Message Sense Maker: Engineering a Tool Set for Customer Relationship Management

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
College of Business and Public Administration
University of Arizona, Tucson, AZ 85721
lzhao@bpa.arizona.edu

Abstract

To determine the important trends and issues in thousands of comments from customers and make strategic decisions about business operations, managers must go over these messages manually and try to make sense of them in a time consuming and tedious manner. There is an urgent need for technologies that help improve the efficiency of customer message management. We develop new issue identification techniques based on clustering and context aware similarity networks to enable managers to discover knowledge in text messages. We engineer a tool set specifically for exploring short text messages in the context of customer relationship management. In this paper, we report a proof of concept prototype called Message Sense Maker that can assist managers to map the overall sentiment of customers semi-automatically. We further justify the choice of particular technologies and validate our system through a field study of a customer support center in a large university.

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

1. 1. Introduction[1]

Managing customer relationships requires effective communications with customers frequently so that companies can monitor customers' sentiment towards the products and services provided by the companies. One such means of communication is customer survey in which customers rate the products or services in some closed or open questions. Beside numerical ratings, companies also value customers' comments in free texts since they will help companies to discover potential problems in the products or services. In recent years, customer survey data comes often in digital messages via Web survey forms. Making sense of thousands of customer messages is an important task in customer relationship management.

One of the tasks of a customer relationship manager is to collect feedback from customers and analyze the text messages for changes in customer sentiments. This requires a customer relationship manager to sip through hundreds or even thousands of text messages to search for major shifts in customer opinions. This is very time consuming and extremely tedious. As a result, much of the valuable customer feedback is not fully utilized.

Up to our knowledge, no summarizing techniques have been studied in CRM applications. In this paper, we present a toolset called Message Sense Maker (MSM) that is designed to assist CRM managers in analyzing thousands of customer messages and discovering recurrent issues and trends. Our tool integrates existing state of the art artificial intelligence technologies into a user-friendly environment.

This tool would address the well-known problem of information overload in the context of computer mediated communication (CMC) [11], [9]. Hiltz & Turoff [11] suggested imposing structure on the data as a possible technical remedy to the information overload problem in CMC context. Several studies have explored this avenue by automatically summarizing CMC messages by representing them with a list of most representative topics [3], using concept maps [15], or clustering messages into semantically homogeneous groups [17]. The common belief behind those approaches is that automated context summarizing techniques can reduce the cognitive load of communication participants even if manual post-processing is still required.

Our research is based on the business environment in the Center for Computing and Information Technology of a major university (CCIT). 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,

[1] This paper is an extension to Roussinov and Zhao, Making Sense of CRM Messages: an Interactive Toolset, AIS 2002 Americas Conference on Information Systems, August 9-11, 2002, Dallas, TX.

resources for research, resources for administrative systems, and campus networking. One of the important missions of CCIT is to improve the quality of services. To accomplish this mission, CCIT must conduct customer surveys periodically and report the results of analysis of these customer surveys to the upper management.

Customer relationship management is important also 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 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 make 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.

One of our research goals is to develop new techniques for analyzing customer survey messages and validating the prototype system in the CCIT environment. Currently, we have collected 1438 CRM messages collected in the period of several years by the CCIT CRM group.

We are currently working with the CCIT CRM managers in order to evaluate the accuracy of each of the automated techniques involved. A field study will be our next step once all the algorithms and parameters are tuned based on the data that we currently have or will collect in future.

In the rest of the paper, we present the design and implementation of a message analysis toolset for customer relationship management, which we refer to as Message Sense Maker (MSM). One of our unique contributions is the application of a novel text processing technique called *context sensitive similarity networks* [18].

2. Message sense maker

At the highest level the toolset can be viewed as consisting of the components performing the following tasks:

1) Identifying semantic descriptors of the messages.
2) Identifying semantic similarity relationships between descriptors through web mining.
3) Clustering Messages into Issues
4) Categorizing new messages into previously identified issues.

