Text Clustering and Summary Techniques for CRM Message Management

Dmitri Roussinov

School of Accountancy and Information Management

SAIM, College of Business, Arizona State University,

Box 873606, Tempe, AZ 85287-3606

dmitri.roussinov@asu.edu

J. Leon Zhao

Department of MIS

Eller College of Business and Public Administration

University of Arizona, Tucson, AZ 85721

lzhao@bpa.arizona.edu

Abstract

One of Customer Relationship Management (CRM) activities involves soliciting customer feedback on product and service quality and the resolution of customer complaints. Inevitably, companies must deal with large number of CRM messages from their customers either through emails or from work logs. Going through those messages is an important but tedious task for managers or CRM specialists in order to make strategic plans on where to place the resources to achieve better CRM results. In this paper, we present a methodology for making sense out of CRM messages based on text clustering and summary techniques. The unique features of CRM messages are the short message length and frequent availability of correlated CRM ratings. We propose several novel techniques including organizational concept space, web mining of similarity relationships between concepts, and correlated analysis of text and ratings. We have tested the basic concepts and techniques of CRM Sense Maker in a business setting where customer surveys are used to set strategic directions in customer services.

Keywords

Customer Relationship Management, knowledge management, web mining, text clustering.

Introduction

According to a recent book on Customer Relationship Management (Swift, 2001), CRM is "an enterprise approach to understanding and influencing customer behavior thorough meaningful

communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability." The challenge is to continue attracting new and profitable customers whilst forming ever-tighter bonds with existing ones, thus creating a profitable customer base. One of the important aspects of customer retention is to address their problems and queries timely and effectively.

Companies are spending a large amount on service management (SM) and complaint management (CM) systems in order to understand and evaluate which customer issues to act on, and in what order. According to CRMDaily.com, worldwide CRM spending reached US$13.7 billion by the end of 2002 based on a report by the Aberdeen Group and should be close to $20 billion by 2005 (Morphy, 2002). As such, CRM is a strategic matter, and the company must decide how to allocate limited CRM resources for the maximum impact on overall customer satisfaction.

Since customer support inquiries often provide critical information on product defects or service deficiencies companies need automated methods for tracking customer complaints and actions taken to close the gap with the customer. Modern customer relationship management systems provide an integrated, structured connection between customer support centers and back-end engineering, production, or product management organizations (www.remedy.com). It helps companies to capture and store information on customer issues, ideas, and suggestions, as well as to assign and route them to the appropriate groups.

CRM managers must frequently study customer survey data, customer complaint data, and other forms of customer sentiment messages in the form of text. Because the number of those CRM messages is frequently in the hundreds or thousands, reading and summarizing those messages is an overwhelming task. Consequently, there is a great need for an effective tool to assist CRM managers to analyze thousands of customer messages in order to discover recurrent issues and problems based on customers' feedback. This problem falls into a more general problem of information overload in the knowledge management context and even a more general context of computer mediated communication (CMC) (Hiltz & Turoff, 1985; Gallupe & Cooper, 1993).

Within CRM systems, the messages are frequently stored in a digital format. This makes it possible to utilize past research ideas and tools in computer mediated communication (CMC) towards managing

CRM messages. The general CMC approach to Information Overload reduction is to impose structure on the data (Hiltz & Turoff, 1985). Much work in this direction has been done within the Group Decision Support Systems context, specifically in automated summarization of meeting messages, for example by representing them with a list of most representative topics (Chen, et al., 1994), or using concept maps (Orwig & Chen & Nunamaker, 1997), or clustering messages into semantically homogeneous groups (Roussinov & Chen, 1999). The common belief behind those approaches is that automated processing techniques can reduce the cognitive load of meeting participants even if manual post-processing is still required.

While many text-processing techniques exist in the literature and laboratories, few CRM tools have incorporated them in the real world. The main reason is that the existing techniques are not easy to use by an average manager.

In this paper, we address the ability of an organization to understand the customer messages and thereby improve the communication efficiency and effectiveness. We present a conceptually novel toolset called CRM Sense Maker that has been developed to support CRM managers at a customer support center in a large university. Our tool can summarize customer feedback messages using state-of-the-art text processing techniques such as automatic indexing and clustering.

