Dynamic Prediction of Communication Flow Using Social Context

Munmun De Choudhury   Hari Sundaram   Ajita John   Dorée Duncan Seligmann

Arts Media & Engineering, Arizona State University   Collaborative Applications Research, Avaya Labs

Email: {munmun.dechoudhury,hari.sundaram}@asu.edu, {ajita,doree}@avaya.com

ABSTRACT

In this paper, we develop a temporal representation framework for communication and social context to efficiently predict communication flow in social networks. The problem is important because it facilitates determining social and market trends as well as efficient information paths among people. We describe communication flow by two parameters: the *intent to communicate* and *communication delay*. There are three key contributions in this paper. (a) To estimate the intent and delay, we design features to characterize communication and social context. Communication context refers to the attributes of current communication. Social context refers to the patterns of participation in communication (information roles) and the degree of overlap of friends between two people (strength of ties). (b) A subset of optimal features of the communication and social context is chosen at a given time instant using five different feature selection strategies. (c) The features are thereafter used in a Support Vector Regression framework to predict the intent to communicate and the delay between a pair of individuals. We have excellent results (~12% prediction error) on a real world dataset from the largest social networking site, www.myspace.com. We observe interestingly that while context can reasonably predict intent, delay seems to be more dependent on personal contextual changes and latent factors, e.g. 'age' of information and presence of cliques among people.

Keywords

Social networks, Communication flow, Social context, Information roles, Strong and weak ties, MySpace, Feature selection, SVM.

1. INTRODUCTION

In this paper we develop a temporal representation framework for communication and social context to efficiently predict communication flow in social networks. The problem is important because prediction of communication flow can help organizations determine their experts or knowledge points, monitor the dynamics of effective information paths, as well as facilitate understanding community evolution in social networks.

There has been prior work on developing computational models for information diffusion [6,10]. In [6] the authors focus on analyzing the text in blog posts and use an epidemic disease propagation model for determining information diffusion. In [10], the authors present an early adoption based information flow model useful for recommendation systems. The authors in [9] provide simple models for the onset of epidemic behavior in diseases. Prior work [3] also focused on determining the effect on diffusion dynamics in small world networks, due to heterogeneous consumers. There has also been prior work on analysis of emails of software developers [1], to understand the relationship between the email activities and the software roles.

However, there are several limitations of prior work. First, since the semantics of communication are dynamic [4], context plays an important role in the exchange of messages in any communication. However context has not been exploited comprehensively in prior work. Second, the traditional approach to prediction of information flow has only been focused on modeling current context of communication between people [2]. However, communication patterns can be affected by the habitual and network properties that are acquired over time. Third, the factors affecting communication between a pair of individuals evolve over time. From prior psychological studies [8], we know that messages are influenced by the sender's current short-term memory; besides attributes of the current context (topic context, the messages exchanged as well as the neighborhood context.

The main contribution of this work is an effective temporal prediction framework for determining communication flow between members of a social network. Communication flow is described by two parameters [2]: the *intent to communicate* (the probability that a person Alice would communicate with another person Bob) and *communication delay* (the time taken for Alice to send a message to Bob).

In this work, we design features to model social context for the prediction of communication flow. Social context refers to the patterns of participation in communication (information roles) and the degree of overlap of friends between two people (strength of ties). We build upon our earlier work on communication context [2]. We develop a dynamic feature selection algorithm. An optimal subset of the features is chosen at a given time instant by combining five different feature selection strategies. A Support Vector Regression framework is then used for predicting the intent to communicate and delay between a pair of individuals.

We have excellent results on a real world dataset from the largest social networking site, www.myspace.com. Our results indicate that modeling social context is key to determining communication flow. We also notice qualitatively that intent is more affected due to contextual dynamics than delay. Delay seems to be more dependent on other latent factors characterizing communication, including the 'age' of information transmitted and presence of *cliques* among people.

The rest of the paper is organized as follows. In the next two sections, we discuss the communication and social context. In section 4 we present the prediction framework. Section 5 discusses the dataset followed by experimental results. We conclude by discussing the major contributions and directions towards future work in section 6.

