Knowledge Perception Analysis in a Social Network

Nishith Pathak, Sandeep Mane, Jaideep Srivastava
Department of Computer Science
University of Minnesota
Minneapolis, MN, USA
{npathak, smane, srivasta}@cs.umn.edu

Noshir S.Contractor
Department of Speech
Communication and Psychology
University of Illinois at Urbana-Champaign, Urbana, IL, USA
nosh@uiuc.edu

ABSTRACT

Knowledge management in organizations is gaining in importance, especially with the advent of computer networks. Networks foster interaction between individuals, and have become the medium of choice for all types of interactions, both professional and social. In this research, we study the perception of knowledge in an organization's email network. An important aspect of an individual's knowledge is that it may be incomplete and hence any analysis approach must handle knowledge uncertainty. We propose an approach based on the Dempster-Shafer theory of evidence for modeling individuals' perceptions about knowledge, thus enabling the understanding of knowledge in an organization. We show how correlating the knowledge of two or more individuals can help identify the discrepancies between them, and thus identify sources of organizational misperceptions. The proposed approach has been evaluated on the publicly available e-mail logs from the Enron Corporation. For the present study, meaning extraction from e-mail content was done manually. Initial results show that the approach is very promising. Our continuing research is focusing on applying techniques for automated identification of knowledge from email as well as sentiment analysis techniques for automated evaluation of individuals' sentiments.

Keywords

Social network analysis, Knowledge networks, Dempster-Shafer theory, sentiment analysis, Enron email corpus.

1. INTRODUCTION

Social network analysis (Wasserman and Faust, 1994) is a widely studied field of research, where a researcher is interested in understanding "who knows who". Understanding such social networks in an organization is important as it affects the organization's as well as individuals' performance. Such (informal) social networks can be quite different from the organizational (formal) hierarchy, and can be a powerful channel for spread of information (or misinformation). An important aspect of social network analysis is thus to know which individual (who) has what information (knowledge) and also how this knowledge spreads in the network. Thus, a step further in social network analysis is to understand "who knows what", which is referred to as *knowledge network* (Contractor, 2000). Knowledge networks are important in an organization as proper exchange of information among different individuals fosters research collaboration.

The widespread use of computer networks in organizations is enabling members in far-flung locations to carry on significant amounts of interaction. This has enabled individuals to form social relationships (contact) with other individuals without geographical constraint. Identifying knowledge sources among such individuals in a social network and making sure that proper information is being transmitted is useful for the managers in the organization. Electronic communication such as electronic-mail (e-mail) and instant messages (chats) between employees are usually logged in an organization. The text fields, such as message text in e-mail and text in instant message, provide an insight into the knowledge exchanged between individuals. The use of such data for identifying 'the knowledge and the knowledgeable' in an organization is the main problem of interest in this research. We concentrate on the analysis of knowledge in email exchanges between individuals in an organization. We explain how Dempster-Shafer theory of evidence can be applied to model individuals' knowledge acquisition process based on what they observe, and then illustrate how this can be used to understand the overall knowledge in the organization. In this paper, we concentrate on a socio-centric analysis of a knowledge network. This is still a work in progress research, and we show preliminary experimental results using the Enron email corpus. The results are promising and further research is being pursued.

The rest of the paper is organized as follows – Section 2 provides a brief background on knowledge networks and related research that has used the Enron email corpus. Section 3 describes the proposed approach using Dempster-Shafer theory of evidence. Section 4 illustrates the preliminary results obtained using the Enron email data. Section 5 concludes the paper and discusses the current research in progress.

2. RELATED WORK

Knowledge management in an organization has been an active research field for the last few years. Hansen (2002) shows how an organization may benefit from using knowledge residing in its different sub-units. A knowledge network is basically a network which connects individuals (or actors) to resources (or knowledge) (Seufert et al 1999). Another definition of a knowledge network is – "a knowledge network is a special case of social network, where the links represent shared or related knowledge" (Jones, 2001). Different mechanisms have been proposed to be the driving factors for the evolution of linkages in a knowledge network (see Contractor et al, in preparation). Palazzolo et al (in preparation) show how the Theory of Transactive Memory helps to understand how nodes seek knowledge from other nodes in a knowledge network. An important point to be considered in case of a knowledge network is that the linkages between individuals (actors) and knowledge are imprecise, and hence computational models used for analyzing such networks should allow for incorporation of this uncertainty

The Enron Email corpus was made public in 2003, and since then it has been used for different kinds of analyses. Initial research concentrated on methods for cleaning this real-world data. Bekkerman et al (2005) illustrated a method for automated classification of such email data into folders. A lot of research effort has been directed to identifying social networks and structures in such networks (see Workshop on Link Analysis, Counterterrorism and Security in SIAM International Conference on Data Mining, 2005). For example, Yitao et al (2005) illustrate how link analysis can be used for discovering structures in social networks among email users. Shetty and Adibi (2005) apply an information-theoretic model to this email dataset to find centrally important nodes and closed groups around them. Diesner and Carley (2005) showed that the communication intensity between individuals increased during the Enron crisis. In this paper, we are interested in understanding how individual knowledge perceptions evolve in the email network. However, till now due to lack of availability of such large datasets about individuals' knowledge, not much research has been done to use data mining techniques for such analysis.

