Web Mining of Hotel Customer Survey Data

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

and

Qingyu ZHANG
Department of Computer & Information Technology, College of Business
Arkansas State University
Jonesboro, AR 72467-0130 USA
E-mail: qzhang@astate.edu

ABSTRACT

This paper provides an extensive literature review and list of references on the background of web mining as applied specifically to hotel customer survey data. This research applies the techniques of web mining to actual text of written comments for hotel customers using Megaputer PolyAnalyst®. Web mining functionalities utilized include those such as clustering, link analysis, key word and phrase extraction, taxonomy, and dimension matrices. This paper provides screen shots of the web mining applications using Megaputer PolyAnalyst®. Conclusions and future directions of the research are presented.

Keywords: Web mining, hotel customer survey data, Megaputer PolyAnalyst®

1. LITERATURE BACKGROUND

Huang and Chen [3] used data mining techniques to analyze and retrieve customers' traveling patterns from the database in the traveling system built in the World Wide Web environment. Murphy et al. [8] investigated website generated market research data by tracking the tracks left behind by visitors. O'Connor [9] explored the collection of data by the websites of the top 100 hotel brands.

Lau et al. [4] and Lau et al. [5] both discussed text mining for the hotel industry to develop competitive and strategic intelligence. Lau et al. [4] compiled information about Hong Kong hotels and would-be travelers and performed mined customer demographics and attitudes with reasonable accuracy from newsgroup postings.

Olmeda and Sheldon [10] analyzed the potential uses of data mining techniques in Tourism Internet Marketing and electronic customer relationship management. The data mining techniques addressed by Olmeda and Sheldon [10] included customer profiling, inquiry routing, e-mail filtering, on-line auctions, and updating e-catalogues.

Scharlr [12] et al. studied an integrated approach to measure web site effectiveness in the European hotel industry. The study of Scharlr [12] employed a novel method of Web content extraction and analysis to investigate the e-business of competitive travel and tourism. Wober et al [25] discussed the success factors of European hotel web sites. Lau et al. [6] discussed web site marketing for the travel-and-tourism industry. Liu and Wu [7] discussed web usage mining for electronic business applications.

Raghavan [11] surveyed the some of the recent approaches of using data mining in the context of e-commerce. Anderson Analytics [1] investigated web mining in the leisure industry with respect to hotel brands and loyalty programs for the frequent traveler. Stumme et al. [22] discussed the state of the art and future directions of semantic web mining and argue that the two areas of data mining and semantic web need each other to fulfill their goals but the full convergence of these two areas in not yet realized.

Preliminary research of the authors are in the applications of data mining is in Segall [13] Segall and Zhang [20], [21], [22], and Segall et al. [15]; for microarray databases for biotechnology in Segall and Zhang [19] and Zhang and Segall [25]; for four selected data mining software in Segall and Zhang [16]; for text and web mining in Segall and Zhang [14]; for teaching web mining with an overview of web usage mining in Segall

and Zhang [18]. The later reference of Segall and Zhang [17] has implications to web mining usage such as for hotel a customer which is the topic of this paper.

2. SOFTWARE

The web mining software utilized in this research is that of Megaputer PolyAnalyst®. This software is an enterprise analytical system that has unique feature of integrating web mining together with data and text mining. PolyAnalyst® can load input data from web site sources for disparate data sources including all popular databases, statistical, and spreadsheet systems. In addition, it can load collections of documents in html, doc, pdf and txt formats, as well as load data from an Internet web source.

Megaputer PolyAnalyst® has the standard data and text mining functionalities such as Categorization, Clustering, Prediction, Link Analysis, Keyword and entity extraction, Pattern discovery, and Anomaly detection. These different functional nodes can be directly connected to the web data source node for performing web mining analysis.

3. WEB MINING OF SURVEY DATA

Figures 1 to 12 below are screen shots of PolyAnalyst® 6.0 as applied to actual text of written comments of hotel guest as provided by Megaputer Intelligence Inc. Figure 1 shows the workspace created by "drop-and-drag" icons from node palette to workspace for keyword extraction, spell check, replace terms, phase extraction, and others.

Figure 2 shows the source properties window with some of the data for attributes of code, purpose, hotel, region, and comments. Figure 3 shows keyword extraction report with the selected term highlighted of "breakfast" with the corresponding extraction report with relevance index, comment, and gender.