2. COMMUNICATION CONTEXT

In this section we briefly review communication context as proposed in [2] since it is useful when predicting communication flow. Communication context [2,10] refers to the set of attributes that affect communication between two individuals. In [2] we identify three aspects that affect communication – (a) neighborhood context, (b) topic context and (c) recipient context. For details readers are referred to [2]. In the subsequent paragraphs, we use the following running example. Assume that we have two users Alice and Bob and a group of people – Bob's contacts. Alice wants to discuss topic Λ with Bob.

Neighborhood context refers to the effect of the user's social network on her communication. There are two network effects of interest – backscatter and susceptibility. Backscatter refers to the fraction of the messages received by Alice from her contacts that are about a topic Λ. Susceptibility measures whether the social network that Alice interacts with is interested in the topic that she plans to communicate on. Intuitively, if a network is susceptible to communication on a certain topic, then Alice is more likely to send a message on topic Λ to her network.

Topic context refers to the effect of the semantics of a user's *past* communication on the topic Λ on her future communication. We are interested in four measures – (a) message coherence (b) temporal coherence (c) topic relevance and (d) topic quantity.

Message coherence refers to consistency in message semantics and the semantic relationships of the messages with the current topic Λ (e.g. 'movies'). Temporal coherence is defined as the correlation of the time-stamps of the messages on a topic received by Alice. High coherence of messages in a recent past would increase Alice's intent to communicate and vice versa. Topic relevance for user Alice on a topic Λ refers to the relationship between *topics in her past communication* to the topic Λ. Topic quantity is the number of topics on which Alice has received messages in the recent past. The effect of topic quantity for a topic Λ on the intent to communicate for user Alice is inversely related to the number of topics k on which Alice has communicated.

Recipient context refers to effect of the recipient identity on Alice's intent to communicate. There are three measures of interest – (a) reciprocity, (b) communication correlation and (c) communication significance.

Reciprocity refers to the ratio of the messages received from the recipient to those sent to the recipient, on the intended communication topic. Communication correlation refers to the topical alignment between a user Alice and her contact Bob with whom she wants to communicate. Communication significance refers to the fraction of past messages to the specific contact v on the current communication topic.

However, in [2], only the attributes that are part of current context have been considered. The communication patterns of people can also be affected by the habitual and network properties that are acquired over time; e.g. *who* is communicating with *whom* and what is the *strength* of relationship shared between them. In the following section we define these contextual factors.

3. SOCIAL CONTEXT

In this section, we discuss social context. *Social context* is the set of attributes that refers to *who* is communicating with *whom* and what is the *strength* of relationship shared between them. These are the habitual and network properties acquired by a person over time. There are two features of interest that we discuss in this section: information roles and strength of ties. Communication flow between individuals is affected by the communicative behavior of each person. For example, Alice might be a person who is very active in sending messages to her contacts. In that case the probability that Alice would communicate with Bob is high. Further, suppose Alice and Bob have several friends in common, which implies they share a strong bond. The strength of the tie between Alice and Bob would also affect the probability that they would communicate.

3.1 Information Roles

Information role is a contextual attribute acquired over time which impacts a person's communication behavior. We formally define three different categories of roles of people: (a) *generators*, people who generate information by themselves or from other sources (e.g. external events like the American Idol or iPhone release), (b) *mediators*, people who act as transmitters of information between people, and (c) *receptors*, people who mostly receive messages. Drawing an analogy with the hyper-linking structure of the web, we notice that the generators and the receptors act like *authorities* and *hubs* respectively in the social network. The roles presented are exhaustive but not mutually exclusive; clearly a person can play different roles, depending on the context.

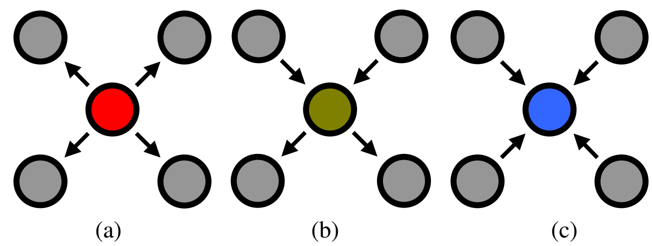

Figure 1: (a) Generators (red) (b) Mediators (green) (c) Receptors (blue). The people shown in gray can belong to any of the three roles.