We claim contributions along several lines. First, we applied a previously studied document clustering approach to CRM domain. Second, we significantly improved the usability of the techniques due to following:

1) Semi-Automatic approach: Allowing user interference at the crucial stages to improve the quality of the outcome.

2) Applying the notion of organizational concept space (OCS) (also referred as the similarity network or thesaurus approach) to address a very well known vocabulary diversity problem, i.e., different words used to refer to same or similar meanings (Furnas et al., 1987).

3) Integrating a web mining technique we have developed recently (Roussinov & Zhao, 2003) to obtain OCS automatically from Internet with high reliability. A typical collection of CRM messages, although formidable to analyze manually, does not contain sufficient number of words

and phrases to derive a reliable organizational concept space (OCS) needed for analyzing the messages. That is why we believe Internet mining is a crucial technique in this context.

The rest of the paper is organized as follows. Section 2 introduces the concepts and techniques of the CRM Sense Maker. Section 3 discusses a business case in which the CRM Sense Maker is applied and tested. Section 4 concludes the paper and points out directions of future studies.

CRM Sense Maker

The CRM Sense Maker consists of the following three steps: (1) identifying descriptive terms, (2) identifying semantic relationships between them, and (3) grouping messages into clusters of related issues. The following subsections explain in more detail what each step does and why it is necessary.

Identifying Descriptors

The content of each text message is described by words and phrases that this message contains, which is currently the most effective and efficient representation known for information scientists and artificial intelligence researchers. Those content bearing words (called *terms*) are identified through a process called *automatic indexing*. The general purpose of automatic indexing is to identify the contents of each textual document automatically in terms of associated features, i.e., words or phrases. Automatic indexing first extracts all words and possible phrases in the document. Then it removes words from a "stop-word" list to eliminate non-semantic bearing words such as "the", "a", "on", and "in".

As in the state of the art in text-processing technologies, after automatic indexing, each message (document) is represented by a vector. Each coordinate in the vector space corresponds to a term. If a term is present in the document, the coordinate is set to *1*, otherwise to *0*. For computational efficiency and accuracy of representation, only the specified number of most frequent terms is used for vector representation. According to Chen, et al. (1994), Orwig & Chen & Nunamaker (1997), Roussinov and Chen (1999), this approach works best with small collections consisting of short text messages, since it provides the greatest overlap in representations.

The accuracy of this vector representation is crucial for every text technology involved, specifically automatic clustering, categorization, retrieval or summarization. Apart from its statistical properties in

the collection of documents (messages), each term is treated the same way, regardless of its semantic meaning, which apparently results in problems. Some terms do not help to represent messages since they may have too general meaning for the context at hand. Hence, we suggest that manual cleaning of context bearing terms selected for vector space representation is necessary for the technologies to be applicable in real-life (e.g. managerial) applications.

take in Figure 1

Thus, the first step in the process is an interactive review of the automatically suggested terms. Currently, it is implemented using MS Access database software as shown on Figure 1. The user has three options: 1) discard a term as non descriptive ("not useful") (e.g. TECHNCIAL SUPPORT is too general and not useful in this context since all messages are related to technical support anyway), 2) identify a term as a definitely descriptive ("Useful"), e.g. TELEPHONE, 3) do not provide any feedback on a term (default option). Once, the user is finished, the system gives higher weights to the descriptive terms in the vector space representation of the messages which promotes their influence on the clustering outcome.

Grouping Descriptors into Concepts

The vector space model has another serious limitation since it does not take similarities between different words and phrases into account. For example, `customer` and `user` would be treated as different words, although in our CRM context they are nearly synonyms.

This problem has also been noticed in a more general domain of text technologies and traditionally known as vocabulary problem (Furnas et al., 1987). However, there has not been an effective solution to it. Since natural languages are very ambiguous and diverse, solving this problem would require knowing semantic relationships between all possible words and phrases. This task is believed to be "AI-complete," (Ide & Véronis, 1998) which means solving it would require solving all the other AI (Artificial Intelligence) tasks such as natural language understanding, common sense reasoning and logical thinking.

