Text Mining e-Complaints Data From e-Auction Store With Implications For Internet Marketing Research

Kuan C. Chen, Purdue University Calumet, USA

ABSTRACT

This study seeks to analyze the effectiveness of the text mining process. Complaint forums on various consumer report websites will be analyzed using text and data mining software. Data from feedback forums will be compiled and analyzed using a text miner software program. The relationships and patterns among keywords and their associations will be cluster-analyzed to gain a deeper understanding of the data. A case study will also be conducted to assay the effectiveness of text mined. The data of Internet complaint forum, http://www.planetfeeback.com, will be text mined. The decision to use an Internet complaint forum as the case subject was made because of its easy access and reputation as storage medium for large sources of data. The main goal of this study is to gauge the effectiveness of text mining. The complaint forum will be text mined to find relationships. The results will then be analyzed and then interpreted to determine the effectiveness of the text and data mining process.

Keywords: Text data mining, Cluster analysis, Association rules, e-Auction.

INTRODUCTION

When people talk about marketing research, they are usually referring to quantitative research. This includes such applications as mail surveys, in-person interviews, and telephone surveys. Less frequently used research methods such as focus group, by contrast, acquires qualitative data that are less generalizable, hence less useful for mass marketing. The introduction of information technology into marketing, especially the internet, significantly changed that orientation.

Equipped with the internet and databases, marketing experts have come a long way from targeting their customers with cold entries and figures in the past to interacting with customers in natural language and in real time now. Stored usually in data warehouses, these communication records are goldmines for any company leaders who are attempting to profile their customers' minds and deeds in order to consolidate business relationships with them. In general, data mining comprises three major techniques: data analysis, system optimization and computer applications (Romeu 2001). Consequently, fields of research that utilize any or all of these techniques can be linked to data mining.

Text data mining (TDM) is a natural extension of data mining (DM) (Hearst 1999), and follows steps similar to those in DM (Groth 2000). In essence, the researcher still is required to transform the text data into some organized numerical form so that a computer program can be used so explore the data. The qualitative difference in text mining, however, is that TDM processes data from natural language text rather than from structured databases of facts. What can be inferred, then, from such an endeavor? What if the natural language is in a form other than the English language?

DATA MINING

The advent of data mining has improved the effectiveness of gathering information from raw data. Data mining involves representing information in a new fashion so previously undetected relationships, tendencies, patterns, and trends can be identified (Chaffey, 2004, pp. 572). It is a process that finds hidden and additional value from data. It packets this new data in a fashion that allows end users to work with current, up-to-date, and suitable information (Chittaluru, Hunter, and Thompson, 2004,). Data mining software uses complex computational processes such as clustering, data cleaning, decision trees, artificial intelligence, neural networks, textual distance measurement, and regressions to thoroughly probe and outline large data sets involving multiple sources. The key and main reason for the use of data mining is its ability to identify relationships and trends that are not readily seen by the naked eye or other analytical devices. Data mining also allows the user to input his or her unqiue variables and specify search criteria, weighting those variables and criteria deemed to be most important to increase the results' accuracy and relevance. (McCue, Stone, and Gooch, 2003).

TEXT MINING

Text mining, data mining's sister practice, is often used in tandem with the aforementioned process. Whereas data mining is often used on databases, text mining is more flexible in its use. Text mining software does not require data to be in database form for analysis like data mining software; rather it may be used on raw unformatted blocks of text. Companies use text mining software to draw out the occurrences and instances of key terms in large blocks of text, such as articles, Web pages, complaint forums, or Internet chat rooms and identify relationships (Robb, 2004). The software converts the unstructured data formats of articles, complaint forums, or Web pages into topic structures and semantic networks which are important data drilling tools. Often used as a preparatory step for data mining, text mining often translates unstructured text into a useable database-like format suitable for data mining for further and deeper analysis (Cerrito, 2004).

