Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2011 Proceedings - All Submissions

8-5-2011

Customer Feedback Analysis using Collocations

Tomas Eklund
Abo Akademi University, tomas.eklund@abo.fi

Jarmo Toivonen
Abo Akademi University, Jarmo.Toivonen@gmail.com

Hannu Vanharanta
Tampere University of Technology, Hannu.Vanharanta@tut.fi

Barbro Back
Abo Akademi University, barbro.back@abo.fi

Follow this and additional works at: http://aisel.aisnet.org/amcis2011_submissions

Customer Feedback Analysis using Collocations

Tomas Eklund
TUCS / Department of Information Technologies,
Åbo Akademi University, Turku, Finland
Tomas.Eklund@abo.fi

Jarmo Toivonen
Department of Information Technologies, Åbo
Akademi University, Turku, Finland
Jarmo.Toivonen@gmail.com

Hannu Vanharanta
Department of Industrial Engineering and
Management, Tampere University of Technology,
Pori, Finland
Hannu.Vanharanta@tut.fi

Barbro Back
TUCS / Department of Information Technologies,
Åbo Akademi University, Turku, Finland
Barbro.Back@abo.fi

ABSTRACT

Today's ERP and CRM systems provide companies with nearly unlimited possibilities for collecting data concerning their customers. More and more of these data are more or less unstructured textual data. A good example of this type of data is customer feedback, which can potentially be used to improve customer satisfaction.

However, merely getting an overview of what lies in an unstructured mass of text is an extremely challenging task. This is the topic of the field of computational linguistics. Collocation analysis, one of the tools emerging from this field, is a tool developed for this task in particular. In this paper, we use the collocation analysis to study a text corpora consisting of 64,806 pieces of customer feedback collected through a case company's online customer portal. Collocation analysis is shown to be a very useful tool for exploratory analysis of highly unstructured customer feedback.

Keywords

Opinion mining; CRM; collocation analysis; collocation networks; customer feedback; computational linguistics

INTRODUCTION

Modern Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) systems are immensely powerful tools for collecting information about customers. There is a great need for tools to help companies learn more about their customers, and by extension, to improve customer satisfaction and thereby customer retention (Buttle, 2004). Analytical tools in the form of Business Intelligence tools have improved greatly during the last years and in particular the tools for analysis of quantitative data.

One of the most significant challenges within CRM, however, concerns the analysis of textual information, such customer feedback, online community discussions, and other highly unstructured masses of text. These masses represent a wealth of information that is generally underutilized in industry and commerce today. Harnessing this information is the topic of the field of opinion mining (Pang and Lee, 2008), an area often utilizing text mining tools. However, many of the constantly developing methods available are quite complex, and choosing a method understandable by business management is not a trivial task.

The purpose of this paper is to apply collocation analysis to the analysis of customer feedback through a Finnish online company portal. Collocation analysis is a computational linguistics method for visualizing and analyzing the structure and relationships of words in a large text mass. The case company has been collecting customer feedback through the system, but until recently, no systematic overview of the collected data has been performed. The data contains many categories of feedback, not necessarily clear from the classification performed by the submitter of the feedback, unlike the case with online reviews. This favors an exploratory approach to the analysis. The paper shows that collocation analysis represents an easy to understand and potentially highly visual approach for analyzing customer feedback, and that the results are also easily presented and explained to a business-oriented audience.

The rest of the paper is organized as follows. In the next section, collocation analysis and collocation networks are presented. In Section 3, the case company, company portal, and data are discussed, and preprocessing carried out on the data is

presented. Some examples of the results are shown in Section 4, and limitations and future research are discussed in the conclusions.

METHODOLOGY

Opinion mining has been the topic of great interest in the research community for some time now. With the great advances made in the field of text mining in recent years, semi-automated analysis of customer feedback data is becoming a feasible approach within the field of CRM. Opinion mining (often also referred to as sentiment analysis) is essentially the process of using computational methods to extract opinions, sentiments, or subjectivity from textual data (Pang and Lee, 2008). Opinion mining can be seen as a subset of text mining, which itself is an umbrella term for employing methods from e.g., statistics, machine learning, and computational linguistics for the analysis of textual data.