Let us consider Alice, Bob and Charlie to be part of a social network. The roles of the three (as generators, mediators and receptors) emerge from three different communication structures shown in Figure 1. Alice is a generator characterized by several out-going communication links; Bob is a mediator with comparable number of incoming and out-going links and Charlie represents a receptor with large frequency of incoming links.

We define the information role of a person to depend on the net communication activity, in which she participates. This is given by the following message frequency ratio:

$$R(u,t_i) = \frac{\sum_v n_{u\to v}(t_i)}{\sum_v n_{v\to u}(t_i)}, \qquad <1>$$

where $n_{u\to v}$ is the number of messages sent by u to a contact v at time slice t_i and $n_{v\to u}$ is the number of messages received by u from a contact v at time slice t_i. In our experiments one time slice is assumed to be equivalent to one week. Now we define the following conditions to define the roles: (a) If R (u, t_i) is significantly greater than 1, then the person is a generator, (b) If R (u, t_i) is approximately equal to 1, then the person is a mediator, and (c) If R (u, t_i) is significantly less than 1, then the person is a receptor.

Intuitively, the information roles of a pair of communicators would change over time. Hence their probability of communication would also change due to change of roles. If the duration of past communication of the two communicating people is divided into i time slices, we therefore need to determine their roles as well as the probability that they would communicate after i time slices.

We define a role transition matrix $\mathbf{P}^1$ over the three roles s_1, s_2, s_3 such that $P(s_m, s_n)$ gives the probability that a person in role s_m would communicate with another person in role s_n at the first time slice. The values of the transition matrix are determined empirically from the past communication based on frequency of messages exchanged. Now given this initial role transition matrix, our goal is to determine their probability of communication after i transitions (or time slices). We use the Chapman-Kolmogorov equation which defines a technique to compute i-step transition probabilities by multiplying the initial transition probability matrix i-times. This gives us the role transition matrix $\mathbf{P}^i$ for the two people involved in communication at time slice i. We can now easily determine the two roles at time slice i using eqn. < 1 > and the probability that these two roles will communicate from $\mathbf{P}^i$.

3.2 Strong and Weak Ties

It is well known from prior work [7] that the nature of relationship between two people affects communication. This argument is based on evidences in [7] which suggest that the pattern of relationships between actors (people) reveals the likelihood that individuals will be exposed to particular kinds of information. We categorize the relationships between two people using the strength of shared ties: strong ties and weak ties, similar to [5]. The strength of a tie is proportional to the number of common friends between two persons. The work in [5] emphasizes the real world observation of evidence of strong friendship (or bond) between two persons when there is frequent communication between them. To take an example, suppose Alice and Bob are friends and Alice and Charlie are also friends. The work in [5] considers the induction of a 'psychological strain' for the two pairs. This is because both Bob and Charlie attempt to make their communication congruent with their common friend Alice. This depicts the introduction of a positive tie between Bob and Charlie.

Therefore, we consider the strength of a tie between two people to be dependent on the overlap of their friends' circles. For example, as in the above example, if Alice and Bob have ten common friends then they share a strong tie, while if Alice and Charlie have two common friends, then the share a weak tie. Prior work on the strong and weak ties [5] reveals two claims as follows.

The first claim is that the exchange of *new information* (an external event like London bombings) is higher along weak ties. This is explained by the intuition that the new information can traverse greater social distance along weak ties. Suppose Alice sends a message on a recent movie review to her friends, and those friends send messages to their friends. Many of these people would form a small clique, sharing strong ties. As a result the information traveling through such ties will likely be limited to a small clique of friends. On the other hand, if the new information is transmitted across weak ties, the less overlap of friends is likely to make the new information reach other cliques.