The sections below present more details.

2.1 Identifying semantic descriptors

In order to evaluate automatically the similarity between messages or to categorize messages into pre-existing categories, each of the messages has to be digitally encoded. Since there has not been any convincing evidence so far that any other way is better than the commonly accepted "bag of words" approach [19] and the resulting Vector Space model, we opted for adopting them. Thus, each message is represented by words and phrases that it contains through the process called *automatic indexing* which extracts all words and phrases occurring more than once in the collection. Upon removing words from a "stop-word" (like "the", "a", "on", "in") the messages are encoded according to Vector Space Model [19]. Each coordinate in the vector space corresponds to a possible term (word or phrase), set to 1 if a term is present in the document, and to 0 otherwise. We preserve only the specified number of the most frequent terms for computational efficiency. According to [3], [15]who studied collections of meeting messages that are similar in size and style to the CRM messages, this approach seems the most promising. In addition, we suggest that manual cleaning of context bearing terms is a very useful step, which we have implemented using MS Access database shown on Figure 1. A CRM manager who works with MSM 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 results in those terms influencing clustering and categorization decisions more than the other terms.

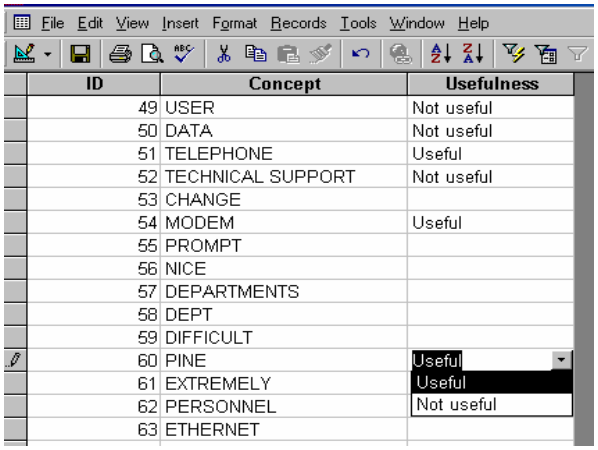


Figure 1. Manual refinement of descriptive terms.

TDT technologies are partially based on the relatively more mature text categorization technology. Text categorization is the task of building software tools capable of classifying text (or hypertext) documents under predefined categories. Text categorization has witnessed a booming interest in recent times, due to the availability of larger numbers of documents in digital form and to the growing needs to organize them. A number of learning techniques have been applied to text categorization, including multivariate regression, nearest neighbor classifiers, probabilistic Bayesian models, decision trees, and neural networks. Lewis and Hayes [13] gave a comprehensive overview of the related works. In each, a text document is represented by the words and phrases that it has (see next section for more details) and the learning algorithms are presented with a number of positive and negative examples for each category. The accuracy exceeding 90% has been reported with the commonly used testing collections such as Reuters newswires.

Message Sense Maker uses Text Categorization technology to decide if a new customer message falls into any of the existing issues (categories). Since the issues have been manually refined and enough positive and negative examples accumulated, we are expecting the high accuracy to be provided by this component.

2.2 Mining semantic similarity network from the Web

Clustering short text messages imposes a severe limitation on using existing text clustering and representation techniques because they typically expect the documents to be sufficiently large and share at least some terms in order to be treated as semantically similar. The notorious vocabulary problem [8] caused by the fact that different people use different words for same or very similar concepts, makes automated clustering extremely difficult and the outcome not always intuitive. After studying customer messages we observed that the vocabulary problem is clearly demonstrated in our CRM data, for example the words *user*, *customer*, *teacher*, *student*, *people* refer to the same concept of a *customer*. Thus, messages mentioning customers or users would not have been necessary placed into the same cluster by MSM, if it did not use the semantic similarity network to alleviate the vocabulary problem as explain in the sections below.