Pang and Lee (2008) divide opinion mining applications into two main groups, *classification* and *summarization*. To these, Lee et al. (Lee, Jeong and Lee, 2008; Pang and Lee, 2008) add *development of linguistic resources*. In classification, the objective is classify or categorize text based upon the opinions or sentiments expressed, for example, positive or negative product or service reviews. Often, supervised machine learning tools are used to identify the sentiment or opinions based upon a number of pre-classified examples. Classification approaches are often word vector based, and involve heavy preprocessing such as feature extraction, dimensional reduction (e.g., latent semantic indexing), etc. Examples of often used methods are Bayesian classifiers (Nauck, Ruta, Spott and Azvine, 2006) and Support Vector Machines (Lakshminarayan, Yu and Benson, 2005; Nauck, Ruta, Spott and Azvine, 2006).

In opinion or sentiment summarization, the task is to gain an overview of the opinions or sentiments expressed in a single document or collection of documents, for example, an online review site. Opinion classification might even form a sub task of summarization (Lee, Jeong and Lee, 2008). Many of the applications of opinion summarization use methods from the greater field of text summarization, including methods from computational linguistics. In particular, in recent years, much attention has been focused on various approaches for visual summarization of opinion feedback (e.g., Gamon, Aue, Corston-Oliver and Ringger, 2005; Oelke, Hao, Rohrdantz, Keim, Dayal, Huag and Halldór, 2009; Viegas, Wattenberg and Feinberg, 2009; Wu, Wei, Liu, Au, Cui, Zhou and Qu, 2010).

Opinion mining is a challenging area, as customer feedback occurs in many widely varying forms, such as market research surveys, focus groups, personal interviews, observations, and free form feedback. – Of these, free form web surveys, such as multi-topic feedback fields, represent the richest data, but are extremely challenging to analyze (Lakshminarayan, Yu and Benson, 2005). They are often the most common data available, as most companies with an online presence collect some form of feedback on their web pages. As such, an exploratory data analysis approach is a natural way to approach these data. Most opinion mining applications, however, deal with pre-classified texts (e.g., positive, negative, or neutral feedback concerning movies, books, or other products), and are ill-suited for exploratory analysis of free-form customer feedback. Often, free-form customer feedback even concerns multiple topics, and therefore, can be difficult to categorize per topic or negative or positive. In addition, many tools within opinion mining are still in their infancy and the results of these are often quite difficult to interpret and validate. Machine learning-based methods also typically require heavy preprocessing and modification of the data. There is a risk of erasing the structure of a text when using machine learning-based, categorization approaches, which is a clear danger in exploratory analyses.

However, there are a number of methods within computational linguistics that provide quick overviews of the contents of text corpora. One such tool is collocation analysis, and it's visual extension, collocation networks. Collocation analysis is a method that has emerged from the field of linguistics, specifically the subfield of corpus linguistics, which concerns the analysis of text corpora (collections) (Vechtomova, Robertson and Jones, 2003). Collocation analysis is based upon the idea that words often occurring within a meaningful distance of each other in a text imply some sort of semantic association. A collocation can be defined as "the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991). The field has emerged due to the difficulties in assessing the meaning of single words in a context; instead focus is on the circumstances under which they occur. The significance of a collocation is measured by how often words occur together compared to what would be expected by coincidence. A commonly used measure in computational linguistics is the *mutual information (MI) score* (Church and Hanks, 1990). A further development of collocation analysis is in the actual visualization of the central words and the links between them. This can be done using collocation networks (Williams, 1998).

Collocation analysis has been widely applied in the scientific literature. For example, Böhm et al (2002) used collocation analysis to provide basic information on large text corpora, to serve as a basis for further text mining. Magnusson et al. (2005) used collocation analysis to study changes in the textual parts of annual reports and found indications of financial performance changes preceding the actual realization of the changes in the actual financial figures. Ku et al. (2009) used collocation analysis as one tool in a palette for analyzing online product opinions. Magnusson (2010) used collocation

networks for temporal visualization and analysis corporate communication. Collocation networks have also been applied as the basis for concept grouping, e.g., clustering of massive amounts of documents based upon similarity (Li, Sun and Wang, 2009; Veling and van der Weerd, 1999).