The second claim is that strong ties are better for transmission of *existing information* which is often characterized by execution of an action in the external world (e.g. referral for a job position). This claim works on the ground that weak ties work well when there is a lot of friction among the people (since they stay manageable and provide a fresh perspective). As this friction gets reduced in a social network with hundreds of contacts, weak ties become overwhelming and people ignore the information to cope with information overload. In this case strong ties are more reliable means of transmission.

We therefore observe that the probability of information across strong or weak ties depends upon latent factors, e.g. the 'age' of the information. Determining the age of information in an online social network like MySpace is a challenging problem. Instead in this work, we *learn* the impact of strength of ties on communication by defining the overlap in the friends' circles of the two people. Strength of tie is a symmetric measure, given by,

$$\psi(u,v,t_i) = \left(L(u,t_i) \cap L(v,t_i)\right) / \left(L(u,t_i) \cup L(v,t_i)\right), \quad <2>$$

where $L(u, t_i)$ and $L(v, t_i)$ are respectively the friend lists (vector of contact names/ids) of u and v at time t_i.

In the following section, we introduce a Support Vector regression framework for predicting communication flow which uses the features of communication context discussed so far.

4. PREDICTION

Having modeled the features, we now discuss the prediction framework for determining the intent and the delay in communication. The intent to communicate and delay can be modeled as a regression problem [2] where the relationships between the different model parameters can be learnt over time and for specific individuals. Our regression model for predicting the intent and delay is based on a Support Vector Regression based unsupervised learning as in [2]. The details of the algorithm are provided in [2].

In order to model the temporal aspects of context, it is important to eliminate some features dynamically. This is because the importance of a specific feature towards capturing the communication context at a specific time instant can vary over time. Secondly, the features may be inaccurately estimated. Both of these conditions worsen the ability to predict communication flow. Therefore context is modeled dynamically in our problem by performing feature selection at the beginning of each time slice. We analyze our prediction performance against several popular static and dynamic feature selection techniques: Correlation coefficient, Mutual information, Decision trees, Principal Component Analysis and *k*NN based estimation. We use a voting strategy to determine the best k features at any time instant.

5. EXPERIMENTAL RESULTS

In this section, we discuss the experimental results of prediction. The dataset used for our experiments is similar to [2] and comprises approximately 20,000 users from MySpace who have exchanged about 1,425,010 messages in the time snapshot from September 2005 to April 2007. Topics of communication were detected using WordNet using hierarchical clustering procedures as in [2].

Choosing Optimal *k* features: In this section we describe the experiments performed to find the optimal features (k) that need to be chosen for prediction of the intent to communicate and the delay. We construct 11 different training sets using 1, 2, 3, …, 11 features. The training set spans over a period of 70 weeks (averaged over four different topics A ('person, someone'), B ('entity, abstraction'), C ('event, happening') and D ('party, gathering')). For each of the 11 training sets, we determine the mean errors E_1, E_2, E_3, …, E_{11} in prediction (using the corresponding 1, 2, 3, …, 11 features) over the next 10 weeks. The mean minimum error (~10-15 %) in prediction of intent and delay is found to occur for a set of five features.

However, since context is dynamic, these five features would vary over time. Moreover, the *k* (five) optimal features chosen by each strategy will be different. We therefore identify the optimal five features that were selected at each week (during testing phase) by each of the five feature selection strategies (ref section 4). We thereafter follow a voting strategy to pick the optimal $k = 5$ features over all five strategies. In every time slice (week), for every feature, we determine if the feature has been chosen by at least three strategies. If true, then it is appended to the list of five optimal features. We observe that certain features, e.g. susceptibility, backscatter, information role and strength of tie are selected by all the five strategies in most time slices. This implies the importance of these features in the prediction process.

