Making Sense of CRM Messages: an Interactive Toolset

Dmitri Roussinov
School of Accountancy and Information Management
College of Business, Arizona State University,
Box 873606, Tempe, AZ 85287-3606
dmitri.roussinov@asu.edu

J. Leon Zhao
Department of Management Information Systems
School of Business and Public Administration
University of Arizona, Tucson, AZ 85721
lzhao@bpa.arizona.edu

Abstract

CRM managers are frequently overloaded with large number of text messages from their customers, and making sense of those messages is a difficult task. We introduce an interactive toolset that can facilitate the exploration of CRM data semi-automatically. This toolset is based on the state of the art text processing technologies and allows CRM managers to discover interactively the re-occurring issues and trends in vast amounts of customer messages. Our design of the toolset is based on an iterative sequence of steps: 1) identifying descriptive terms, 2) identifying semantic relationships between the terms, and 3) grouping CRM messages into clusters of related issues. This paper presents a prototype implementation of the toolset, the justification for the specific toolset design choices, and an ongoing field study with CRM managers at a computer customer support center in a large university.

Keywords: Customer relationship management, computer mediated communication, information retrieval, text clustering, text mining.

1 Introduction

CRM managers are frequently overwhelmed with information, making effective customer relationship management a difficult task because manual exploration of all messages from customers is very often time prohibitive, biased, redundant and is lacking comprehensiveness. As a result, there is a great need for an effective toolset 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 information system context (Hiltz & Turoff, 1985; Gallupe & Cooper, 1999).

On the other hand, nowadays CRM data is typically supplied through computer mediated communication (CMC), thus many of the existing CMC tools and research results may apply to CRM. The general CMC approach to Information Overload Reduction is to impose structure on the data (Hiltz & Turoff, 1985). Much work on imposing structure in CMC domain has been done in a Group Decision Support Systems context. Numerous studies have explored 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 et al., 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. Our research integrates existing text clustering techniques into a user-friendly toolset for CRM managers and conducts a field study with CRM managers and other users to validate the toolset. Furthermore, we have also developed a new text summary methodology based on our recent work on automatic discovery of concept similarities to combat the notorious vocabulary problem.

This paper presents a proof of concept prototype along with its theoretical justification (next section) and the ongoing field study with CRM managers at a computer customer support center in a large university.

2 A System Framework

Our approach proceeds through the following 3 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.

2.1 Identifying Descriptors

Following the commonly accepted the "bag of words" approach (Salton & McGill, 1983), the content of each text message is described by words and phrases that this message contains. 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".

Using the Vector Space Model (Salton & McGill, 1983), which is still 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 the most frequent terms is used for vector representation. According to Chen, et al. (1994), Orwig et al. (1997), Roussinov and Chen (1999), this approach works best with small collections consisting of short text messages.

The accuracy of this vector representation is crucial for every text technology involved, being it 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, will likely to be necessary for the technologies to be applicable in real-life (e.g. managerial) applications.

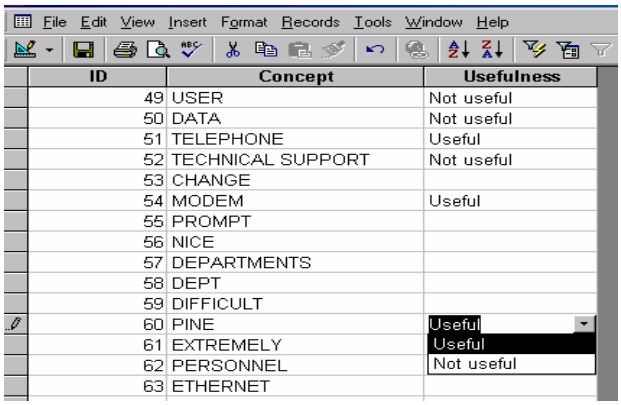

Figure 1. Interactive refinement of descriptive terms.

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.

2.2 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.

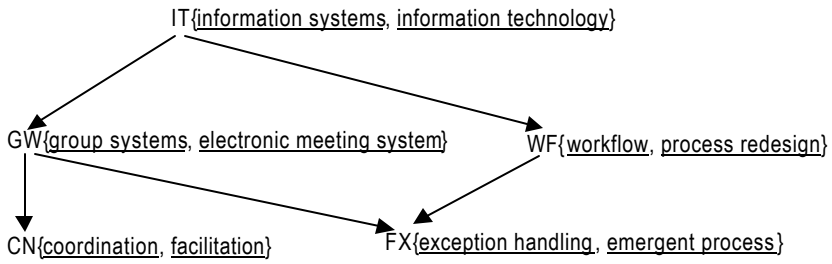

Figure 2. A Generic Similarity Network (Adopted from Zhao et al., 2000).