3. PROPOSED APPROACH

The basic idea behind knowledge perception analysis is to analyze the knowledge network at various time instances as knowledge propagates in the underlying social network. We record the various stages of the evolution of a knowledge network across time. Analysis of these various stages of the knowledge network then provides valuable insight as to how knowledge perceptions evolve in the underlying social network. We analyze a social network where the medium of communication is electronic-mail (email).

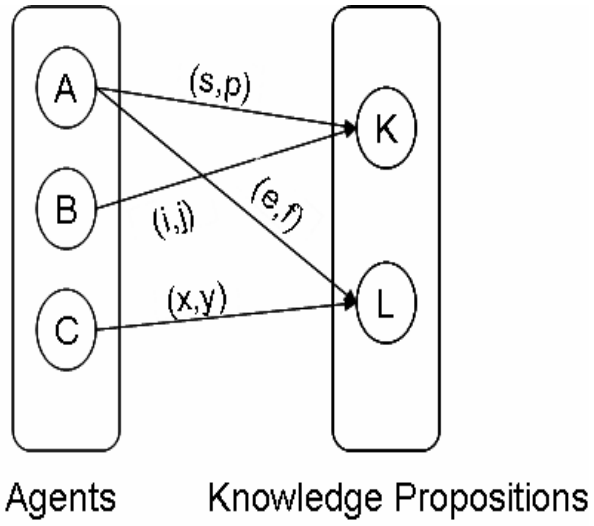

Figure 1. Knowledge Network as a Bipartite Graph.

We represent the knowledge network as a bipartite graph; where one partite set consists of nodes corresponding to the agents participating in the social network and the other partite set consisting of nodes corresponding to the *knowledge propositions* (see Figure 1). We represent knowledge as a collection of statements which can be true or false. A knowledge proposition is one such statement which can be either true or false. People can believe it to be true or false with a certain probability and they can also be uncertain about this probability value. An example of a knowledge proposition would be a statement like *'The company image is good'*. This statement can be believed to be either true or false with some uncertainty, by different agents participating in a social network in an organizational environment. An edge from an agent A to a knowledge proposition K implies that agent A has a certain belief regarding the truthfulness of K. Each edge has a binary tuple associated with it. The tuple is of the form *(s, p)* where $0 \leq s \leq 1$, $0 \leq p \leq 1$, and $s \leq p$. This tuple corresponds to the belief state of an agent A regarding a knowledge proposition K. The tuple quantifies how *probable* it is that agent A believes K is true and also of how *certain* agent A is of this chance of K being true. Here, *s* and *p* are called the *support* and *plausibility* values respectively, of agent A perceiving K to be true. A detailed explanation of these support and plausibility values and the tuple is provided later in this section.

3.1 Dempster-Shafer Theory

In our proposed approach, the knowledge network graph is constructed and updated using Dempster-Shafer theory (Shafer, 1976). Dempster-Shafer theory is a generalization of the Bayesian theory of subjective probability (Dempster, 1968). It is also known as theory of beliefs. The theory is a valuable tool for combining evidences obtained from multiple sources. In Dempster-Shafer theory there exists a set of mutually exclusive alternatives called the *frame of discernment* Θ. For example, if we were reasoning about an organization's public image, then Θ consists of the following mutually exclusive alternatives –

$$\Theta = \{good, bad\}.$$

A function m: $2^{\Theta} \rightarrow [0, 1]$ is called basic probability assignment if it satisfies the following constraints,

$$m(\phi) = 0, \quad \sum_{A \subseteq 2^{\Theta}} m(A) = 1$$

The quantity *m(A)* is defined as A's basic probability number. It represents the exact belief in the proposition K. The basic probability assignments can be inferred from various evidences by using the combination rules of Dempster-Shafer theory. For example, suppose we have evidence that tells us that the company image is good with a confidence of 0.6 and another piece of evidence that says that the company image is bad with a confidence of 0.3. Let us consider each of these evidences in detail. For the first evidence we have a basic probability assignment

$m_1(\phi) = 0$
$m_1(\text{'good'}) = 0.6$
$m_1(\text{'bad'}) = 0$
$m_1(\text{'good or bad'}) = 0.4$

For the second evidence the basic probability assignment is,

$m_2(\phi) = 0$
$m_2(\text{'good'}) = 0$
$m_2(\text{'bad'}) = 0.3$
$m_2(\text{'good or bad'}) = 0.7$