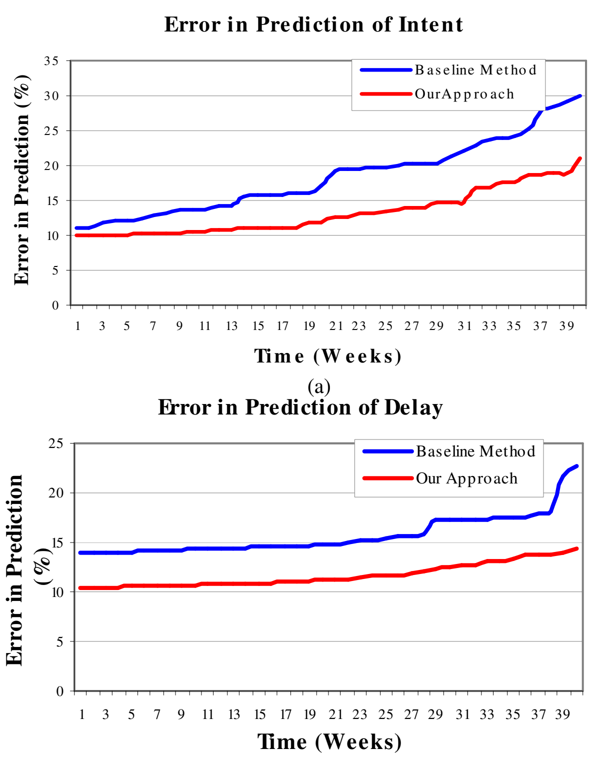

Figure 2: The plots show variation in error across time averaged over eight contacts of a user. We compare our dynamic feature selection technique against a baseline static feature selection method. (a): Comparison of error in intent prediction (b) Comparison of error in delay.

Analytical Comparison with Baseline Method: We, now present some results of prediction of the intent to communicate and delay against a baseline approach where no dynamic feature selection is done – i.e. all 11 features are used for prediction. The experiments are shown for the baseline framework and our approach in Figure 2 over a period of ten weeks and averaged over eight different contacts of a person Charlie and the four topics A-D. We observe significant improvements in prediction accuracy over [2]. The mean error in prediction of intent (with respect to actual communication) is ~12 % and for delay is ~13% using our approach; whereas it is ~19% for intent and ~18% for delay using the baseline method.

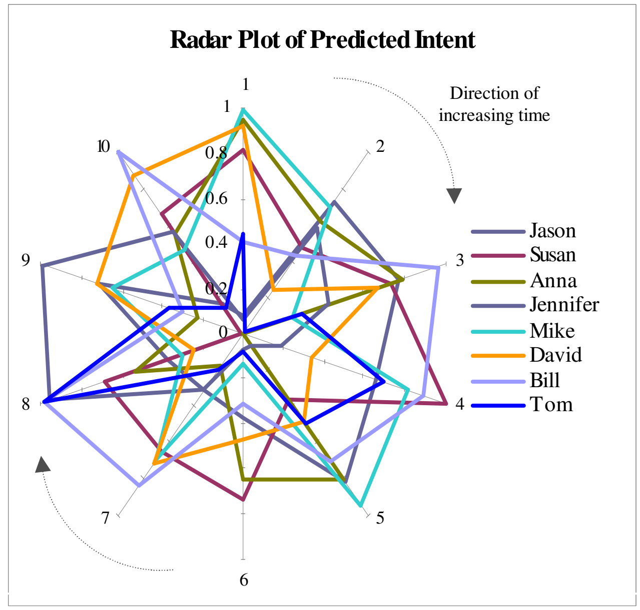

Figure 3: Dynamics of predicted intent across time and contacts for Charlie. The concentric axes are the ten weeks; each axis contains the normalized intent values between 0 and 1. Time is increasing in the clockwise direction.

Analysis of Predicted Intent: In this section we present some qualitative results of the predicted intent to communicate for a person Charlie and his social network comprising eight contacts. The experiments are shown over the first ten weeks of testing in Figure 3.