ISSUES SURROUNDING DATA AND TEXT MINING

Although the two practices work very closely together, text mining faces issues that data mining does not. Data mining is an older and more established practice, and text mining is still finding its footing in the industry. Text mining does have unique problems. However, both do share some of the same problems. Both data and text mining tools need to be valid. If the software is valid, it should not identify a relationship or pattern that has no bearing on the purpose of the mine. Both involve complex software programs and require programmers with extensive experience with the software programs (King and Linden, 2002).

One of the most difficult aspects surrounding information mining is result interpretation. It is a difficult aspect because result interpretation is dependent on the skill of the software technician. The greater the skill of the technician, the more effective the data or text mine. Even if a skilled technician is very successful with the data mine, the data mine still may not reach its potential as the user may not have the analytical skills to interpret the results of the text mine.

As in most other projects, data and text mining projects may fall victim to scope creep and change in project definition during the life of the project. A data or text mine needs a clear definition and success criteria established long before the mine begins. An unclear or changed definition would cripple an information mine from the start. (King and Linden, 2002).

One of the main reasons for the slow adoption of text mining programs is cost. Text mining programs are very expensive, often costing thousands of dollars for a single license with multiple licenses required. It is a steep investment in a young industry. The issues and problems unique to text mining software can be attributed to the fact that text mining as a process is still in its infancy. Text mining may also have a limited scope of use. While text mining is invaluable in information mining, information mining is not as important to certain functional business units as it is to others. While text mining would be valuable to marketing or advertising departments, accounting or finance would not share the need.

Other reasons for the lag of text mining adoption are related to its sister process, data mining. Data mining software have been around for several years longer than text mining software and have had the time to be modified and adapted. Text mining software does not yet match the effectiveness of data mining tools. This hindrance has made text mining a niche tool not yet ready for the mainstream. The lack of programmer skill and familiarity using text mining programs also contributes to the limited reception of text mining. As its shortcomings are addressed with improved software editions, text mining is gaining in acceptance and popularity (Robb, 2004).

CASE BACKGROUND

E-commerce has burst onto the scene the past ten years to become an important revenue stream for businesses. It certain instances, it has become the main revenue stream for companies. The era has also been characterized by increased communication. "E-word of mouth" is much the same as regular word of mouth; with the difference being that e-word of mouth is spread via an electronic medium. Whether it is email, instant messenger services, online articles, e-message boards, or feedback forums, e-word of mouth detailing customer experience is spread. e-WOM can be communicated to a greater number of people than normal word of mouth and at a faster pace. While a person may retell a story concerning an experience with a business to several friends in regular word of mouth. In e-WOM , that story is posted on a feedback forum available to not only family and friends of the poster, but to the online community as well. The story would be available to many more people than the person could ever physically tell. The reach of the story is much greater because of e-WOM.

RESEARCH PROCESS

An independent e-feedback forum, Planet Feedback (www.planetfeedback.com) will be text mined using TextAnalyst 2.3 by Megaputer Intelligence to gauge the effectiveness of the text mining process. . Planet Feedback is a large online feedback forum featuring thousands of comments, complaints, and compliments from consumers about various companies. The effect of e-WOM is felt as some companies are featured by Planet Feedback for their attention to the comments posted on the forums.

The main aim of this research is to conduct a successful text mining analysis. This study seeks to analyze the ease-of-use and effectiveness of the text mining process and software. The results, particularly the semantic network and topic structure will also be studied to learn the reason for the complaints. The category selected from Planet Feedback to be text mined is the credit card industry, more specifically complaints against "Billing/Payment" policies. The credit card industry was selected because it provided a large bank of complaints enveloping numerous companies at no cost. The complaints found with credit card companies' billing/payment plans will also be analyzed. The results will be used to assay the behavior of the complainer. The results of the text mine will be studied, evaluated and understood.