Figure 4 shows "link term" report for the term "room" as being the origin node for the links. Figure 5 shows for the selected expression of "noise and traffic" the corresponding comments, gender, age, code, purpose, and hotel name.
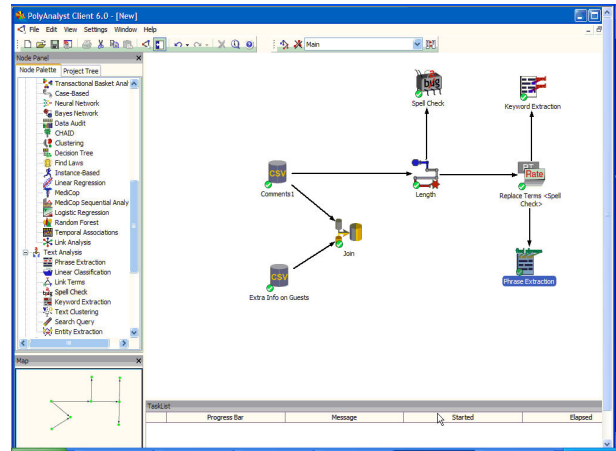

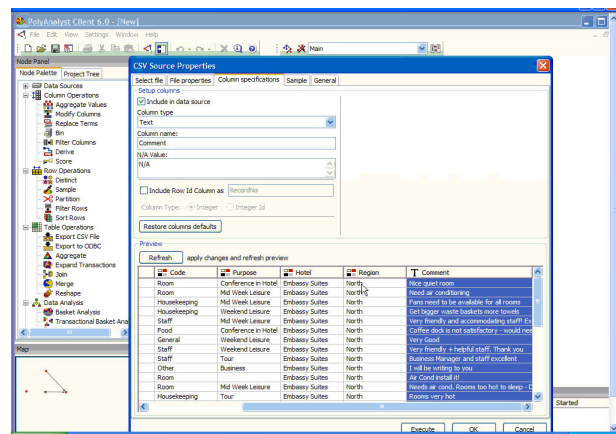Figure 1: Initial workspace of Megaputer PolyAnalyst® for hotel customer survey data

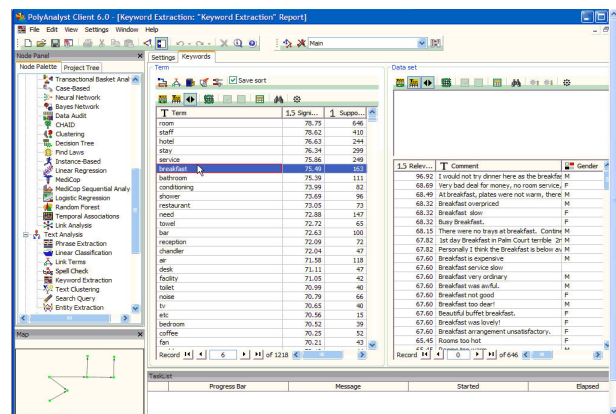Figure 2: Source Properties window of Megaputer PolyAnalyst®

Figure 3: Keword Extraction Report of Megaputer PolyAnalyst®
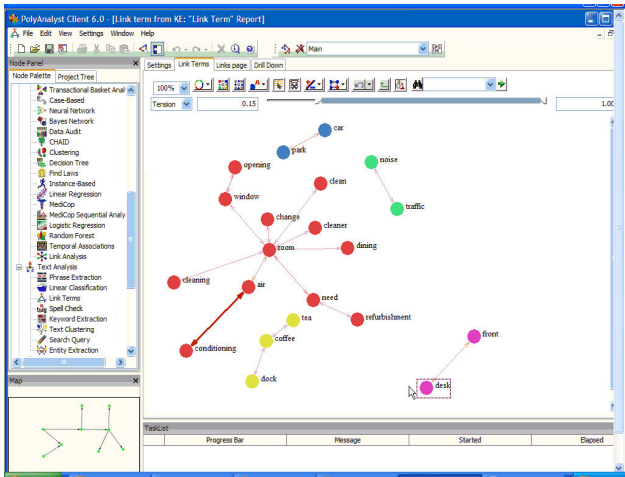
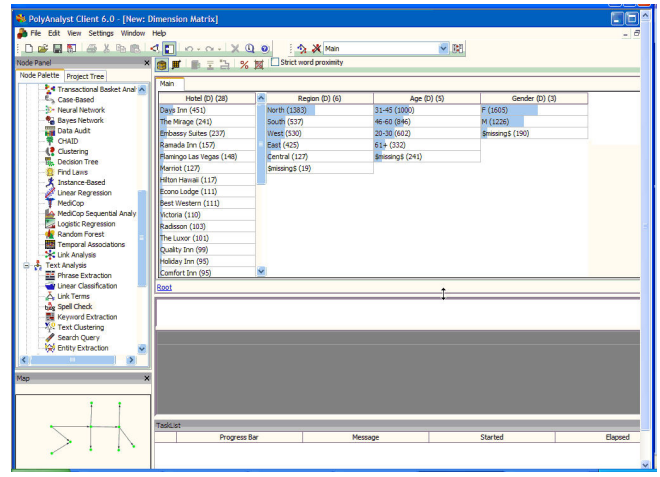Figure 4: Link Term Report of Megaputer PolyAnalyst®