Note that in probability theory, we have *m('bad')=1-m('good')*. However, according to Dempster-Shafer theory, the first evidence talks only in favor of the company image being good but does not say anything about the image being bad therefore, the probability *1-m_1('good')* is assigned to the event when we are uncertain about the status of the company image, i.e. 'good or bad'. In other words, even if this evidence is flawed, it is still possible that the company image is good and not necessarily bad. Therefore, if the evidence turns out to be flawed, then we are uncertain of the company's image, and hence the 0.4 chance of the evidence being flawed is assigned to *m_1('good or bad')* instead of *m_1('bad')*. Similarly, in case of the second evidence, the probability *1-m_2('bad')* is assigned to *m_2('good or bad')*. This is the main difference between probability theory and Dempster-Shafer theory. In Dempster-Shafer theory, we have the option of a state of uncertainty whereas in Bayesian probability theory we do not. Dempster-Shafer theory also allows us to combine the two evidences to provide a basic probability assignment, which takes into account both the evidences. The Dempster-Shafer theory rule for combining two evidences is,

$$m(A) = \frac{\sum_{X \cap Y = A} m_1(X) \cdot m_2(Y)}{1 - \sum_{X \cap Y = \phi} m_1(X) \cdot m_2(Y)} \qquad (1)$$

Here, m_1 and m_2 are two basic probability assignments of different evidences, on the same frame Θ. If the denominator is zero, then the two evidences are said to be *totally contradictory* and cannot be combined. The log of the denominator is called the *weight of conflict* between the evidences. After combining the two evidences under consideration we obtain the following basic probability assignment,

$m(\phi) = 0$
$m(\text{'good'}) = 0.51$
$m(\text{'bad'}) = 0.15$
$m(\text{'good or bad'}) = 0.34$

The probability number *m('good')* is called the *support* for the company image being good and the value *1-m('bad')* is called the *plausibility* for the company image being good. In our context where knowledge propositions can be either true or false, we have the support value *s* as the probability with which we can strongly state a proposition to be true and the plausibility value *p* as the maximum possibility of the proposition being true. Note that the *support* is always less than or equal to *plausibility* and the difference *plausibility-support* is the "uncertainty" in the proposition being true. In probability theory, we have no state of uncertainty, hence plausibility = support = the probability of the proposition being true.

The generalized formula for combining n evidences is,

$$m(A) = \frac{\sum_{\cap X_i = A} \prod_{1 \le i \le n} m_i(X_i)}{1 - \sum_{\cap X_i = \phi} \prod_{1 \le i \le n} m_i(X_i)} \qquad (2)$$

3.2 Knowledge Network Construction and Updating

In our model for a given agent A and a given knowledge proposition K, we associate a frame of discernment $\Theta_{A,K}$ = {true, false} corresponding to the mutually exclusive alternatives of proposition K being true or false. In the knowledge network graph, this is represented by an edge from agent A to the knowledge proposition K carrying a label (s, p) (see figure 2). For the edge (A, K), s and p are respectively the support and plausibility values for K being true. Note that since we have only two elements in $\Theta_{A,K}$, the tuple (s, p) can be used as a representation for the basic probability assignment for K with respect to agent A's perspective, where,

$m(\phi) = 0$
$m(\text{'good'}) = s$
$m(\text{'bad'}) = 1 - p$
$m(\text{'good or bad'}) = p - s$

Thus, the tuple (s, p) can be treated as the belief state of agent A regarding the knowledge proposition K.

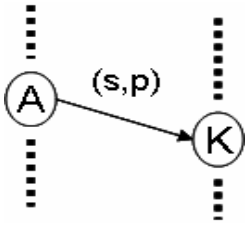

Figure 2. Agent A is associated with Knowledge Proposition K

The higher is the value of *s*, the more confident is agent A's belief in K being true and the higher the difference *p-s*, the more uncertain is A regarding this belief for K. Suppose agent A receives an email from another agent B which provides evidence for or against proposition K being true. If the evidence is in favor of *x*, where *x* can be good or bad, with a confidence *s*, then the basic probability assignment for this evidence is done by letting $m(\phi) = 0$, $m(x) = s$, $m(x^c) = 0$ and $m(\text{' } x \text{ or } x^c \text{ '}) = 1-s$. If the sentiment of an email is neutral, then no evidence is extracted from the email. Suppose the email from agent B speaks in favor of the proposition K being true with a confidence of 0.9. We treat this email from agent B as an evidence of K being true which carries a confidence of 0.9. Agent A himself entertains some initial basic probability assignment reflected by the tuple *(s, p)*, corresponding to his perception regarding the verity of K. Using Dempster-Shafer theory's rule of combination we can combine this new evidence with agent A's belief state to obtain the updated belief state of agent A, represented by the new tuple *(s´, p´)*, which now becomes the new edge label of edge (A, K) in the knowledge network graph. The email sent by agent B is also used as an evidence for/against the knowledge proposition K to update agent B's belief state. This is because since agent B has sent the email, the evidence obtained from the email must also be reflected in the sender's belief state. Thus, the evidence obtained from an email is combined along with the recipients' as well as the sender's belief states. For both sender and recipient, the confidence for the evidence is same and the belief states are updated using Dempster-Shafer theory as explained above. The following example further elucidates the use of Dempster-Shafer theory for combining evidence with an agent's belief state.