Nevertheless, we believe that some progress in the right direction can be made. While solving the problem in the most general setting does not seem to be feasible in the nearest future, alleviating it

within a particular organization or a particular task, such as CRM, by applying Organizational Concept Space (OCS) (Zhao, Kumar & Stohr, 2000) has been shown to be possible. OCS is an organization specific framework, that among the other data structures, includes a so-called *similarity network*, a collection of similarity relationships between the important concepts. Figure 2 illustrates a simple similarity network with generalization (up) – specialization (down) hierarchy. All concepts in the same node of the network or connected by arcs are believed to be strongly semantically related.

take in Figure 2

Roussinov & Zhao (2003) presented and empirically validated Web mining approach that is capable of discovering semantic relationships between specified concepts, and as a result, helps to organize messages produced during electronic meetings supported by Group Decision Support Systems. In their study OCS was successfully "text mined" from the World Wide Web.

take in Figure 3

In our current project, we combine automated mining with the manual user feedback to build and maintain real size organizational concept spaces for the purpose of Customer Relationship Management. Figure 3 shows an example of manual refinement of OCS implemented as editing a specially formatted text file using Notepad editor from MS Windows. Each concept is placed on a new line, and related concepts immediately follow and are indicated by indentation (e.g. TRAINING is related to CLASSES etc.). The initial relationships are built automatically through the co-occurrence based text mining (Roussinov & Zhao, 2003). Then, a CRM manager can refine them.

Clustering Messages into Issues

Recently, information visualization techniques have revived interest in text clustering. The idea behind many of these techniques that are able to visualize large collections of documents is to agglomerate similar documents into clusters and present a high-level summary (e.g. via a list of the most representative terms) of each cluster. This way, the user does not need to go through similar documents or through entire documents in order to become familiar with the collection. This greatly reduces redundancy and cognitive demand. Examples of such visualization systems are Scatter/Gather (Cutting et al., 1992), WebBook (Card & Robertson & York, 1996), and SenseMaker (Wang

Baldonado & Winograd, 1997). Hearst (1997) gives a comprehensive overview of such systems and the ideas behind them.

Our final step organizes messages (clusters) into groups of similar issues through a semi-automatic interactive procedure. Figure 4 shows an example of a file containing CRM messages organized into clusters. Each cluster is described by automatically identified most representative terms (e.g. CLASSES or EXCELLENT JOB) and started with a marker "*** New Issue". The messages are separated by empty lines. First, the initial grouping and assigning labels to clusters is done automatically. Then, the user (a CRM manager) can manually clean up the groups or just glance over them to identify re-occurring issues.

take in Figure 4

Application of the CRM Sense Maker

Currently, we are testing our tool with 1438 CRM messages collected in the period of several years by a computer customer support center in a large university (CCIT). We are in the process of a field study with CCIT CRM managers. Since our prototype system includes several components mentioned above, each of them is being tested and validated empirically. One of our research goals is to develop new techniques for analyzing customer survey messages and validating the prototype system in the CCIT environment.

CCIT provides comprehensive services to hundreds of units and tens of thousands of users throughout the university community. Its services include e-mail & computer accounts, resources for teaching, campus telephone service, resources for research, resources for administrative systems, and campus networking.

CCIT itself is not an independent organization by itself but rather a sub-organization within a large University which primary goal is to provide college level and graduate education. Due to competitive nature of modern education market, Universities strive to achieve the reputation for high quality and degree of customer satisfaction, in which computer services play important role. That is why one of the important missions of CCIT is to perpetually provide and improve the quality of its services. In

order to evaluate the quality and elucidate potential drawbacks, CCIT must periodically conduct customer surveys and report the results of analysis of these customer surveys to the upper management as many similar service based units within organizations do.

Customer relationship management is also important because CCIT needs to manage customer perception of the importance of CCIT services. For instance, many university employees benefit from improved CCIT services; however, in their feedback, many those people say that their do not use CCIT services at all. This finding made the CCIT management realize that they need to do a better internal marketing and committed resources for this cause. A CRM group was put in place permanently to work with the help desk and the strategic planning group.