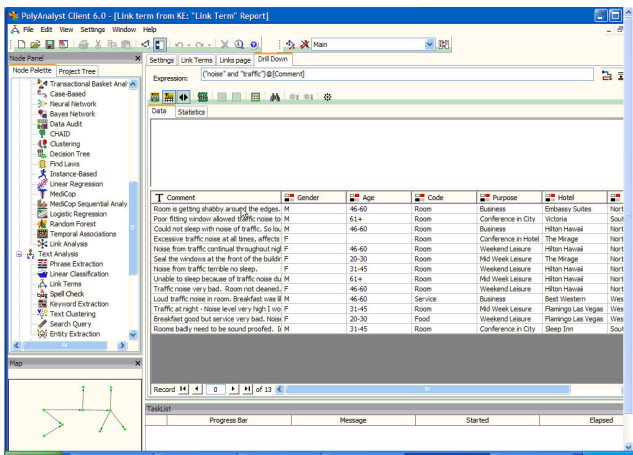Figure 6: Dimension Matrix for hotel data

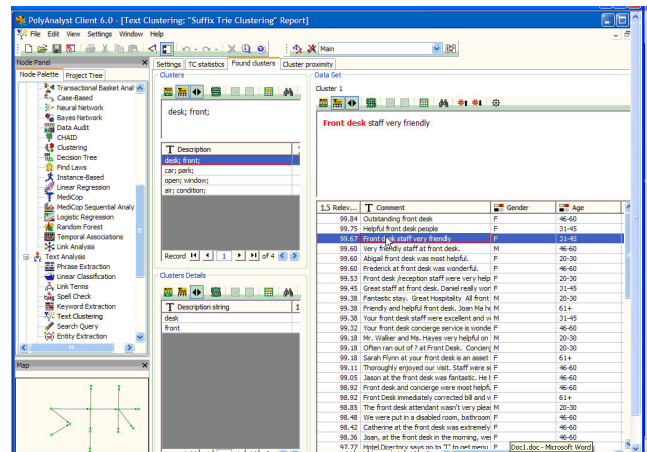Figure 5: Link Term Report for hotel customers' comments of "noise" and "traffic"

Figure 7: Text clustering report of hotel customer comments with "desk" and "front"

Figure 6 shows the new dimension matrix for dimensions of hotel name, region, age, and gender. Figure 7 shows the clustering report with found clusters indicated as "desk" and "front". This clustering report in Figure 7 includes attributes of relevance index, comment, gender and age.

Figure 8 show the taxonomy window with the root node indicated as "breakfast" and the corresponding relevance index, comments, gender and age obtained upon performing web mining using a drill down.
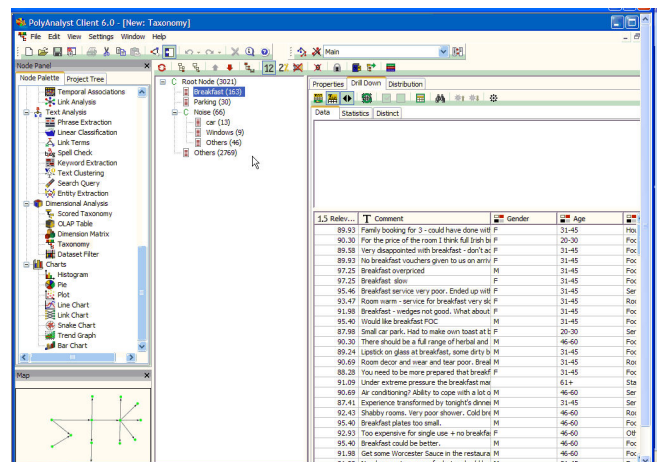