However, collocation analysis has not been as widely applied in industry and commerce, and to our knowledge, has not previously been applied as an exploratory approach to the analysis of multi-category customer feedback. As an easily explainable method suitable for exploratory text analysis, collocation analysis could potentially be a suitable tool for customer feedback analysis.

THE BUSINESS CASE

Case company

The case company is a national retail and service chain. The chain has an online portal through which customers can provide feedback about products, services, or particular stores, and also get answers to any questions they may have or follow up on their loyalty program status. Each feedback or question received through the online portal is manually classified and forwarded to the relevant part of the chain. The case company has not yet performed any kind of systematic overview of the contents of the feedback database since it was created. Also, the categorization of feedback and enquiries was originally specified when the online portal was first developed. Thus, a systematic overview would also help the company to better classify the types of feedback and enquiries arriving through the online portal.

Data

The data consisted of all customer feedback provided through the company's online portal during the period 12.12.2003 - 11.2.2009, in total 64,806 pieces of feedback. The data were provided as a tab-delimited text file, in the following format:

> Date and time | title | subject | detailed subject | feedback | checkbox

The checkbox indicates if the customer would like a representative of the company to contact them concerning the feedback. An example of one piece of feedback is the following[1]:

> 12.1.2004 12:01:12|Empty shelves|Feedback about our stores|Store name||Unfortunately, I have to provide negative feedback concerning your store in [city] […]|checkbox

The majority of the feedback was in Finnish, but a number were also in Swedish or English. For the purposes of this study, feedback and questions in other languages than Finnish were ignored.

Preprocessing

As is typical in computer-aided natural language processing, the raw data require a significant amount of preprocessing because of error such as typos and character errors, word ambiguity, specialty terms or slang, stop words, etc. (Brier and Hopp, 2011; Gamon, 2004; Lakshminarayan, Yu and Benson, 2005; Manning and Schütze, 1999). In this case, we created a filter to perform the following steps:

1. Split each feedback or enquiry so that each sentence is on a separate line

2. Remove noise, such as keystroke errors and errors relating to Scandinavian letters (Å, Ä, and Ö)

3. Remove sensitive data, such as personal IDs and bank account numbers, and replace with XXXX

4. Round all numbers to magnitude

5. Replace synonyms with one specified word

6. Identify and remove stop-words, i.e., words that are frequent and have little meaning (e.g., *the, a, that* in English). In practice, a list of Finnish stop words was used and new words added where necessary until meaningless words were removed (Brier and Hopp, 2011).

During the process of preprocessing, a number of problems with the data were noted. In addition to keystroke and other input errors, there were a number of spam messages, as well as test messages relating to the implementation and testing of the

---

[1] Freely translated from Finnish

company portal, in the dataset. In addition, the previously noted messages in different languages needed to be dealt with. There were a number of incomplete messages, either aborted by the user or interrupted due to technical errors. In some cases, these resulted in (partially) duplicate messages when the user reentered the same message. Finally, there were also some problems with words being split up in the middle. This was probably due to technical limitations in the feedback input form.

RESULTS AND ANALYSIS

The case company provided a number of keywords for the analysis, the purpose being to look for collocations to this list. The list was divided into two sub lists; one to assess how often and in which circumstances certain words were used (e.g., disappointed, suggestion), and the other to study positive and negative collocations to certain words (e.g., store name, prices). For each word, a number of synonyms were provided. The list of keywords can be found below[2].