CASE STUDY

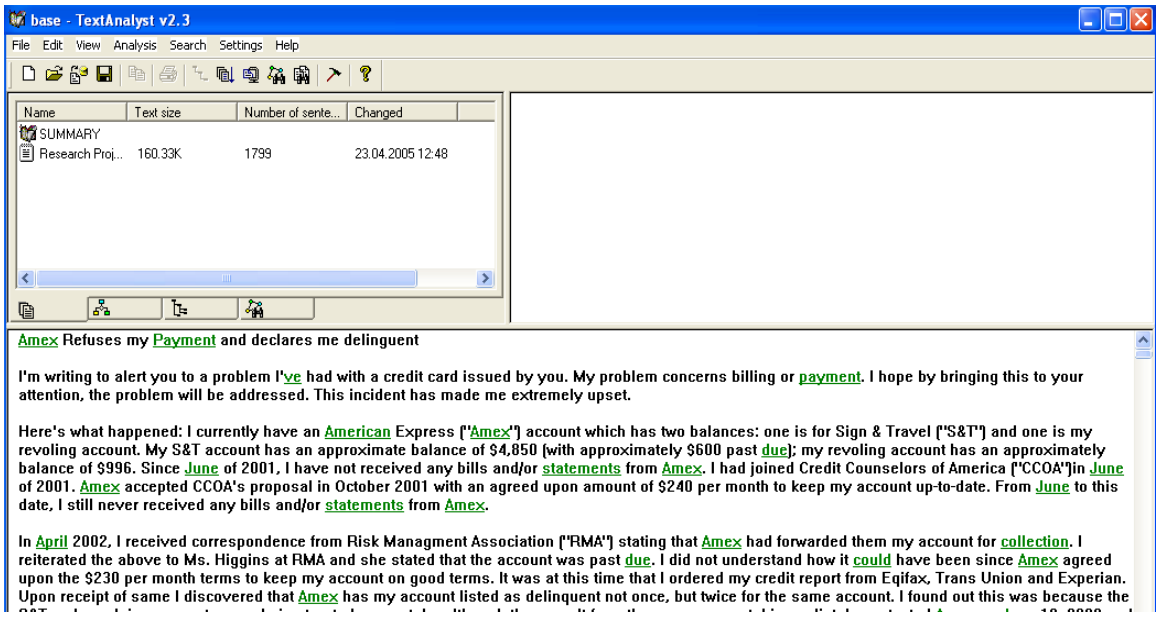
The data for the text mine was gathered from the website Planet Feedback. Planet Feedback is website that provides a feedback forum for consumers to post questions, concerns, complaints, or compliments about an experience with a product, brand, or company. The website features numerous industry categories ranging from airlines to online department stores to pet products manufacturing to online toy stores to water utilities.

Planet Feedback has a dual purpose. First, it serves as a popular medium for consumers to voice their opinions and experiences about a particular company as well as reference for consumers who are considering purchasing an item from a particular company. Consumers are able to view the one hundred most recent complaints at no charge. Second, Planet Feedback allows companies another outlet to view consumer feedback. The feedback is organized first by industry. Once the industry is selected, the feedback is organized by the date, company, type of comment, and category. Planet Feedback charges companies for the use of their information.

The category selected for data mining in this study is the credit card industry. More specifically, the complaints levied against several credit card companies and their billing/payment system will be text and data mined. The industry, companies, and complaint category were selected because they provided a solid sample size at

no cost. The companies examined in the study are the American Express Company, Fleet Boston Credit Card Services, Aspire Visa Card Services, General Electric Capital Corporation, Aria Visa, GM Card, AT&T Universal, Discover Card, Next Card, Inc., and MBNA, There is no particular bias or aversion to any of the companies included in this study, the companies just happened to have complaints levied against them. It was possible to select 100 complaints in the billing/payment complaint category free of charge to text and data mine.

Figure 1: Example of TextAnalyst 2.3



The initial size of the document entered into TextAnalyst 2.3 (see Figure 1 above) was 160.33 KB. The document contained 1799 sentences. TextAnalyst 2.3 preprocesses the data to remove supplementary words such as "a", "an", and "the." The words are removed because they have no semantic meaning. They are dead weight to a text mine and considered worthless. The preprocessing also identifies word stems and separates them from prefixes and suffixes. The text mine wants to analyze stems alone to provide a more lucent and clear picture of their relationships within the document.