Figure 8: Taxonomy report for hotel customer written comments

Figure 9 show a window of link analysis properties with the available attributes indicated as TaxNodeType, TaxLevelName2, gender, code, purpose, hotel, region, and length; and independent attributes of TaxLevelName1 and age.

Figure 10 shows the window of PolyAnalyst with the desktop icons additional used in the web mining of the hotel customer data. These additional icons include those for dimension matrix, dimension matrix of interest, link term from kernel, taxonomy, scored taxonomy, and link analysis. The smaller window on the left bottom of this window shows the skeleton of the web mining process.

Figure 11 show the results of a search query with the selected comment of "noise." The search tree shows the roots in Figure 11 as being "noise", "noise and parking", "noise or parking", "related (noise)", and "(noise) and 'traffic' ".

Figure 12 shows the actual data set of written comments of hotel guests. The attributes of this data set that were collected include gender, age, code (e.g. room, housekeeping, staff, food, general), purpose of visit, hotel name, region (e.g., North, South, East, West) and written text of comments.
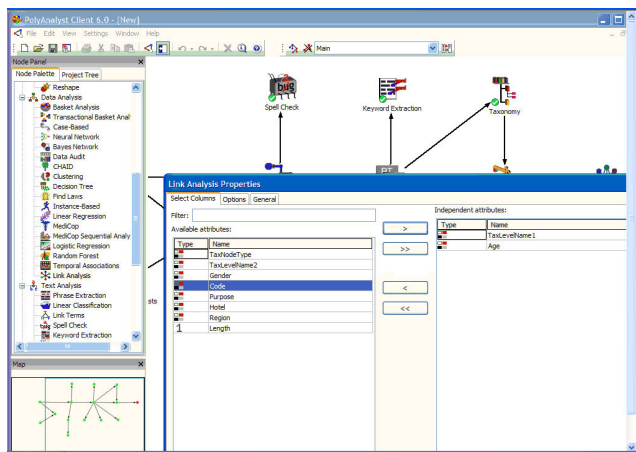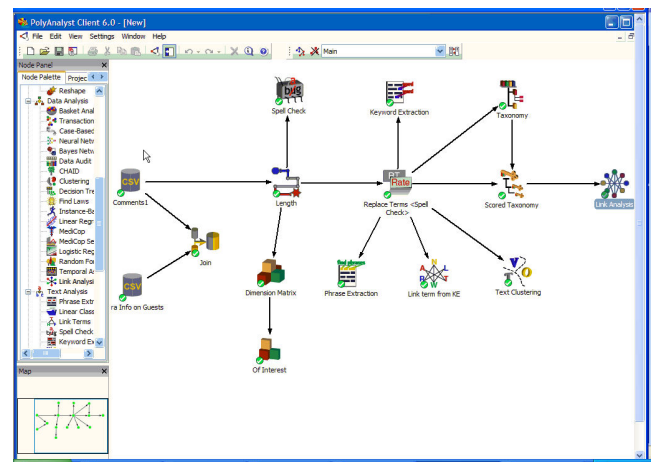