While solving the problem in the most general setting does not seem to be feasible in the nearest future, we alleviate it by using an Organizational Concept Space (OCS) [23] framework. OCS includes a so-called *similarity network*, a collection of similarity relationships between the important concepts (words and phrases). For example, OCS would note the similarity between `customer` and `user`.

Roussinov & Zhao [18] 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 "mined" from the World Wide Web.

MSM also uses the OCS approach to alleviate the vocabulary problem. Below, we detail the steps involved.

STAFF	LABS
CAMPUS	TECHNICAL
STUDENTS	COMPUTING
EXCELLENT	USERS
TRAINING	EXPERIENCE
CLASSES	LAB
PERSON	RESPONSE
STUDENT	FRIENDLY
NETWORK	EMPLOYEES
DEPARTMENT	SORRY
OFFICE	ACCOUNTS
CUSTOMER	ACCOUNT
FACULTY	ASSISTANCE
RATE	SYSTEMS
SLOW	TECHNOLOGY

Table 1. Top 30 most frequent concepts in the CRM data set in order of frequency.

2.5.1. Representing the business context. According to [18], OCS is effective only in specific business context, e.g. inside one organization, or one specific meeting. E.g. in our CRM context of the words *customer* and *student* are almost synonyms, which may not be the case in the more context of education. MSM represents the context automatically by the top most frequently occurring

concepts (words or phrases). Table 1 lists 30 such concepts for our CRM data.

2.5.2. Downloading context specific pages from the Web for mining**.** In order to perform effective data mining, Message Sense Maker automatically downloads thousands of pages from the Web that are semantically close to the identified context. It constructs queries for the underlying commercial search engine (www.altavista.com in the current implementation) in such a way that the matching document have to include the top 30 concepts and very likely to include the other concepts describing the context. Table 2 presents a fragment of such queries. The spider component included in Message Sense Maker downloads 200 pages from each query. More details on how to automatically interface with commercial search engines can be found in [18].

+STAFF CAMPUS STUDENTS EXCELLENT TRAININ(
+CAMPUS STAFF STUDENTS EXCELLENT TRAININ(
+STUDENTS STAFF CAMPUS EXCELLENT TRAININ(
+EXCELLENT STAFF CAMPUS STUDENTS TRAININ(
+TRAINING STAFF CAMPUS STUDENTS EXCELLEN
+CLASSES STAFF CAMPUS STUDENTS EXCELLENT
+PERSON STAFF CAMPUS STUDENTS EXCELLENT
+STUDENT STAFF CAMPUS STUDENTS EXCELLENT
+NETWORK STAFF CAMPUS STUDENTS EXCELLEN
+DEPARTMENT STAFF CAMPUS STUDENTS EXCEL

Table 2. A fragment of queries generated by Message Sense Maker for AltaVista search engine from the top 30 most frequent concepts in the CRM data set.

Many researchers and practitioners believe that the World Wide Web is a gold mine filled with useful information. Indeed, such vast amount of textual and multimedia information was not available for researchers before the mid 90s. According to Lyman and Varian [14], the Web currently contains more than 2.5 billion of pages, consisting of at least 10 terabytes of textual information. Although there exist multiple definitions, in this study, under "Web mining" we mean automated discovering of semantic associations between the specified terms. For a comprehensive review of the Web mining literature, please refer to Cooley et al. [4].

2.5.3. Indexing mining collection**.** The HTML pages are converted into plain text using and truncated to the first *20,000* bytes of text in order to avoid overly long web pages. They are processed by the same automatic indexing procedure as mentioned in the preceding section and represented by vectors. We use 0/1 encoding instead of popular TF-IDF weighting for the mined collection and normalize the vectors to the unit length afterwards. Since conversion is a one-pass algorithm, it is runs relatively quickly so downloading is currently the only bottleneck for performance. However, if performed in parallel, it can be sped up considerably [16], so that the entire process can be implemented to run in real time. Also, if search engine index is available through Application Interface (API), the downloading would not be necessary at all. Currently, many commercial search engines starting to allow this kind of API.