A customer relationship manager is considered to be a bridge between the customers and the company managers, which must identify important trends in customer sentiment and communicate them convincingly to other corporate managers. The corporate managers then need to take actions to mitigate the problems.

The task of customer feedback analysis requires important several steps:

(1) Gather the customer feedback messages through call center or Web survey.

(2) Read the customer messages (sometimes repeatedly) to discover sensitive and recurrent themes.

(3) Summarize the recurrent themes while reading the messages.

(4) Categorize the customer messages to support the most important themes the manager(s) consider as issues.

(5) Compare with previous customer relationship initiatives to identify improvements in customer sentiments.

(6) Propose corporate actions and estimate resources and impacts of such actions.

Steps (2) to (5) are very time consuming and may take days of work to go over a few hundreds of messages. As our interviews with CRM managers indicated, going over thousands of messages is close to impossible because of time constraints. Consequently, any tools that can help speed up the message analysis process and improve the quality of message categorization will be invaluable to customer relationship management.

Our empirical design includes control and test groups, who are both given the same amount of time to familiarize with the CRM messages. Only some of the test groups have access to our toolset. We are comparing the outcomes to see how well each group is able to analyze the same collection of messages. The metrics used in the field study include the number of valid issues identified, the proportion of correct answers to a set of specially designed questions based among others.

We are currently working with the CCIT CRM managers in order to evaluate the quality of automated pre-processing at each step. A more extensive field study will be our next step once all the algorithms and parameters are tuned based on the data that we currently have.

Conclusions

We have analyzed the possibility to alleviate information overload in CRM data contained in a large collection of text messages. We have built a proof of concept prototype and validated the outcome of each step, thus exploring the applicability of the modern text technologies. The existing CRM systems do not contain such tools and therefore are inadequate in assisting CRM managers to analyze their valuable customer information. We believe that the CRM Sense Maker is a first step towards resolving the information overload problem in CRM message analysis.

The CRM sense maker can be used in more advanced CRM message analyses such as longitude analysis of CRM messages. By partitioning a collection of CRM messages into moving windows and applying the CRM Sense Maker in the moving windows, one can observe changes in the message patterns in time. This longitude analysis can be applied in two ways. First, significant changes in customer sentiment represent either a worsening or an improvement of an existing problem. This can be used as a generic management surveillance system. Second, when companies have invested in certain area of business strategically and expect certain business results, the CRM Sense Maker can be used to ascertain if the strategic goal has been achieved or not and to what degree as measured by customer sentiment.

Currently, we are continuing with more field tests and study visual representations such as semantic maps and the us e of customer ratings. Semantic maps will help managers to observe the results of analysis in a graphical manager, thereby making the CRM Sense Maker easier to use. The use of

customer ratings can potentially make the analysis more accurate since the same phrase used in two CRM messages can have different implications if the two messages have significantly different ratings. For instance, a negative phrase in a CRM message accompanied by high rating of service might be a humorous statement while the same phrase in another CRM message linked to a low rating need to cause an alarm to the CRM manager.

References

Card, S.K., Robertson, G.G., & York, W. (1996), "The WebBook and the Web Forager: An Information Workspace for the World-Wide Web," *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems,* pp. 111-119, Vancouver, Canada.

Chen, H., Hsu, P., Orwig, R., Hoopes, L. and Nunamaker, J.F. (1994), "Automatic concept classification of text from electronic meetings," *Communications of the ACM*, Vol 37 No 10, pp. 56-73.

Cutting, D.R., Karger, D.R., Pedersen, J.O., & Tukey, J.W. (1992), "Scatter/gather: A cluster-based approach to browsing large document collections," *Proceedings of the Fifteenth Annual International ACM Conference on Research and Development in Information Retrieval,* pp. 318-329.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987),"The Vocabulary Problem in Human-System Communication," *Communications of the ACM*, Vol 30 No 11, pp. 964-971.