Figure 10: Complete final workspace of Megaputer PolyAnalyst® for hotel customer survey data
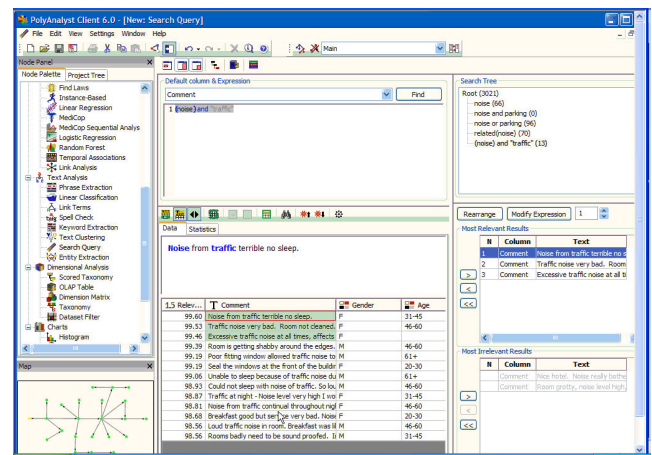
Figure 11: Query report window for search of hotel customer survey data with "noise" comment
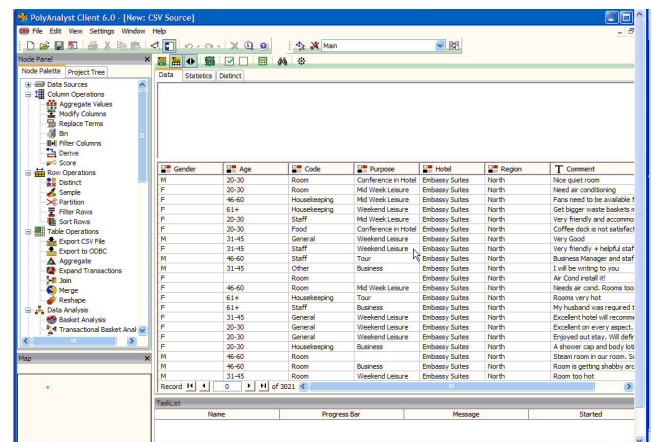

Figure 9: Link analysis properties window

Figure 12: Dataset of hotel written comments

Figure 13 is a pie chart showing frequency of origin for hotel customers per region, where the possible origins are those zones described and used in Figure 12 of North, South, East and West. As can be seen in the upper left region of this figure, the user has options to create several other pie charts (e.g. gender, age, code, purpose, hotel name, and comment). These other pie charts would likewise indicate the frequencies for each possible category of that attribute of data (e.g. gender: (male, female)).

Figure 14 is a dimension matrix including additional dimension of hotel to those of region, age, and gender. Similarly each of these dimensions has frequency counts for each of the list sub-categories. It appears from Figure 14 that the most frequent region is North with frequency 160, the most frequent age is 31-45 with frequency of 562, and the most frequent gender is female with frequency of 1605.

Figure 15 shows actual written comments of hotel customers related to traffic noise. The screen includes a root figure with branches of categories and their frequencies such as "noise and parking", a list of the most relevant and irrelevant results of the search for the combination of terms of "noise" and "traffic". Some of the most relevant results shown in Figure 15 are "Noise from traffic terrible, no sleep" and "excessive traffic noise at all times, affects sleep." The latter comment was highlighted and thus also made visible in the center window of this screen as shown in Figure 15.
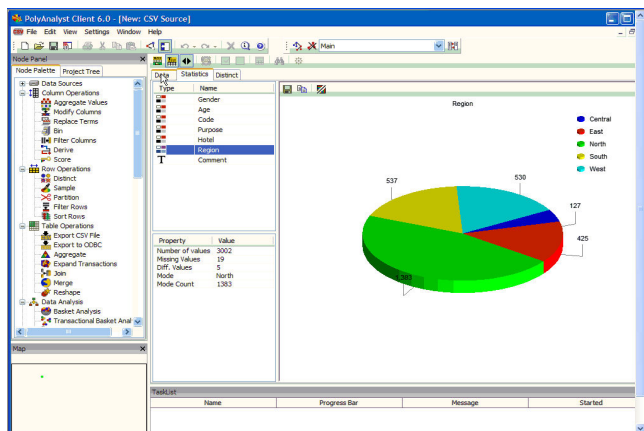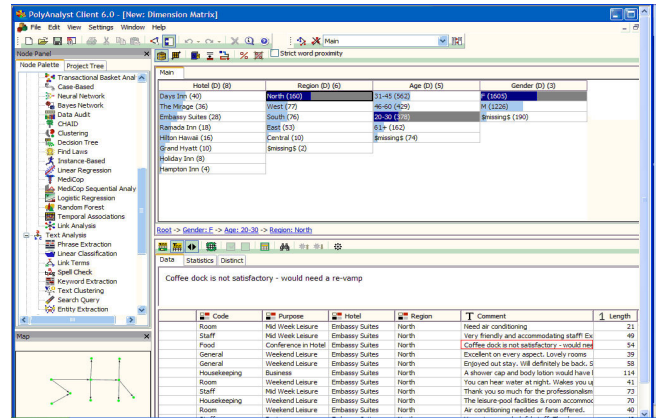