Consider an agent A with belief state (0.8, 0.9) regarding the truthfulness of some knowledge proposition K. Now its belief state is to be combined with evidence that speaks against the truthfulness of K with a confidence of 0.7. The *s* and *p* values in agent A's belief state are 0.8 and 0.9 respectively. From these, we get the basic probability assignment for agent A's belief state to be – $m_A(\phi) = 0$, $m_A(\text{'true'}) = 0.8$, $m_A(\text{'false'}) = 0.1$ and $m_A(\text{'true or false'}) = 0.1$. The basic probability assignment for the given evidence is $m_E(\phi) = 0$, $m_E(\text{'true'}) = 0$, $m_E(\text{'false'}) = 0.7$ and $m_E(\text{'true or false'}) = 0.3$. Using Dempster-Shafer theory's rule of combination, we get the combined basic probability assignment to be –

$$m(\phi) = 0$$

$$m(\text{'true'}) = \alpha \cdot (m_A(\text{'true'}) \cdot m_E(\text{'true'}) + m_A(\text{'true'}) \cdot m_E(\text{'true or false'}) + m_A(\text{'true or false'}) \cdot m_E(\text{'true'}))$$

$$m(\text{'false'}) = \alpha \cdot (m_A(\text{'false'}) \cdot m_E(\text{'true'}) + m_A(\text{'false'}) \cdot m_E(\text{'true or false'}) + m_A(\text{'true or false'}) \cdot m_E(\text{'false'}))$$

$$m(\text{'true or false'}) = \alpha \cdot (m_A(\text{'true or false'}) \cdot m_E(\text{'true or false'})$$

Where,

$$\alpha = \frac{1}{1 - (m_A(\text{'true'}) \cdot m_E(\text{'false'}) + m_A(\text{'false'}) \cdot m_E(\text{'true'})}$$

On solving we get,

$m(\phi) = 0$
$m(\text{'true'}) = 0.54$
$m(\text{' false'}) = 0.39$
$m(\text{'true or false'}) = 0.07$

The new support $s' = m(\text{'true'}) = 0.54$, and
new plausibility $p' = 1 - m(\text{' false'}) = 0.61$

Thus, the updated belief state for agent A is the tuple *(s´, p´) = (0.54, 0.61).*

Input:

- Set of agents **A**
- Set of Knowledge Propositions **KP**

Output:

- Set of belief states **B,** where b ∈ **B** corresponds to an agent a ∈ **A** updated perception regarding knowledge proposition kp ∈ **KP**,

Pseudo code:

/* *Set initial s and p values to 0 and 1 respectively**/

1. For (a,k) in (**A** × **KP**) do
2. KN_Graph[a,k].s = 0
3. KN_Graph[a,k].p = 1

/* Update_Knowledge_Network, i.e. combine Evidence e with agent a's perception of knowledge proposition k */

/* *First, determine basic probability assignment for evidence e* */

1. If e.sentiment = true then
2. $m_e(T)$ = e.confidence
3. $m_e(F)$ = 0
4. else
5. $m_e(T)$ = 0
6. $m_e(F)$ = e.confidence
7. end if
8. $m_e(T \text{ or } F)$ = 1-e.confidence

/* *Determine basic probability assignment for agent a's belief state* */

9. $m_a(T)$ = KN_Graph[a,k].s
10. $m_a(F)$ = 1-KN_Graph[a,k].p
11. $m_a(T \text{ or } F)$ = KN_Graph[a,k].p-KN_Graph[a,k].s

/* *Combine the evidence and agent a's belief state to get updated basic probability assignment for agent a's belief state* */

12. $\alpha = 1 / [\, 1-(m_e(T)m_a(F) + m_e(F)m_a(T)) \,]$
13. $m(T) = \alpha\{ m_a(T)m_e(T) + m_a(T)m_e(T \text{ or } F) + m_a(T \text{ or } F)\, m_e(T) \}$
14. $m(F) = \alpha\, [\, m_a(F) \times m_e(F) + m_a(F) \times m_e(T \text{ or } F) + m_a(T \text{ or } F) \times m_e(F) \,]$
15. $m(T \text{ or } F) = \alpha\, [\, m_a(T \text{ or } F) \times m_e(T \text{ or } F) \,]$

/* *Update agent a's belief state in the knowledge network* */

16. KN_Graph[a,k].s = m(T)
17. KN_Graph[a,k].p = 1-m(F)
18. End for
19. Return KN_Graph[*,*].b

Algorithm 1 Evidence Updating in Knowledge Network.