Keywords for association analysis:

- Disappointed
- Enquire
- Suggestion
- Help
- Unanswered enquiry
- Survey

Keywords for positive / negative association analysis:

- Online service
- Web page
- Wlan
- Online customer portal
- Loyalty program
- Online application for loyalty program
- Store name + prices
- Store name + selection
- Development + online

Firstly, a heuristically determined search was performed for collocations including each word on the list. Based upon the results of searches, Boolean operators such as * were used to capture as large a part of the sought data as possible. A window of +/- five words was determined to be suitable in this case. Then, frequencies were calculated for each collocation, as well as counts of position within the +/- five word window. The list was ordered according to the frequency of the collocations. We also performed lemmatization on the words in the first list, i.e., transforming the words to the base form (e.g., walking -> walk).

For example, for the keyword DISAPPOINTED, most of the closest matches were indicative of the magnitude of the disappointment, such as VERY (n=543), BIG (n=251), and EXTREMELY (n=198), in most cases appearing on the left hand of the collocation as would be expected. These were also followed by the subject of the disappointment, such as the LOYALTY CARD SYSTEM (n=50), STORE NAME, SERVICE (n=82), PRODUCT (n=85), etc. The words AGAIN (n=29) and OFTEN (n=27) signified repeated frustration. The appearance of certain specific store names more often than others, combined with specific complaints, could indicate potential problems within the chain that need to be solved. However, the number of negative collocations such as these was fairly small considering the size of the text corpora.

---

[2] The list is a free translation from Finnish and thus is not entirely comparable to the Finnish original

Another example is the collocations for the keywords STORE NAME and PRICES. The most interesting collocations are TODAY (n=19), YESTERDAY (n=19), SAME (n=19), RISE (n=9), GOOD (n=9), and WONDER (as in "I wonder if/what…") (n=8). Specific store names or locations, as well and competitors, are also mentioned, indicating that some kind of comparison is being made. However, the number of collocations including these keywords appearing in the text is small, indicating that prices are somewhat surprisingly not mentioned very often in the feedback provided by customers.

In addition, a collocation network (Williams, 1998) was created based upon the base form (lemma) of the name of a particular chain of stores in the case company. We used the *t-score* (Barnbrook, 1996) measure to determine the significance of each collocation, according to the equation Eq. 1.

$$t = \frac{J - \dfrac{F_1 * F_2}{F_t}}{\sqrt{J}},$$ (Eq. 1)

where F_1 is the frequency of the store name in the text, F_2 is the frequency of its collocation word, F_t is the frequency of all collocations in the text, and J is the frequency of the store name collocation. This resulted in 11 initial collocations, which were then further branched into a network using the same formula. Thus, the second branch was allowed to freely develop based upon appearing collocations. The network was cut off at a level determined to be suitable for out analysis, based upon the t-score. In this case, region names or individual store indicators were left out of the network in order to simplify the analysis, but this type of application could in itself be interesting to pursue in the future.