Figure 14: Dimension matrix including additional dimension of hotel
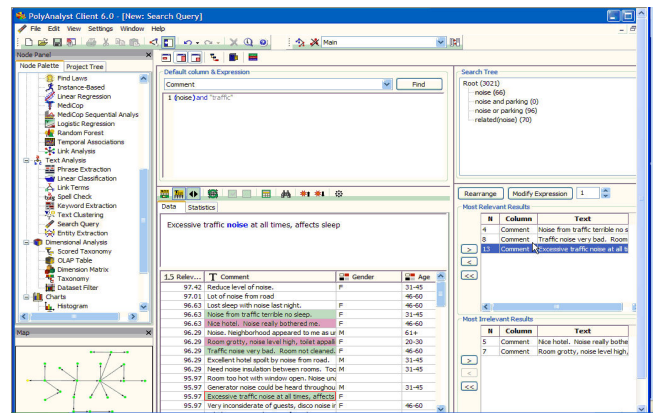
Figure 15: Written comments of hotel customers related to traffic noise

Figure 13: Pie chart showing frequency of origin for hotel customers per region

4. CONCLUSIONS AND FUTURE DIRECTIONS OF RESEARCH

This paper introduces the reader to the applications and background of web mining to hotel customer survey data. Screen shots of actual web data of hotel customer comments are provided using the data mining software of Megaputer PolyAnalyst® to provide the reader of an overview of this software and its usefulness in analyzing web mining data. This paper has shown the usefulness of web mining in extracting key words, link analysis of key words, dimension matrices, text clustering, taxonomy, and search queries all from hotel customer survey data. Web mining of the hotel customer responses makes it known to management what hotel services need to be improved.

The potential of future applications of web mining to other types of web data is made evident by the visual outputs obtained from using Megaputer PolyAnalyst®. The future directions of the research includes the

applications of web mining to other types of data sets and/or hotel customer survey data and the use of additional types of web mining software.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] Anderson Analytics (2006), Does Membership really have its privileges? Web mining in the leisure industry: Hotel brands and loyalty programs through the eye of the frequent traveler, www.marketresearchworld.net/index.php?option=content &task=view&id=961&Itemid=

[2] Ho, J. (1997), Evaluating the world wide web: A global study of commercial sites, **JCMC**, Vol. 3, No. 1, http://jcmc.indiana.edu/vol3/issue1/ho.html

[3] Huang, Y-F and Chen, Z-N (2001), Arrangement of traveling plans and web mining of a traveling system, **Proceedings of Taiwan Academic Network Conference**, http://www.ccu.edu.tw/TANET2001/TANET2001_Paper s/W103.pdf

[4] Lau, K-N., Lee, K-H, Ho, Y. (2005), Text Mining for the hotel industry, **Cornell Hotel & Restaurant Administration Quarterly**, goliath.ecnext.com /coms2/gi_0199-4589101/Text-mining-for-the-hotel.html

[5] Lau, K-N. and Lee, K-H. (2005), Text mining for the hotel industry, **Cornell Hotel and Restaurant Administration Quarterly**, Vol. 36, No. 3, pp. 344-362.

[6] Lau, K-N., Lee, K-H., Lam, P-Y, and Ho, Y. (2001), Web-site marketing for the travel-and-tourism industry, **Cornell Hotel and Restaurant Administration Quarterly,** Vol.42, No. 6, pp.55-62.

[7] Liu, J-G., and Wu, W-P.(2004), Web usage mining for electronic business applications, **Proceedings of 2004 International Conference on Machine Learning and Cybernetics**, August 26-29, Vol. 2, pp.1314-1318.

[8] Murphy, J., Hofacker, C.F., and Bennett, M., Website-generated market-research data: tracing the tracks left behind by visitors, **Cornell Hotel and Restaurant Administration Quarterly**, Vol. 42, No. 1, pp. 82-91.

[9] O'Connor, P. (2006), Who's watching you? Data collection by hotel chain websites, **Journal of Information Technology in Hospitality**, Vol. 4, No. 2-3, pp. 63-70.

[10] Olmeda, I., and Sheldon, P.J. (2002), Data mining techniques and applications for tourism Internet marketing, **Journal of Travel & Tourism Marketing**, Vol. 11, No. 2/3, pp.1-20.