Preprocessing is part of the process the document undergoes before it is text mined. It can be likened to a patient being scrubbed and prepared for surgery. Preprocessing prepares the document for its statistical analysis. The text miner assesses the occurrences, distances, and relationships between words. It results in the creation of a tree-like structure that contains the weights of both words sharing the joint occurrence. TextAnalyst 2.3 also calculates statistical weights for the individual terms and relationships.

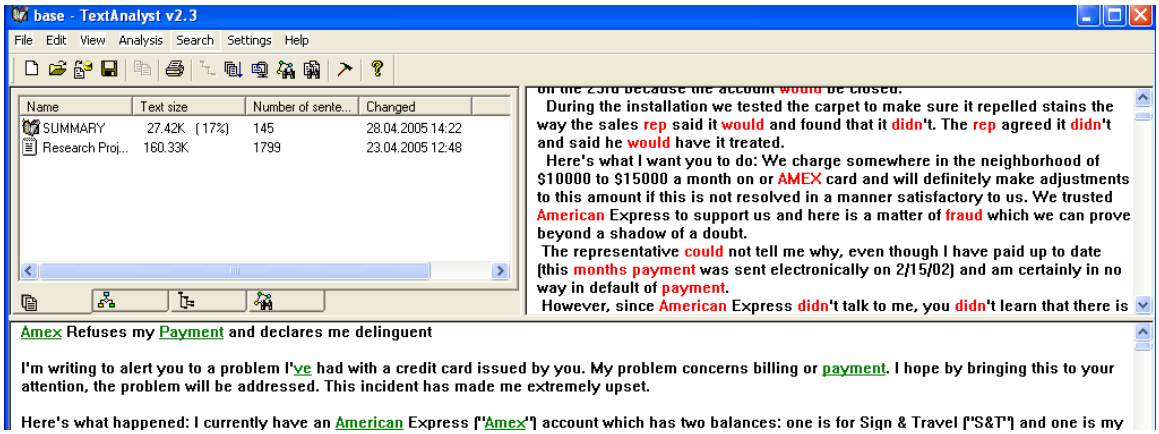
The final component of the analysis preparation is renormalization. In this final stage, the text miner gives the statistical weights a final adjustment. The readjustment of the weights changes them from statistical weights to semantic weights. The semantic network is constructed during renormalization and the semantic relationships are finalized. After renormalization is finished, the semantic network is ready for use.

TextAnalyst 2.3 creates a semantic analysis and summarization of the article submitted for text mining.

The summary (see Figure 2) is a condensed version of the document containing only the most pertinent information in the document. To view the summary, look at the upper most window on the left. In this instance, the summary of the file was 27.42 KB (roughly 17% of the size of the original) and contained 145 sentences as opposed

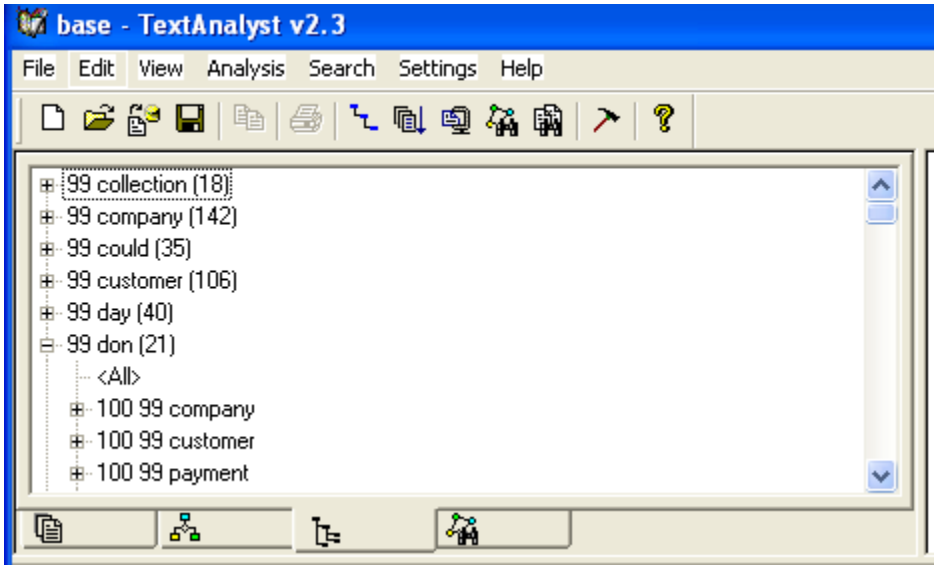
to the 1799 of the original. It allows the analyst to gain an understanding of the data without having to pore through the entire document or set of documents. The window on the right contains the physical summary.