The resulting network can be found in Figure 1. It is important to note that the t-score limits results to significant collocations. Thus, commonly occurring words and their collocations receive little gain, and thus, the list of collocations obtained previously may look very different from the network built using the t-score.
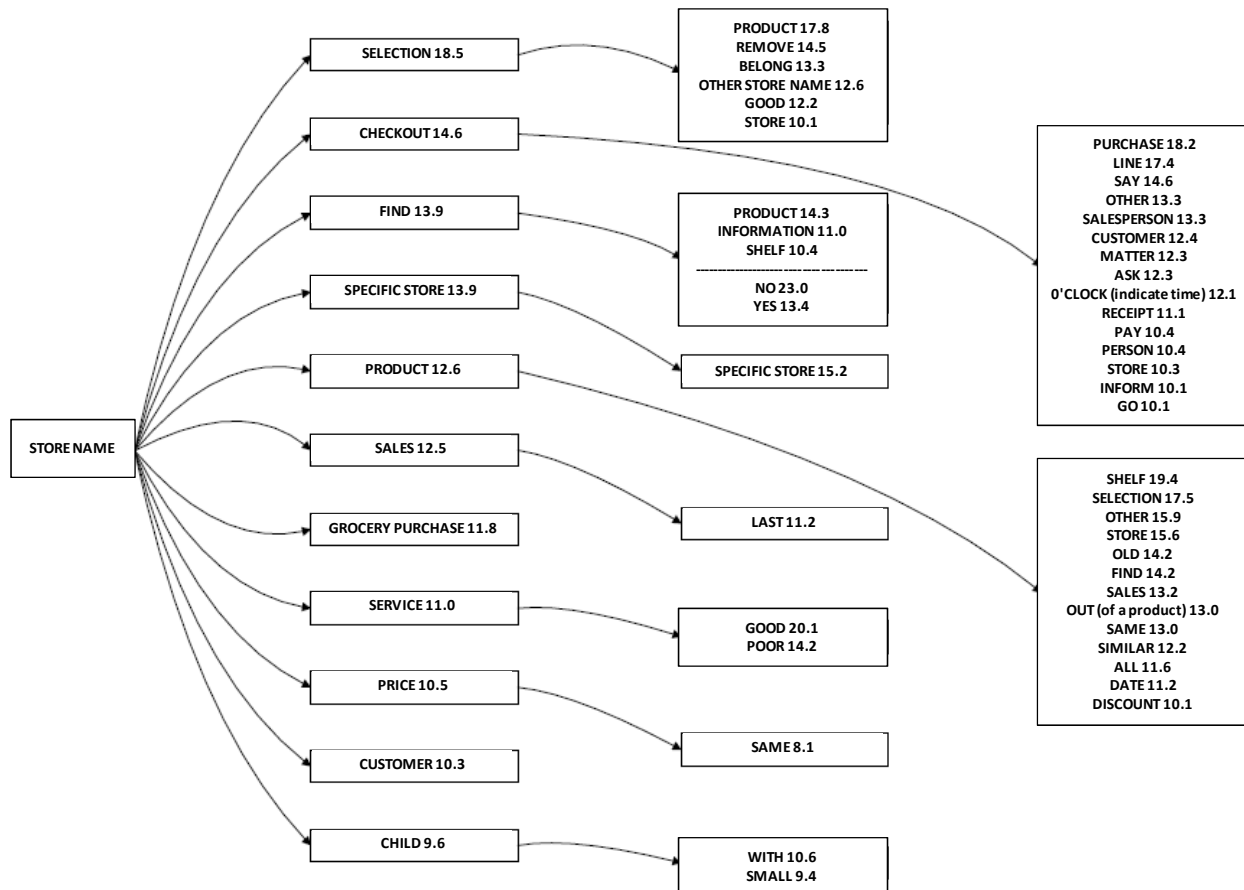

Figure 1. The final collocation network based upon the store name.

There are a number of interesting links in the formed networks. Most links seem to form around the words SELECTION, CHECKOUT, and PRODUCT, linked to STORE NAME. For example, there is an obvious link between the collocation STORE NAME – SELECTION – GOOD, indicating positive feedback.

Examples of neutral or undetermined collocations include STORE NAME – SELECTION – REMOVE and BELONG, indicating feedback about the selection but not if it is negative or positive. Likewise, the collocation STORE NAME – PRODUCT – SELECTION contains no indication of negative or positive feedback.

On the other hand, there are also collocations that can be interpreted as negative, such as STORE NAME – CHECKOUT – LINE, and indications of times (O'CLOCK), with very strong t-scores. This would indicate that there has been feedback complaining about long lines at specific times. Other negative collocations include STORE NAME – PRODUCT – OLD, DATE, and OUT, indicating negative feedback about products or stocking levels, and STORE NAME – FIND – PRODUCT, which likely indicates that the customers had difficulties locating a particular product. PRICE contains no specific collocations indicative of positive or negative feedback.

Finally, one of the most significant collocations from the perspective of the case company is STORE NAME – SERVICE – GOOD (t-score 20.1) or POOR (t-score 14.2), indicating a majority of positive feedback concerning service.