If agent A does not have any prior beliefs regarding K and this is the first time it comes across any evidence regarding K then for the sake of applying combination we assign agent A's default belief state to be that of complete uncertainty by letting $s = 0$ and $p = 1$. Therefore, in the knowledge network, if there is no edge between a given agent and a given knowledge proposition, we assume the s and p values for such a pair to be 0 and 1 respectively. Algorithm 1 shows the pseudo code for knowledge network construction and updation.

Thus, we can construct and update the knowledge network graph using Dempster-Shafer theory. As emails are exchanged in the network, we extract evidences and their confidences from them for the different knowledge propositions of interest. Then, using Dempster-Shafer theory as illustrated in this section, we update the knowledge network graph for all pairs of agents and knowledge propositions. If no evidence regarding a knowledge proposition is perceived by an agent, then its belief state regarding that proposition does not change. The updated knowledge network graphs at the end of regular time instances are recorded and used to analyze the perception of knowledge in the underlying social network. In the next sub-section, we explain how to obtain the evidences and their confidence values from the email text.

3.3 Evidence Acquisition from Email Text

The first task in our approach is extracting the knowledge propositions of interest from the e-mail's text message. A related work is by Berry and Browne (2005), who show an interesting approach for automatically identifying semantic features (topics) from emails as well as clustering emails, thus removing the need for manually reading the emails. Further research into knowledge proposition extraction is being pursued by us. However, for our initial experimental results shown later, we have manually defined the knowledge proposition of interest.

Once the knowledge propositions of interest have been defined, a piece of evidence consists of a Boolean sentiment value indicating whether the actor speaks for or against (positive or negative) the knowledge proposition and a confidence value or degree (between 0 and 1) for this claim (similar to Turney and Littman, 2003). Thus, an important task is to extract the sender's opinions about the knowledge propositions of interest. For this, we believe that sentiment classification techniques will be a useful tool. To provide a brief background, sentiment analysis using automated techniques has gained interest in computer science research community with the advent of internet. The availability of data about the user's opinions/reviews on the Web has triggered the evaluation of several (machine

learning, NLP or text mining) approaches for sentiment analysis (see Turney, 2002; Pang et al, 2002; Wiebe et al, 2001; Bai et al, 2004). Pang and Lee (2005) illustrate an approach for converting sentiments into a rating measure. However, sentiment analysis is a difficult task and more research is still remains to be done. For example, Pang et al (2002) illustrate the difficulty in using bag-of-words machine learning methods for sentiment classification. In addition, sentiment analysis still remains a very domain-specific problem, where classifiers trained for one domain may not perform well in others. (Aue and Gamon, 2005).

In our analysis, automatic sentiment classification from the email will enable us to evaluate large amounts of data. However, in present stage of our research, we manually identified the sentiments of users for a pre-defined knowledge proposition. One justification for this is that the number of emails, that we found to have the knowledge proposition of interest, was small. Hence, we are not sure how the current sentiment detection techniques will perform for our experiments. Further research in this problem is currently being pursued.

3.4 Model Architecture

We now combine the concepts discussed so far and present the complete model architecture which can be used to construct and update the knowledge network, as knowledge propagates in the underlying social network.

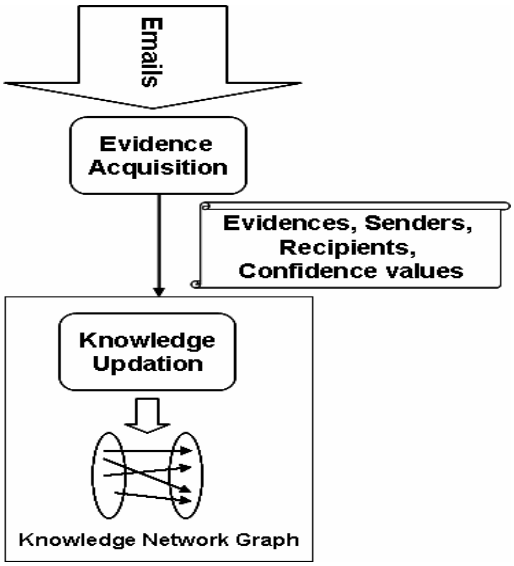

Figure 3. Model Architecture

As shown in Figure 3, the proposed model architecture consists mainly of three components. It is essential that this model is allowed to observe all the email communication that takes place in the network. A good idea would be to have the model residing in the email server.

As an email arrives, the evidence acquisition module analyzes the text of an email message and acquires the evidences for interesting knowledge propositions from them. An email can act as at most one piece of evidence for a given knowledge proposition. If the email talks about more than a single knowledge proposition, then we can extract multiple evidences, one piece of evidence for each of the knowledge propositions that the email talks about. The knowledge updating module takes as input all the evidences compiled from an email, the recipients and the sender of these evidences and the confidence values of these evidences, and updates the knowledge network graph using Dempster-Shafer theory rule of combination (as illustrated previously). As more emails are exchanged in the network, new evidences are gathered from these emails. The updated knowledge network graph is recorded at regular time intervals for analysis for knowledge perception in the social network.