Figure 2: (Example of Summary)



The semantic analysis is a very useful tool and one of the most important aspects of text mining. It is a tree structure of the concepts of the article and depicts the relationships identified in the text. It is a visual presentation of the relationships the text miner identified in the document.

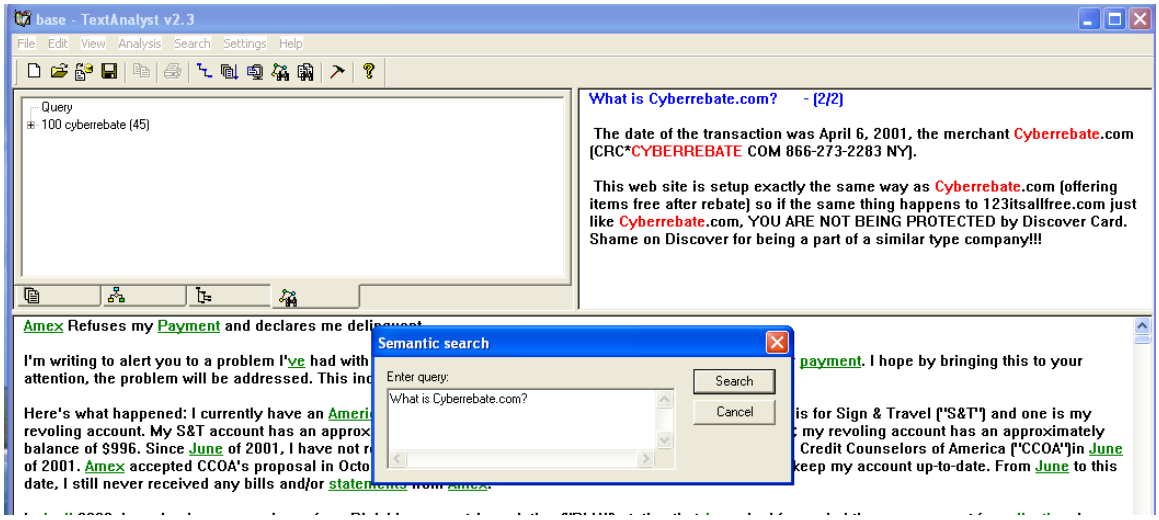
Figure 3: (Example of Semantic Network)



The semantic network (see Figure 3 above) is the foundation of all future analyses. It is a list of topics and their semantic weights. The semantic weight is the measure of probability of the words contextual importance. The semantic network lists the words alphabetically and includes the number of times the word appears in the document. Among the words TextAnalyst 2.3 placed in the semantic network were "collection," "payment," and "statement". All three terms had a semantic weight of 99. The semantic network allows a further investigation into each word

appearing in the network. Collapsible lists allow a further investigation into each word. The list contains the relationships between the key word of the semantic network and other words appearing on the list. Each concept is preceded by a pair of numbers. The first set of numbers is the semantic relationship between the parent and concept. The second number is the individual semantic weight of the concept. For example, the previously mentioned "statement" is a parent of the term "customer." The two share a semantic relationship with a strength of 55 while the semantic strength of the "customer" term is 99. There are well over 50 concepts in the semantic network, each with a varying semantic strength. Several terms have a semantic strength of 99, while the words towards the bottom of the semantic network have semantic strengths of less than 20.