CONCLUSION

Opinion mining is fundamental to the process of improving the communication between the company and its customers today, yet at the same time, much of the information that it is available for this task is underutilized because of the difficulty involved in analyzing it. In this paper, we have presented the application of collocation analysis and collocation networks to the analysis of customer feedback and questions obtained through a case company's online customer portal. Collocation analysis is a simple and quick method for identifying significant associations in a large text corpus. Results are intuitive to assess when dealing with a specific domain of language, such as customer feedback and questions, and findings that differ from what intuitively would be expected are potentially interesting for further analysis. The results show that it is possible to exploratively analyze a difficult and varying dataset such as this with collocation analysis.

Collocation analysis has several advantages over more complex text mining methods. One of the most significant advantages of collocation analysis is that it can be used with very little *a priori* information about the contents of a text, and works solely based upon word and collocation frequencies to generate overviews. In this case, collocation analysis has indeed been used exploratively, to gain a general overview of the dataset. Secondly, collocation analysis retains many of the relationships in a text, allowing the user to analyze the structure of the text. This is an advantage over methods based upon word frequencies, as is typical, for example, within opinion mining using machine learning methods. Thirdly, preprocessing, although still demanding, is often based upon standard filtering approaches, such as stop word lists and stemming methods, which are readily available as commercial and non-commercial tools on the Internet. For example, latent semantic indexing or challenging quantification of the data is not required, as opposed to many machine learning approaches, which is an important advantage when considering a business-oriented audience. Additionally, the text analyzed in this case requires an exploratory approach, as it is even less structured than the online reviews typically addressed in opinion mining (e.g., online car reviews: Gamon, Aue, Corston-Oliver and Ringger, 2005; printer reviews: Oelke, Hao, Rohrdantz, Keim, Dayal, Huag and Halldór, 2009). In this case, the feedback can relate to anything from the online portal itself, to an individual store or product.

Limitations

There are of course limitations of this research that should be addressed. Most importantly, preprocessing of the text corpus is still a delicate problem and can significantly affect the results of the study. Firstly, synonyms and word form problems need to be dealt with. Keystroke errors, language errors, online slang, incomplete or duplicate messages, and technical problems leading to incorrect input are common problems, as well as spam and test messages. Messages in different languages also affect the results. In all, as with most computational linguistics approaches, significant filtering is required.

Another limitation of this approach is the time required for "tweaking" the analysis. Much of the preprocessing is heuristically based upon the iteratively obtained results. For example, setting the cutoff points (window of words and t-score) is difficult as it is a balance between rigor and relevance; the more detailed the analysis is, the more there is a risk of losing the overall picture. It is also worth noting that the analysis is not entirely exploratory, as a list of keywords was provided by the case company. However, as was mentioned earlier, there are a number of tools available for performing much of filtering required.

FUTURE RESEARCH POSSIBILITIES

The results of this study open a number of development options in the future. Firstly, collocation analysis could be used to provide better a classification of feedback and questions, leading to new subject categories, better handling of messages, and reduced personnel workload. This could in turn also be used to develop the online portal itself, improving service and customer satisfaction.

Another potential advantage is the automatic cleaning of the input data performed, making the data more suitable for long term storage and analysis.

Further, time series analysis on the data could be performed, to see evolving or resolved issues, seasonal variance, and trends. This would require visualization tools for efficient dissemination to managers and decision makers. The results could also be combined with quantitative data, such as sales figures and key performance ratios. External information, such as online forum discussions and web pages could also be included by using web crawling techniques, but this presents new challenges concerning preprocessing.

Finally, emphasis should be placed on weak signals appearing in the text, i.e., words appearing with a low frequency. Currently, such words are largely ignored by this method, but might contain useful information for analysis.

ACKNOWLEDGMENTS

REFERENCES

1. Barnbrook, G. (1996) *Language and Computers: A Practical Introduction to the Computer Analysis of Language,* Edinburgh University Press, Edinburgh.

2. Böhm, K., G. Heyer, U. Quasthoff, and C. Wolff. (2002) Topic Map Generation using Text Mining, *Journal of Universal Computer Science,* 8, 6, 623-633.

3. Brier, A., and B. Hopp. (2011) Computer Assisted Text Analysis in the Social Sciences, *Quality & Quantity,* 45, 1, 103-128.