Roussinov & Zhao (2002) 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.

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, 2002). Then, a CRM manager can refine them.

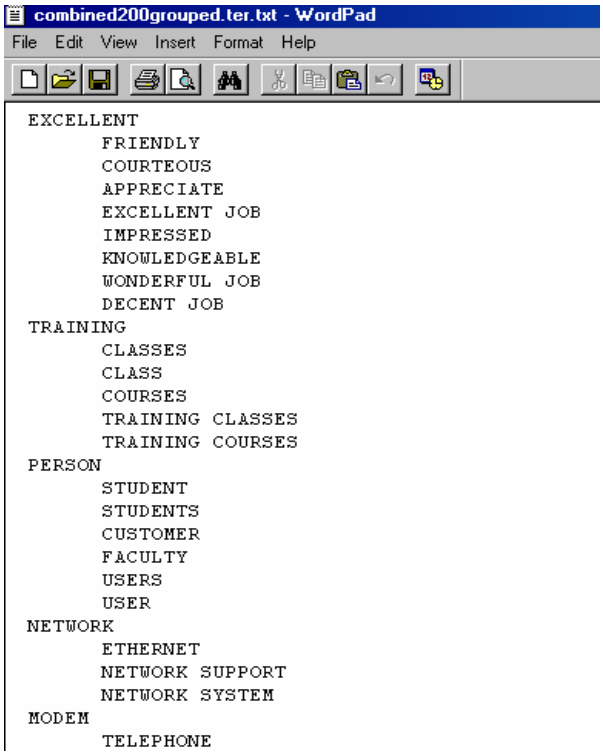

Figure 3. An example of a file showing related concepts grouped together.

2.3 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, 1996), and SenseMaker (Wang & 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 the most representative term (e.g. CLASSES or EXCELLENT JOB) and started with a marker "*** New Issue". The messages are separated by an empty line. Once again, the initial grouping and assigning labels to clusters is done automatically. Then, the user (CRM manager) can manually clean up the groups or just glance over them to identify re-occurring issues.

*** 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.

3 Empirical Evaluation

Since our prototype system includes several components mentioned above, each of them is being tested and validated empirically. Currently, we have 1438 CRM messages collected in the period of several years by a computer customer support center in a large university. We are currently working with CRM managers in order to evaluate quality of automated pre-processing at each step. A field study will be our next step once all the algorithms and parameters are tuned up based on the data we currently have or will collect in future.

We are also designing an experiment, which will include control and test groups who will analyze the same CRM messages with similar training. Only some of the test groups will have access to our toolset. We will then compare the outcomes to see how well each group is able to analyze the same collection of messages. The metrics used in the field study includes the number of valid issues identified, the proportion of correct answers to a set of specially designed questions based among others.

4 Conclusions and Future Research

We have analyzed the possibility to alleviate information overload in the analysis of large collections of CRM messages. We proposed a framework consisting of several semi-automatic steps and have built a proof of concept prototype, thus exploring the applicability of the modern text technologies to as information overload reduction technique in CRM applications.

Since the achievements reported in the past mainly related to completely automated approaches, our work is novel and can lead to an unexplored and promising venue of research. In the future, we will continue with more field tests and also study visual representations such as semantic maps and the use of non-text attributes (such as date and time, customer rating, etc).

While the focus of this paper is on customer relationship management, our study has also implications to other Knowledge Management applications, especially in situations when Information Overload is a major issue, such as in Group Decision Support System, corporate workflow, email analysis and filtering.

References

1. 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.
2. 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*, 37(10), pp. 56-73.

3. 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).
4. Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(11) (pp. 964-971).
5. Gallupe, R.B., and Cooper, W.H. Brainstorming Electronically, *Sloan Management Review*, v35n1, Fall 1993, pp. 27-36.
6. Hearst, M.A. (1997). Interfaces for Searching the Web. *Scientific American*, March (pp. 68-72).
7. Hiltz, S.R., and Turoff, M. (1985). Structuring Computer-Mediated Communication Systems to Avoid Information Overload, *Communications of the ACM*, 28(7), pp. 680-689.
8. Ide, N. and Véronis, J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, 241, pp. 1-40.
9. 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*, 48(2) (pp. 157-170).
10. Roussinov, D., and Chen, H., (1999). Document Clustering For Electronic Meetings: An Experimental Comparison Of Two Techniques, *Decision Support Systems*, (27)1-2, pp. 67-79.
11. Roussinov, D., and Zhao, L. (2002, forthcoming), Automatic Discovery of Similarity Relationships through Web Mining, *Decision Support Systems*.
12. Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York. McGraw-Hill.
13. Wag 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.
14. 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.