4. INITIAL EXPERIMENTAL RESULTS

4.1 Enron Email Corpus

The Enron email corpus is the set of emails belonging to 158 users, mostly senior management of Enron. The emails were exchanged (and hence logged) during the time period mid 1998 to end 2001 (approximately 3.5 years), thus spanning the Enron crisis which broke out in October 2001. The email corpus is publicly available and can be downloaded from the website http://www.cs.cmu.edu/~enron/. This is a cleaned version of the dataset after the original dataset had been subject to various processes including removal of all email attachments and resolution of multiple email-ids of the same person into one. It consists of about approximately 200,399 email messages. A brief description of the data is given by Klimt and Yang (2004) and a statistical report is provided by Shetty and Adibi (2004).

4.2 Preliminary results

For our preliminary experiments, we used the labeled dataset made available by the Enron Email Analysis Project at the University of California, Berkeley (http://bailando.sims.berkeley.edu/enron_email.html). This dataset is a subset of about 1700 labeled email messages. These emails focus on business-related content as well as content relating to the California Energy Crisis, and emails that occurred later in the collection, avoiding very personal messages, jokes, etc.

Table 1. Designation of for/against sentiment and confidence values for different categories of email evidences

Knowledge Proposition K: 'The company image is/remains good'		
Category Label	**Sentiment (for/against)**	**Confidence Value**
Very Good	For	0.9
Good	For	0.5
Slightly Good	For	0.1
Neutral	NA	NA
Slightly Bad	Against	0.1
Bad	Against	0.5
Very Bad	Against	0.9

Table 2. Number of emails in each category

Category Label	# of emails
Very Good	5
Good	15
Slightly Good	22
Neutral	43
Slightly Bad	15
Bad	15
Very Bad	3
	Total = 118

These emails are classified into 8 major categories namely, business-related, purely-personal, personal but in a professional context, logistic arrangements, employment arrangements, document editing/checking, empty messages due to missing attachments and empty messages. The email messages in the first category i.e. the business-related category are further classified into sub-categories namely, regulations and regulators, internal projects, company image (current), company image (changing), political influence/ contributions/contacts, California energy crisis/California politics, internal company policy, internal company operations, alliances/partnerships, legal advice, talking points, meeting minutes, trip reports.

For preliminary analysis, we chose to use only one knowledge proposition of interest, i.e. *'The company image is/remains good'*. We chose to use only those emails in the company image (current) as well as company image (changing) sub-categories. These two sub-categories consisted of a total of 118 emails. The contents (text message) of each of these emails were examined manually and judged for its sentiment regarding the company image.

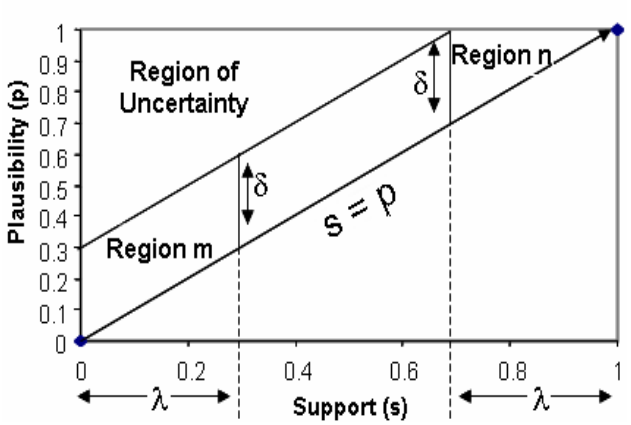

Figure 4. Knowledge Network as a scatter-plot

The emails were manually classified into seven categories indicating the impact on the company image namely, very good, good, slightly good, neutral, slightly bad, bad and very bad. With each of these categories, we designated a specific confidence and for/against sentiment. This designation is shown in table 1. Table 2 shows the number of emails assigned to each category.

The reader is reminded that emails that display neutral sentiment are discarded and no evidences are extracted from them. Since, there is only one knowledge proposition of interest, only one piece of evidence is extracted per email. The evidences extracted from these emails along with their sender and recipients were then fed to the knowledge updating module (in increasing order of time) and the users' beliefs were allowed to evolve over time. A total of 118 users were identified to be involved in these email exchanges. The initial *s* and *p* value for each user was taken to be 0 and 1 respectively. Due to the small size of the data set we chose to record the knowledge network graph at the end of every year i.e. 1999, 2000 and 2001. Among the 75 non-neutral emails, the number of emails belonging to the year 1999, 2000 and 2001 were 3, 21 and 51 respectively. The timeline of these emails was from December 1999 to October 2001.