Gallupe, R.B., and Cooper, W.H. (1993),"Brainstorming Electronically," *Sloan Management Review*, Vol 35 No 1, 1993, pp. 27-36.

Hearst, M.A. (1997), "Interfaces for Searching the Web," *Scientific American*, March 1997, pp. 68-72.

Hiltz, S.R., and Turoff, M. (1985), "Structuring Computer-Mediated Communication Systems to Avoid Information Overload," *Communications of the ACM*, Vol 28 No 7, pp. 680-689.

Ide, N. and Véronis, J. (1998), "Word sense disambiguation: The state of the art," *Computational Linguistics*, Vol 24 No 1, pp. 1-40.

Morphy, E. (2002), "Global CRM Poised for Takeoff," *http://www.crmdaily.com/perl/story/18659.html*, *CRMDaily.com*, July 18, 2002.

Orwig, R.E., Chen, H., & Nunamaker, J.F. (1997), "A graphical, self-organizing approach to classifying electronic meeting output," *Journal of the American Society for Information Science*, Vol 48 No 2, pp. 157-170.

Roussinov, D., and Chen, H., (1999). "Document Clustering For Electronic Meetings: An Experimental Comparison Of Two Techniques," *Decision Support Systems*, Vol 27 No 1-2, pp. 67-79.

Roussinov, D. and Zhao, J. L. (2003), "Automatic Discovery of Similarity Relationships through Web Mining," *Decision Support Systems*, Vol 35 No 1, pp. 149-166.

Swift, R.S. (2001), Accelerating customer relationships, Using CRM and Relationship Technologies, Prentice Hall.

Wang Baldonado, M.Q., & Winograd, T. (1997), "SenseMaker: An information-exploration interface supporting the contextual evolution of a user's interests," *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, pp. 11-18. Atlanta, GA.

Zhao, J. L., Kumar, A., and Stohr, E. A. (2000), "A Dynamic Grouping Technique for Distributing Codified-Knowledge in Large Organizations," *Proceedings of the 10th Workshop on Information Technology and Systems,* December 9-10, 2000, Brisbane Australia.

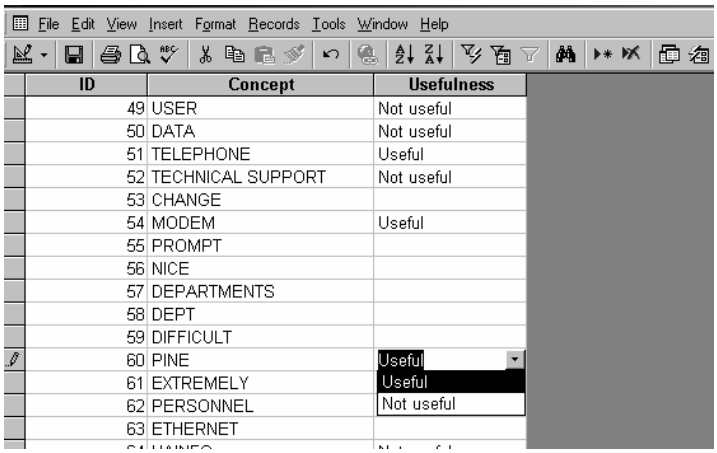

Figure 1. Interactive refinement of the descriptive terms.

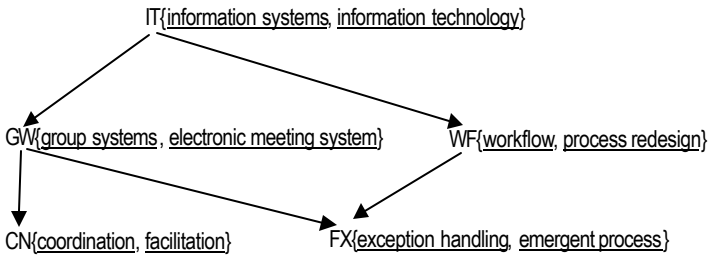

Figure 2. A Generic Similarity Network (Adopted from Zhao & Kumar & Stohr, 2000).