2.5.4. Performing data mining in the downloaded collection**.** It has been known for a long time that the relationships between concepts (words or phrases) can be discovered by their co-occurrence in the same documents or in the vicinity of each other within documents. Firth, a leading figure in British linguistics during the 1950s, summarized the approach with the memorable line: "You shall know a word by the company it keeps." [7] The classical work of van Rijsbergen [20] initiated the use of co-occurrence information for text retrieval and categorization. To obtain the numerical value of similarity, we used the previously suggested [20], [19]

term:COLLABORATION related: COLLABORATIVE 3.877262e-001 COLLABORATIVE SYSTEMS 3.381566e-001 SYSTEMS 2.656920e-001 term:FACILITATION related: COLLABORATION 1.341769e-001 FACILITATORS 1.665471e-001 term:HARDWARE related: NETWORK 2.911064e-001 SUPPORT 2.496007e-001 SYSTEMS 3.189309e-001 term:MEETING related: FACILITATORS 2.335312e-001 INFORMATION 2.949944e-001 MEETINGS 6.949747e-001term:REMOTE related: EXAMPLE 2.259506e-001 NETWORK 2.501757e-001 SYSTEMS 2.463332e-001 term:VOICE related: TECHNOLOGIES 2.254068e-001 TECHNOLOGY 2.486785e-001 NETWORKS 2.637258e-001 SYSTEMS 2.067017e-001

Table 3. A fragment of a listing with mined similarity relationships from [18].

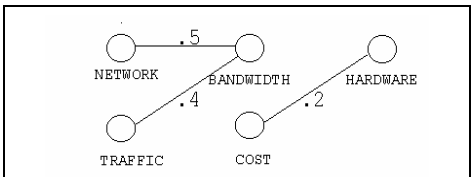
formula:

$S_{ij} = \mathbf{t_i} \cdot \mathbf{t_j} / |\mathbf{t_i}| |\mathbf{t_j}|$,

where S_{ij} is the similarity between terms *i* and *j*; $\mathbf{t_i}$ and $\mathbf{t_j}$ are vectors representing occurrences of the terms i and j in the documents in the mined collection. In order to make computation more tractable Message Sense Maker truncates for each term the mined relationships to only *three* (*3*) other most strongly related terms. A fragment of the mined Concept Space from a different dataset (Roussinov & Zhao, 2002) is shown in as example Table 3, where each term is followed by 3 most closely terms along with the similarity between them.

5) *Applying Similarity Network*. As it is done in [18], Message Sense Maker modifies each message vector **V** by the following transformation: **V** = **V** + *a* **O** ´ **V**, where **O** is the matrix representing OCS, and *a* is the adjustment factor. The product **O x V** is the vector **V** multiplied by the matrix **S** defined by a common linear algebra.

For example, the message "bandwidth concerns -- impact of remote collaboration" originally represented by concepts `BANDWIDTH, CONCERNS, IMPACT, REMOTE` and `COLLABORATION` would also receive the concept `NETWORKS` if Message Sense Maker captured semantic similarity between the concepts of `NETWORKS` and `BANDWIDTH` (as shown in Figure 2 below.



	N	Tr	P	Cu	Te	B	Co	H
Network (N)	1	0	0	0	0	.5	0	0
Traffic (Tr)	0	1	0	0	0	.4	0	0
Problem (P)	0	0	1	0	0	0	0	0
Current (Cu)	0	0	0	1	0	0	0	0
Technology (Te)	0	0	0	0	1	0	0	0
Bandwidth (B)	.5	.4	0	0	0	1	0	0
Cost (Co)	0	0	0	0	0	0	1	.2
Hardware (H)	0	0	0	0	0	0	.2	1

Figure 2. An example of Similarity Network and its matrix representation from [18].