Figure 4: (Query Example)



TextAnalyst 2.3 offers a query function (see Figure 4 above). It is very useful for locating and identifying particular information. It becomes a tool of greater importance the larger the size of the document as navigation becomes more difficult. Queries may be entered in conversational English and results will be given based on the content of the query rather than focusing on keywords. The results are hyperlinked to their appearance in the text. This is done to allow the view of the text location of the query result.

Turning attention to the text mine, the topic structure, summary, and semantic network provide an outline of the material. The summary provides an overall overview of the document. By glancing over the summary, it appears that most complaints deal with issues surrounding billing statement errors and the implementation of a program with cyberrebate.com. The semantic network also supports these findings. The words "payment," "billing," and "issue" all have semantic weights of 99 and share strong semantic weights with other concepts throughout the semantic network. Much more information can be extrapolated from the text mine, it is dependent the needs of the analyst.

Text mining software is important because it establishes a semantic network of a document. Another key use of text mining software involves the preparation of a document for data mining. Text mining takes a standard document and converts it into a form suitable for the ever more detailed data mine. Without text mining software, a document in .txt, .pdf, or .doc form would not be able to be data mined. The document would not have to be converted manually. The manual conversion is much slower and not nearly as accurate or effective as a text mining program. Whereas information can be learned from a text mine, a data mine is an even more useful and detailed information gathering tool.

SUGGESTIONS FOR FUTURE RESEARCH

In terms of this particular text mine, many of the complaints deal with issues surrounding billing statement errors and the implementation of a program with cyberrebate.com. The semantic network also supports these findings. The words "payment," "billing," and "issue" all have semantic weights of 99 and share strong semantic weights with other concepts throughout the semantic network.

Text mining, although still in its infancy, can provide a wealth of information from raw text. By analyzing these complaints, a greater understanding of the reason for complaints was learned. The text mining software broke down the document into a semantic network that allowed for the complaints to be analyzed in a way that could not be done without the program. By studying the semantic network, one can learn the general tone of the complaints, reasons for complaining, and the common words used and their relationships to other words in the text via semantic weight.

The text mining process was interesting and very thorough. The ease of use of the TextAnalyst 2.3 program was tremendous. The product advertised itself as easy-to-use and more than lived up to the billing. The program also came with a manual that contained detailed instructions as well as a simulated text document that was part of a tutorial. The difficulties expected learning the program were much less than originally anticipated. Although this was a simple case with a program famed for its ease-of-use, it appears text mining programs have greatly improved their initial ease-of-use problems. The surface has barely been scratched on this document. The text miner program has advanced options that allow the data to be drilled down to an even more detailed level.

The issue of cost still plagues text mining software programs. The initial price for TextAnalyst 2.3 was well over $1,500. The purchase was made possible due to an educator's license that reduced the price to nearly one-fourth of the original quote. Even with the discount, most of the grant money was used on the purchase of the text miner and service contract.

Text mining can be an extremely useful process. Much can be learned about a company, its customers, merits, and problems through a successful text mine. It identifies trends, strengths, and weaknesses. It is a wonderful tool if it is used correctly and with reasonable expectations. The software could be used by anyone or organization, but its steep price tag only allows large businesses with a focused effort on information mining to reap its benefits.

AUTHOR INFORMATION

Kuan C. Chen is Head and Professor of Department Information Systems in the School of Management at Purdue University Calumet in Hammond, Indiana. Dr. Chen has extensive experience in MIS research topics. He has authored numerous journal papers on topics varying from project management to Information Technology (IT) economics. He has also been a contributing author on several books and a technical editor on numerous books and journal articles. Dr. Chen maintains an active Web development and database consulting practice in both the U.S. and Taiwan. He has a Ph.D. in Information Systems and another Ph.D. in Applied Economics, both from Michigan State University.