4. Buttle, F. (2004) *Customer Relationship Management Concepts and Tools,* Butterworth-Heinemann, Oxford.

5. Church, K. W., and P. Hanks. (1990) Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics,* 16, 1, 22-29.

6. Gamon, M. (2004) *Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis,* in Anonymous *Proceeding of COLING-04, the 20th International Conference on Computational Linguistics,* Aug 23-27, Geneva, CH, 841-847.

7. Gamon, M., A. Aue, S. Corston-Oliver, and E. Ringger. (2005) *Pulse: Mining Customer Opinions from Free Text,* in A. Fazel Famili, Joost N. Kok, José M. Peña, Arno Siebes and Ad Feelders (eds.), *Advances in Intelligent Data Analysis VI,* Springer Berlin / Heidelberg, 121-132.

8. Ku, L., H. Ho, and H. Chen. (2009) Opinion Mining and Relationship Discovery using CopeOpi Opinion Analysis System, *Journal of the American Society for Information Science and Technology,* 60, 7, 1486-1503.

9. Lakshminarayan, C., Q. Yu, and A. Benson. (2005) *Improving Customer Experience Via Text Mining,* in S. Bhalla (ed.), *Databases in Networked Information Systems,* Springer Verlag, Heidelberg, 288-299.

10. Lee, D., O. Jeong, and S. Lee. (2008) *Opinion Mining of Customer Feedback Data on the Web,* in Anonymous *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication (ICUIMC '08),* Jan 15-16, Suwon, Korea, ACM, 230-235.

11. Li, H., C. Sun, and K. Wang. (2009) *Clustering Web Search Results using Conceptual Grouping,* in Anonymous *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics (ICMLC 2009),* 12-15 July, Baoding, Hebei, China, 1499-1503.

12. Magnusson, C. (2010) Improving Competitive Analysis with Temporal Text Visualization, *Marketing Intelligence & Planning,* 28, 5, 571-581.

13. Magnusson, C., A. Arppe, T. Eklund, B. Back, H. Vanharanta, and A. Visa. (2005) The Language of Quarterly Reports as an Indicator of Change in the Company's Financial Status, *Information and Management,* 42, 4, 561-574.

14. Manning, C., and H. Schütze. (1999) *Foundations of Statistical Natural Language Processing,* MIT Press, Cambridge, Massachusetts.

15. Nauck, D. D., D. Ruta, M. Spott, and B. Azvine. (2006) *A Tool for Intelligent Customer Analytics,* in Anonymous *Proceedings of the 3rd International IEEE Conference on Intelligent Systems (IS ′06),* Sept 4-6, London, UK, IEEE, 518-521.

16. Oelke, D., M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L. Huag, and J. Halldór. (2009) *Visual Opinion Analysis of Customer Feedback Data,* in Anonymous *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009),* Oct 12-13, Atlantic City, NJ, USA, IEEE, 187-194.

17. Pang, B., and L. Lee. (2008) Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval,* 2, 1-2, 1-135.

18. Sinclair, J. (1991) *Corpus, Concordance, Collocation,* Oxford University Press, Oxford.

19. Vechtomova, O., S. Robertson, and S. Jones. (2003) Query Expansion with Long-Span Collocates, *Information Retrieval,* 6, 2, 251-273.

20. Veling, A., and P. van der Weerd. (1999) *Conceptual Grouping in Word Co-Occurrence Networks,* in Anonymous *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI '99),* July 31 - August 6, Stockholm, Sweden, Morgan Kaufmann Publishers, Inc., 694-699.

21. Viegas, F. B., M. Wattenberg, and J. Feinberg. (2009) Participatory Visualization with Wordle, *IEEE Transactions on Visualization and Computer Graphics,* 15, 6, 1137-1144.

22. Williams, G. C. (1998) Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles, *International Journal of Corpus Linguistics,* 3, 1, 151-171.

23. Wu, Y., F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. (2010) OpinionSeer: Interactive Visualization of Hotel Customer Feedback, *IEEE Transactions on Visualization and Computer Graphics,* 16, 6, 1109-1118.