This way the message would be automatically evaluated as similar to the message "`Effective transmission of video over networks`" and would likely to be placed in the same cluster of network related messages even though those two messages do not share any words in common. This example illustrates the alleviation of the vocabulary problem.

It is worth noting that even when *a* is small and the resulting modifications are small, the resulting accuracy increase may be still significant. This is because without the modifications, most messages did not share any common terms and had the same similarity *0*, if dot product used to compute it). This poses a problem for clustering (and other) algorithms that have to resolve many "ties" in order to form clusters. Even small changes in the coordinates can break those "ties" in the right direction, thus considerably helping the clustering algorithm to make more accurate choices.

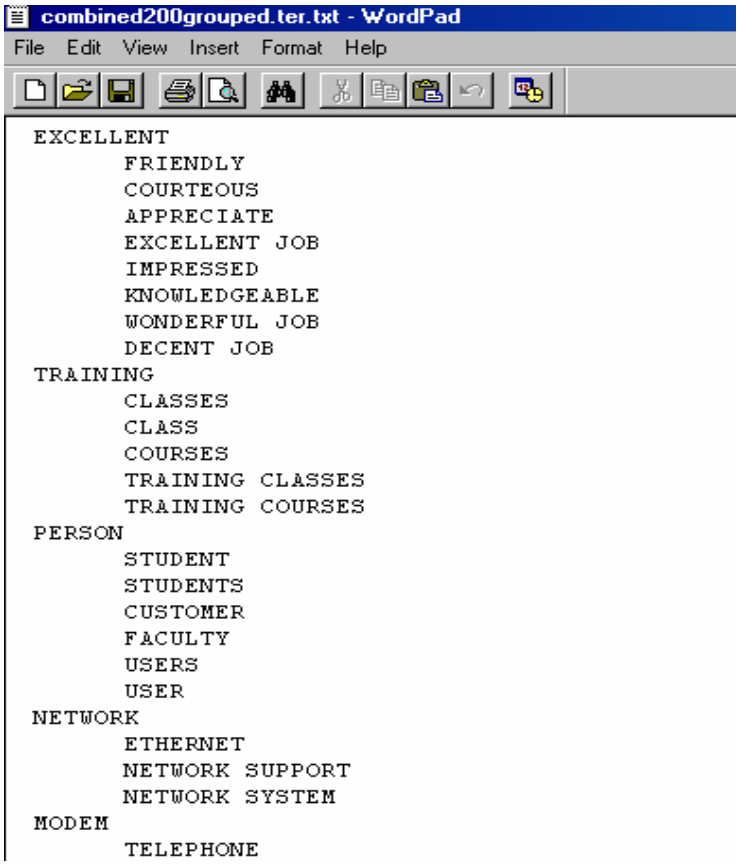


Figure 3. Manual refinement of OCS.

2.5.5. Manual refinement**.** Message Sense Maker combines 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 [18]. Then, a CRM manager can refine them.

2.3 Clustering messages into issues

We defined an issue as a set of messages related through common themes. Since the CRM messages are typically short, they usually have one, but sometimes up to 2 or 3 themes. This way, the messages related to the same issue (theme) have to be semantically similar, thus the issue can be identified through the process of automated document clustering. Everitt [6] defined a cluster as "a set of entities which are alike, and entities from different clusters are not alike." The entities are considered alike or not alike based on the features that

they have (e.g. messages are descried by words and phrases) and the similarity measure (e.g. negative Euclidean distance between vectors). More detailed about features and similarity metrics are in the following sections.

*** 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.
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.
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.

Table 4. CRM Messages organized 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 [5], WebBook [2], and SenseMaker [21]. Hearst [10] gives a comprehensive overview of such systems and the ideas behind them.

Message Sense Maker organizes messages into issues (clusters) through a semi-automatic interactive procedure. Table 1 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. 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.