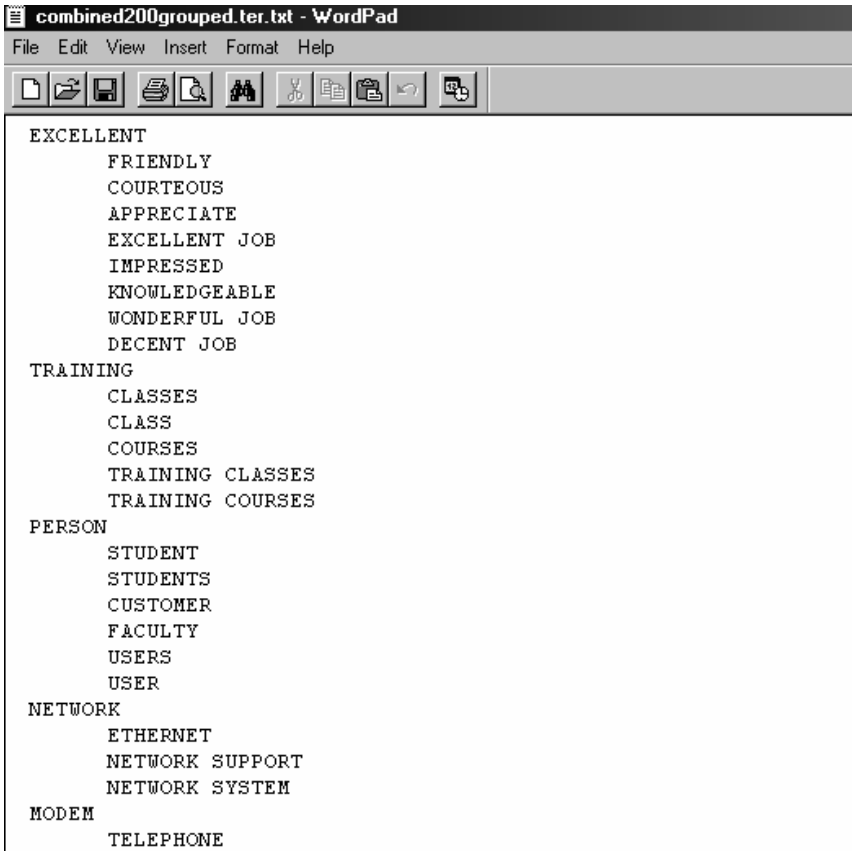

Figure 3. An example of a file showing related concepts grouped together.

*** New Issue: CLASSES

I attended two lecture/classes for computing during the month of November but I'm not sure if they were CCIT classes: 1) Denise Warren - Web Design 2) Copyright Laws (Web) They were both excellent. I look forward to more of the same.

08 In general, my interactions have been very satisfactory. I am thankful to have an efficient and easy access to the internet. However, I was really disappointed when you quit offering your free classes for Macintosh users. There are many of us who use Macs on campus and much prefer them to IBM. Please bring back the Mac classes!

More accessibility to services (I.e., help and other informational aspects). Maybe offer classes to help users with different programs. I am not aware of how useful the CCIT is in enchanting my computer use.

08 I gave it an 8 because last year we got to attend a free "Introduction to Computers" class. To give it a 10 I suggest giving free classes to grounds people on programming irrigation boxes. I am speaking about what helps me. I know almost nothing about computer services outside my department.

*** New Issue: EXCELLENT JOB

Excellent work in meeting UA needs during peak volume for SIS. System went down once, I understand, or I would have rated 10. The center is doing a excellent job.

I think you do an excellent job however it would be nice to be up from 7:00 - 7:00 everyday. Also, more messages to users about downtimes. The help line should have a recording telling us when SIS is expected to be up. We are totally dependent on SIS.

You have been doing an excellent job. However, my office computer is very behind (386) & does not have e-mail or internet. I took a Faculty Development class but the equipment does not measure up to the knowledge.

FYI - name & name did an excellent job of choosing the appropriate computers for our office & helping us to set up Windows NT.

Excellent job. Send chocolate to earn a score of 10!

Figure 4. CRM Messages semi-automatically organized into issues.