We created a scatter plot for the knowledge network at the end of each year. Agents are plotted in the x-y plane with support *s* along the x-axis and the plausibility *p* along the y-axis (Figure 4). The line *s=p* represents points where there is no uncertainty regarding the probability of the verity, of the given knowledge proposition. As we move farther above this line, the uncertainty in the belief increases (region of uncertainty). The region m near the origin i.e. points which have uncertainty less than some δ and with *s* value at most some λ, contains points which believe the proposition more likely to be false with low uncertainty.

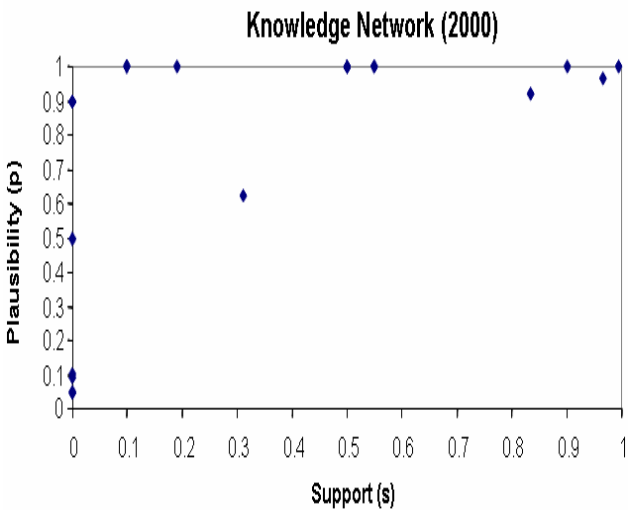

Figure 5. Knowledge Network plot for the year 2000

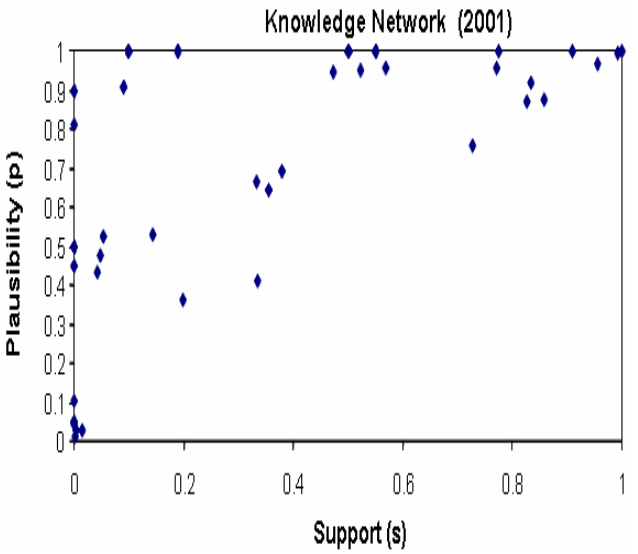

Figure 6. Knowledge Network plot for the year 2001

Similarly, the top-right hand region n, consists of points which believe the proposition more likely to be true with low uncertainty. Since $s \leq p$ always, all points always lie above the line $s = p$.

The graph for 1999 is not shown as not enough data was available to provide anything interesting. During the year 2000, there were many events such as positive press articles, a highly positive article in time magazine and talk of the most innovative company of the year award, that sought to improve the company image, as well as certain events such as miscommunication within company resulting in bad press and bad press due to environmental and human rights violation in overseas ventures, that tarnished the company image. The plot for 2000 is shown in figure 5. During the year 2001, the company image dropped mainly due to negative press generated during the California power crisis. However, things started improving when there was talk of a possible merger with a rival company, Dynergy. The plot for 2001 is shown in figure 6. The interesting part about the result is the lack of consensus among users regarding the company image. Note the existence of a significant number of points in the regions n and m (see figures 4, 5 and 6) for both 2000 and 2001. The number of uncertain people is also quite significant. The fact that the points are pretty much spread out in both the graphs leads us to infer the existence of a lack of harmony among the users' perceptions regarding the company image, throughout both the years.

5. CONCLUSIONS

In this paper, we have presented an approach for analyzing knowledge perception in a social network. We provided an evidence-theory based methodology for constructing and maintaining a knowledge network in an electronic-mail communication environment. Our experimental results using the proposed approach on a subset of the Enron email data has shown some interesting results.

The proposed approach has various applications in an organizational environment. It can be used to monitor the flow of information in an organization, ensure consistent knowledge and resolve misperceptions among participating users. Another important application can be monitoring of employees' sentiments regarding certain sensitive topics such as company image, change in policies etc. The approach is also a better substitute for the various intra-company surveys carried out by organizations, as it does not suffer from traditional manual survey problems such as individual bias.

To summarize, the main contributions of this ongoing research are –

1. We explained the need for incorporating uncertainty in the analysis of knowledge exchange in a social network.
2. We proposed an approach based on combining Dempster-Shafer theory with sentiment analysis for analysis of knowledge perception in an email network.
3. Our preliminary results using this method on a subset of Enron email data are promising and further research is being pursued.
4. The proposed approach is flexible in that it can be extended to incorporate an individual's reliability in determining confidence of evidence provided that user.