Message Sense Maker uses Ward's hierarchical agglomerating clustering technique [22]. The algorithm starts with each document in a cluster of its own and iterates by merging the two most similar clusters until all the documents are merged into a single cluster and the entire collection is represented by a binary tree called a *dendogram*. Then, the resulting dendrogram is converted to a set of non-overlapping clusters (partition). MSM allows users to set the limit to the cluster size. This is useful, since as we observed, users do not create clusters manually with more than 8-15 messages in them.

2.4 Issue detection and tracking

Once the initial set of issues (clusters) has been identified, there is a need to fit new incoming messages into existing set of issues or to recognize new issues when they arise. Message Sense Maker is using the technology recently developed by other researches participating in the Topic Detection And Tracking (TDT) initiative [1], sponsored by Defense Advance Research Program Agency (DARPA) and National Institute of Standards (NIST). TDT investigates the state of the art in finding and following new events in a stream of broadcast news stories. The TDT problem consists of three major tasks: (1) segmenting a stream of recognized speech into distinct stories; (2) identifying those news stories that are the first to discuss a new event occurring in the news; and (3) given a small number of sample news stories about an event, finding all following stories in the stream. Those tasks are very similar to those pursued by CRM managers so we are exploring possible extension of TDT technology to the CRM domain.

3. Conclusions

Managing thousands of customer messages and discovering recurrent themes and changes in customer sentiment are time consuming and tedious. New technologies are needed to assist managers to manage these customer messages efficiently. In this paper, we reported a prototype system called Message Sense Maker that can assist managers to identify trends and issues in the incoming flow of customer messages. Our unique technical contribution is to apply a new text clustering technique based on context sensitive similarity networks.

In the future, we will perform empirical studies and develop visual representations such as semantic maps. Our empirical test will include 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 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.

In addition, we will integrate our text clustering techniques with non-text attributes (such as date and time, customer rating, etc) to improve the precision of message clustering. While this paper focuses on customer relationship management, the basic ideas and techniques stemming from our study can also be applied in other business applications such as information distribution services, knowledge management and computer mediated communication.

References

[1] Boykin, S., and Merlino, A. (2000). Machine learning of event segmentation for news on demand", *Communications of the ACM*, 43(2), February 2000, pp. 35-41.

[2] 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.

[3] 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.

[4] Cooley, R., Mobasher, B. and Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web (with R. Cooley and J. Srivastava), in *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997.

[5] 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).

[6] Everitt, B.S. (1974). *Cluster Analysis*. New York. John Wiley & Sons, Inc.

[7] Firth, J. A (1957). Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis*, Philological Society, Oxford; reprinted in Palmer, F. (ed. 1968) *Selected Papers of J. R. Firth*, Longman, Harlow.

[8] 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).

[9] Gallupe, R.B., and Cooper, W.H. Brainstorming Electronically, *Sloan Management Review*, v35n1, Fall 1993, pp. 27-36.

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

[11] 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.

[12] Ide, N. and Véronis, J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, 241, pp. 1-40.

[13] Lewis, D.D. and Hayes (1994). Special issue of *ACM:Transactions on Information Systems* on text categorization, 12(1), July 1994.

[14] Lyman, P. & Varian, H. (2000). How Much Information? A project report of the Regents of the University of California, available at http://www.sims.berkeley.edu/how-much-info.

[15] 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).

[16] Roussinov, D., & McQuaid, M. (2000). Information Navigation by Clustering and Summarizing Query Results," *Proceedings of Hawaii International Conference on System Sciences (HICSS-33)*, January 4-7, 2000, Island of Maui.

[17] 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.

[18] Roussinov, D., and Zhao, J.L. (2002a). Automatic Discovery of Similarity Relationships through Web Mining, *Decision Support Systems* (forthcoming).

[19] Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York. McGraw-Hill.

[20] van Rijsbergen, C.J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2), pp. 106--119.

[21] 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.

[22] Ward, J. (1963). Hierarchical grouping to optimize an objection function. *Journal of the American Statistical Association.*, 58 (pp. 236-244).

[23] 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.