REFERENCES

1. Ahonen-Myka, H. "Discovery of frequent word sequences in text," Pattern detection and discovery. In D. J. Hand, N. M. Adams & R. J. Bolton (Eds.), Springer-Verlag, Berlin, Berlin, Germany, pp.180-189, 2002.
2. Allen, P. "Business complaint sites give consumers room to vent," *Business Journal* (Central New York), Vol. 15, No. 8, pp. 1-2, 2001.
3. Andreassen, T. W. "What drives customer loyalty with complaint resolution?" *Journal of Service Research*, Vol. 1, No. 4, pp. 324-332, 1999.
4. Andreassen, T. W. "From disgust to delight: Do customers hold a grudge?" *Journal of Service Research*, Vol. 4, No. 1, pp. 39-49, 2001.

5. Blodgett, J. G., Hill, D. J., & Tax, S. S. "The effects of distributive, procedural, and interactional justice on postcomplaint behavior," *Journal of Retailing*, Vol. 73, No. 2, pp. 185-210, 1997.
6. Brown, S. W. "Service recovery through IT: Complaint handling will differentiate firms in the future," *Marketing Management*, Vol. 6, No. 3, 25-27, 1997.
7. Cerrito, Patricia. Inside Text Mining. Retrieved March 24, 2005 http://wilsontxt.hwwilson.com/pdffull/06619/275n6/gs9.pdf, March 24, 2005
8. Chittaluru, P., Hunter, C. L., Thompson, J. F. Bridging Data Islands Efficiently. March 24, 2005 http://wilsontxt.hwwilson.com/pdffull/01036/q63qt/7su.pdf
9. Delgado, M., Martín-Bautista, M. J., Sánchez, D., & Vila, M. A. "Mining text data: Special features and patterns," Pattern detection and discovery. In D. J. Hand, N. M. Adams & R. J. Bolton (Eds.), Springer-Verlag Berlin, Berlin, Germany, pp. 140-153, 2002.
10. Eccles, G., & Durand, P. "Complaining customers, service recovery and continuous improvement," *Managing Service Quality*, Vol. 8, No. 1, pp. 68-71, 1998.
11. Estelami, H. "Competitive and procedural determinants of delight and disappointment in consumer complaint outcomes," *Journal of Service Research*, Vol. 2, No. 3, pp. 285-300, 2000.
12. Estelami, H. "Sources, characteristics, and dynamics of postpurchase price complaints," *Journal of Business Research*, Vol. 56, No. 5, pp. 411-419, 2003.
13. Foster, N., & Botterill, D. (1995). Hotels and the businesswoman: A supply-side analysis of consumer dissatisfaction. *Tourism Management*, 16(5), 389-393.
14. Gelb, B. D., & Sundaram, S. "Adapting to 'word of mouse'," *Business Horizons*, Vol. 45, No. 4, pp. 15-20, 2002.
15. Harrison-Walker, L. J. "E-complaining: A content analysis of an Internet complaint forum," *Journal of Services Marketing*, Vol. 15, No. 5, pp. 397-412, 2001.
16. Harrison-Walker, L.J. & Erdem, S. A. "Consumer complaining behavior: The case of the Internet," In J. J. Hartmann & P. Mallette (Eds.), Proceedings of the Twenty-ninth Annual Meeting of the Western Decision Sciences Institute, April 18-21, Maui, Hawaii, pp. 737-740, 2000.
17. Herr, P. M., Kardes, F. R., & Kim, J. "Effects of word-of-mouth and product attribute information on persuasion: An accessibility-diagnosticity perspective," *Journal of Consumer Research*, Vol. 17, No. 4, pp. 454-462, 1991.
18. Huang, J.-H., Huang, C.-T., & Wu, S. "National character and response to unsatisfactory hotel service," *International Journal of Hospitality Management*, 15(3), 229-243, 1996.
19. Jackson, M. Y., & Brown, M. R. "Creative complaint strategies: Letters aren't the only way to express your dismay," *Black Enterprise*, Vol. 31, No. 7, pp. 204, 2001.
20. Kasouf, C. J., Celuch, K. G., & Strieter, J. C. "Consumer complaints as market intelligence: Orienting context and conceptual framework," *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, Vol. 8, pp. 59-68, 1995.
21. King, James and Linden, Orin. "Data Mining Isn't Magical, But It's Not A 'Cookbook' Procedure, Either," (on-line) http://wilsontxt.hwwilson.com/pdffull/01546/edihq/bs0.pdf, November 19, 2003.
22. Laczniak, R. N., DeCarlo, T. E., & Ramaswami, S. N. "Consumers' responses to negative word-of-mouth communication: An attribution theory perspective," *Journal of Consumer Psychology*, Vol. 11, No. 1, pp. 57-73, 2001.
23. Lee, C. C., & Hu, C. "Hotel customers' complaint behavior on www.ecomplaints.com," In G. R. Jennings (Ed.), Proceedings of the 2002 Annual ISTTE Conference, October 10-12, Salt Lake City, UT, pp. 193-197, St Clair Shores, MI: International Society of Travel and Tourism Educators, 2002.
24. Lennon, R., & Harris, J. "Customer service on the web: A cross-industry investigation," Journal of Targeting, *Measurement and Analysis for Marketing*, Vol. 10, No. 4, pp. 325-338, 2002.
25. Levesque, T. J., & McDougall, G. H. G. "Service problems and recovery strategies: *An experiment," Canadian Journal of Administrative Sciences*, Vol. 17, No.1, pp. 20-37, 2000.
26. Liu, R. R., & McClure, P. "Recognizing cross-cultural differences in consumer complaint behavior and intentions: An empirical examination," *Journal of Consumer Marketing*, Vol. 18, No. 1, pp. 54-75, 2001
27. Mangold, W. G., Miller, F., & Brockway, G. R. "Word-of-mouth communication in the service marketplace," *Journal of Services Marketing*, Vol. 13. No. 1, pp. 73-89, 1999.