Sentiment analysis plays an important role in determining individual perceptions and is the current related research

being investigated by us. Future directions include exploring the use of knowledge predicates instead of knowledge propositions and methods to incorporate users' reliability in evidences provided by them.

6. ACKNOWLEDGMENTS

Nishith Pathak's work is supported by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under Cooperative Agreement number DAAD19-01-2-0014. Sandeep Mane's research is supported by NSF grant (IIS-0431141).

7. REFERENCES

[1] Aue, Anthony and Gamon, Michael. (2005) Customizing Sentiment Classifiers to New Domains: a Case Study. International Conference on Recent Advances in Natural Language Processing, (*in submission)*.

[2] Bai, Xue and Padman, Rema and Airoldi, Edoardo. (2004) Sentiment Extraction from Unstructured Text using Tabu Search-Enhanced Markov Blanket. *Workshop on Mining the Semantic Web*, at the *10th ACM SIGKDD Conference*, Seattle, WA.

[3] Bekkerman, Ron, McCallum, Andrew and Huang, Gary. (2004) Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora. CIIR Technical Report IR-418, University of Massachusetts, Amherst.

[4] Berry, M. and Browne, M. (2005) Email Surveillance Using Nonnegative Matrix Factorization. *Computational & Mathematical Organization Theory* 11, pp. 249-264.

[5] Chapanond, A., Krishnamoorthy, M. and Yener, B. (2005) Graph Theoretic and Spectral Analysis of Enron Email Data. Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005.

[6] Contractor, N. (2000). Presentation of IKNOW: Inquiring knowledge networks on the Web. http://www.spcomm.uiuc.edu/contractor/IKNOW/sld001.htm

[7] Contractor, N. S., Whitbred, R., Fonti, F., Hyatt, A., O'Keefe, B., & Jones, P. Structuration theory and the evolution of networks. *In preparation / under review*.

[8] Dempster, A.P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B* 30 205-247.

[9] Diesner, J., & Carley, K.M. (2005). Exploration of Communication Networks from the Enron Email Corpus. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*.

[10] Andrea Esuli. The Sentiment Classification Bibliography. http://patty.isti.cnr.it/~esuli/research/sentiment/Sentiment.html

[11] Hansen, M. T. (2002) Knowledge Networks: Explaining Effective Knowledge Sharing in Multiunit Companies. *Organization Science*, Vol. 13(3), May. 2002, pp. 232-248.

[12] Jones, P. M. (2001), Collaborative Knowledge Management, Social Networks, and Organizational Learning. In M. J. Smith & G. Salvendy (Eds.). *Systems, Social and Internationalization Design Aspects of Human-Computer Interaction.* Vol. 2, pp.306-309. Mahwah, New Jersey: Lawrence Erlbaum Associates

[13] Klimt, B., Yang, Y. (2004). Introducing the Enron corpus, CEAS.

[14] Monge, Peter and Contractor, Noshir. (2001) Emergence of Communication Networks. F. M. Jablin & L. L. Putnam (Eds.) *New Handbook of Organizational Communication* (2nd Ed.), pp 440-502, Newbury Park, CA: Sage.

[15] Palazzolo, E. T., Serb, D., She, Y., Su C., Contractor, N. S. Co-evolution of Communication and Knowledge Networks as Transactive Memory Systems: Using Computational Models for Theoretical Integration and Extensions. *In preparation*.

[16] Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79-86.

[17] Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the ACL.*

[18] Seufert A.; von Krogh G.; Bach A. (1999) Towards knowledge networking. *Journal of Knowledge Management*, Vol. 3(3), March 1999, pp. 180-190.

[19] Shafer, Glenn (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

[20] Shetty, J., Adibi, J. (2004). The Enron email dataset database schema and brief statistical report. Technical Report, Information Sciences Institute, USC.

[21] Shetty, Jitesh and Adibi, Jafar. (2005) Discovering Important Nodes through Graph Entropy - The Case of Enron Email Database. *Proceedings of Workshop on*

Link Discovery: Issues, Approaches and Applications, ACM SIGKDD Conference, 2005.

[22] Turney, Peter. (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics.*

[23] Turney, Peter D. and Littman, Michael L. (2003) Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, Vol. 21(4), pp.315-346.

[24] Wasserman, Stanley and Faust, Katherine. (1994) *Social Network Analysis – Methods and Applications.* Cambridge University Press.

[25] Wiebe, Janyce, Wilson, Theresa and Bell, Matthew (2001). Identifying Collocations for Recognizing Opinions. *Proceedings of ACL 2001: Workshop on Collocation: Computational Extraction, Analysis and Exploitation,* France.

[26] Yitao Duan, Jingtao Wang, Matthew Kam and John Canny. (2005) A Secure Online Algorithm for Link Analysis on Weighted Graph. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005.*