[11] Raghavan, N. (2005), Data Mining in e-commerce: A survey, **Sadhana**, Vol. 30, parts 2 & 3, April/June, pp.275-289, www.iisc.ernet.in/academy/sadhana/pdf2005AprJun/Pe12 99.pdf

[12] Scharlr, A., Wober, K.W., Bauer, C. (2003), An Integrated approach to measure web site effectiveness in the European hotel industry, **Journal of Information Technology & Tourism**, Vol. 6, No. 4, pp. 257-271.

[13] Segall, R.S. (2006), Data Mining of Microarray Databases for Biotechnology, **Encyclopedia of Data Warehousing and Mining**, Edited by John Wang, Montclair State University, USA; Idea Group Inc., ISBN 1-59140-557-2.

[14] Segall, R.S. and Zhang, Q. (2009), A Survey of Selected Software Technologies for Text Mining, **Handbook of Research on Text and Web Mining Technologies**, Editors Min Song and Yi-Fang Wu, Chapter XLIV, IGI Global Inc.

[15] Segall, R.S., Guha, G-S, Nonis, S. (2008), Data mining of environmental stress tolerances for plants, **Kybernetes: The International Journal of Systems and Cybernetics**, Vol. 37, No. 1, pp. 127-148.

[16] Segall, R.S. and Zhang, Q. (2009), "Comparing Four-Selected Data Mining Software", **Encyclopedia of Data Warehousing and Mining**, Second Edition, Edited by John Wang, Montclair State University, USA, IGI Global, Inc. Chapter XLV, pp. 269-277.

[17] Segall, R.S. and Zhang, Q. (2008), "Teaching Web Mining in the Classroom: with an Overview of Web Usage Mining," **Proceedings of the Thirty-ninth Annual Conference of the Southwest Decision Sciences Institute**, Vol. 39, No. 1, March 4-6, 2008, Houston, TX.

[18] Segall R.S. and Zhang, Q. (2007) "Data Mining of

Microarray Databases for Human Lung Cancer", **Proceedings of the Thirty-eighth Annual Conference of the Southwest Decision Sciences Institute,** Vol. 38, No. 1, March 8-10, 2007, San Diego, CA.

[19] Segall, R.S. and Zhang, Q. (2006), "Applications of Neural Network and Genetic Algorithm Data Mining Techniques in Bioinformatics Knowledge Discovery – A Preliminary Study", **Proceedings of the Thirty-seventh Annual Conference of the Southwest Decision Sciences Institute**, Vol. 37, No. 1, March 2-4, 2006, Oklahoma City, OK.

[20] Segall, R. and Zhang, Q. (2006), "Micro-Array Databases for Biotechnology", **Proceedings of the 2006 Conference of Applied Research in Information Technology**, sponsored by Acxiom Laboratory for Applied Research (ALAR), University of Central Arkansas, March 3, 2006.

[21] Segall, R. and Zhang, Q. (2006), "Data Visualization and Data Mining of Continuous Numerical and Discrete Nominal-Valued Microarray Databases for Bioinformatics", **Kybernetes: The International Journal of Systems & Cybernetics**, Vol. 35, No. 10, pp. 1538-1566.

[22] Stumme, G., Hotho, A., Berendt, B. (2006), Semantic Web Mining – State of the Art and Future Directions,test.websemanticsjournal.org/openacademia//papers/20060405/document2.pdf

[23] Wober, K.W., Scharl, A., Natter, M., Taudes, A. (2002), Success Factors of European Hotel Web Sites, **Information and Communication Technologies in Tourism**, Springer, ISBN 32-2283780-9.

[24] Zhang, Q. and Segall, R.S. (2008) Using Data Mining for Forecasting Data Management Needs", **Handbook of Computational Intelligence in Manufacturing and Production Management**, Edited by Purnendu Mandel and Dipak Laha, IGI Global, Inc., 2008. [ISBN 978-159904582-5]

[25] Zhang, Q. and Segall, R.S. (2007), "Data Mining of Forest Cover and Human Lung Micro-Array Databases with Four-Selected Software", **Proceedings of the 2007 Conference of Applied Research in Information Technology**, sponsored by Acxiom Laboratory for Applied Research (ALAR), University of Arkansas-Fayetteville, March 9, 2007.