journal of the American Society for Information Science and Technology*, 55(5), (2004), 396-413.

14. Turmo, J., Ageno, A., and Catala, N., Adaptive Information Extraction, *ACM Computing Surveys*, 38(2), (2006), 1-47.

15. van Geiist, D., Potharst, R. and van Wezel, M., A support system for predicting eBay end prices, *Decision Support Systems*, 44(4), (2008), 970-981.

16. Woodfield, Terry, *Mining Textual Data Using SAS® Text Miner for SAS9 Course Notes*, SAS® Institute, Inc., Cary, NC, (2004).

17. Yang, H. and Lee, C., Automatic Category Theme Identification and Hierarchy Generation for Chinese Text Categorization, *Journal of Intelligent Information Systems*, 25(1), (2005), 47–67.

18. Yoon, B., Phaal, R. and Probert, D., Morphology analysis for technology roadmapping: application of text mining**,** *R & D Management,* 38(1), (2008), 51-59.

19. Zhang, Q. and Segall, R.S., "Further Continuation of Research on Application of Data Mining Techniques in Knowledge Discovery: an In-Depth Investigation on Algorithms and Heuristics" proposal submitted to and funded by 2007 Arkansas State University (ASU) College of Business Summer Research Grant Committee (2006).

20. Zhang, Q. and Segall, R.S., Using data mining for forecasting data management needs, Chapter XXI in *Handbook of Computational Intelligence in Manufacturing and Production Management*, Information Science Reference, Hersey, PA, ISBN 978-159904582-5, (2008).