28. Maxham, J. G., III, & Netemeyer, R. G. "Modeling customer perceptions of complaint handling over time: The effects of perceived justice on satisfaction and intent," *Journal of Retailing*, Vol. 78, No. 4, pp. 239-252, 2002.
29. McAlister, D. T., & Erffmeyer, R. C. "A content analysis of outcomes and responsibilities for consumer complaints to third-party organizations," *Journal of Business Research*, Vol. 56, No. 4, pp. 341-351, 2003.
30. McCue, Colleen, Stone, Emily S., Gooch, Teresa P. "Data Mining and Value-Added Analysis," Available: http://wilsontxt.hwwilson.com/pdffull/03279/1582t/3sx.pdf, March 24, 2005.
31. Montes-y-Gómez, M., Gelbukh, A. F., & López, A. L. "Text mining at detail level using conceptual graphs," Retrieved December 20, 2002, from *http://citeseer.nj.nec.com/531779.html*, 2002.
32. Morris, S. V. "How many lost customers have you won back today? An aggressive approach to complaint handling in the hotel industry," *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, Vol. 1, No. 1, pp. 86-92, 1988.
33. Murray, K. B. "A test of service marketing theory: Consumer information acquisition activities," *Journal of Marketing*, Vol. 55, No. 1, pp. 10-25, 1991.
34. Nasukawa, T., & Nagano, T. "Text analysis and knowledge mining system," *IBM Systems Journal*, Vol. 40, No. 4, pp. 967-984, 2001.
35. Nyer, P. U. "An investigation into whether complaining can cause increased consumer satisfaction," *Journal of Consumer Marketing*, Vol. 17, No. 1, pp. 9-19, 2000.
36. Robb, Drew. Taming Text. March 24, 2005, http://vnweb.hwwilsonweb.com/hww/jumpstart.jhtml?recid=0bc05f7a67b1790e8bd354a88a41ad89a928d23360302a4959035699f17e2ba8a63e2dd032c73f8a7fmt=H

NOTES