The figure shows normalized intent to communicate of a user Charlie with each one of his contacts across 10 weeks, averaged over the four different topics A-D. We observe an interesting pattern in the values of the intent. The intent is noticed to be consistent over time for certain pairs, e.g. Charlie-Jennifer and Charlie-David. On cross-checking with ground truth communication, we observe that Charlie, Jennifer and David show consistency of information roles over time. The measure of intent for these two pairs is also characterized by their 'habits' with respect to communication, e.g. responding with high intent on messages involving movie reviews. However, for certain other pairs, Charlie-Susan and Charlie-Anna the intent values are observed to be volatile across time. Actual communication reveals that (a) Susan and Anna's information roles depict temporal variability, and (b) bulk of the communication is about Topic D ('party, gathering') which showed temporal volatility in its occurrence pattern.

Analysis of Predicted Delay: In this section, we analyze predicted delays across the contacts of Charlie, topics and time. Figure 4 shows the variation of estimated delay values per contact (Jason, David and Susan) for a period of ten weeks. There are three cells for the three contacts, each comprising four columns representing topics A to D and ten rows on Y axis representing time. There are

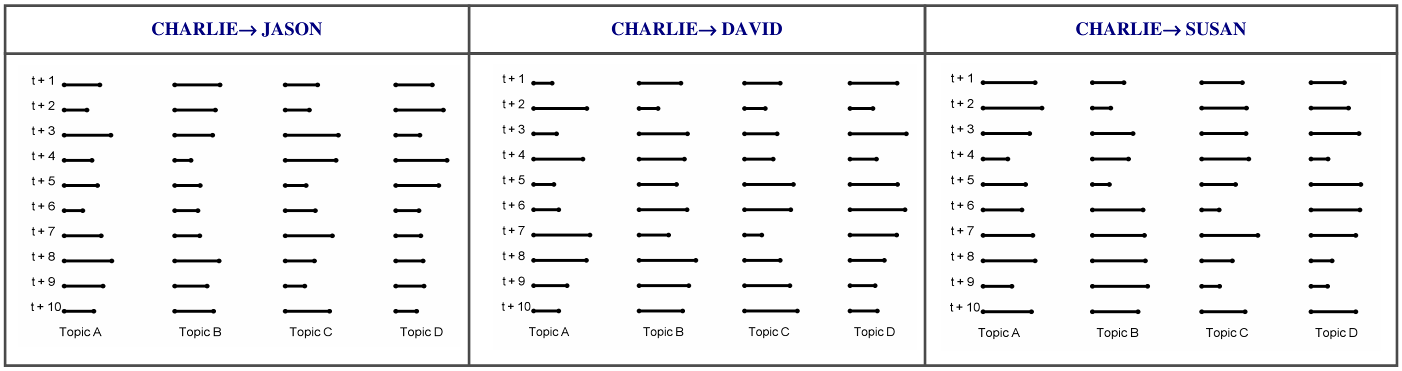

Figure 4: Dynamics in delay between Charlie and three of her contacts. The visualizations show dynamics for ten consecutive weeks (Y axis) and for four topics (X axis). The length of each horizontal line is proportional to the measure of delay.

several interesting insights revealed in this visualization:

1. We notice that for the contact Jason, the delays show high volatility across different topics. Delays in the first two weeks are low; while in the third and fourth weeks it increases. We therefore observe regularity in delay variation. From actual communication, we attribute it to occurrence of an external event (e.g. Friday evening parties) for the pair of users. Such an occurrence reduced the delay for a certain amount of time. When the event's effect got old, the delays got longer.

2. For Charlie and the contact David, we observe that delays whether long or short, persist for a certain amount of time consistently across the four different topics. Actual communication suggests that Charlie and David depict consistency in their communication behavior with respect to a particular topic. Being a generalized topic about 'person, someone', Topics A shows low delays with consistency while B shows longer delays with consistency (except for weeks 2 and 7 in both cases). The inconsistency for weeks 2 and 7 seem to have occurred due to personal contextual changes for the two communicators. Topics C and D refer to events and are therefore characterized by 'bursty' conversations where we have low delays for considerable short periods, followed by consistent high values.

3. We observe an interesting pattern for the contact Susan. For topics A, B and D we notice that the delay values are temporally consistent. This reflects consistency of Charlie and Susan in response behavior for these topics. However for topic C, the delays are volatile. Topic C, being about 'event, happening, occurrence', ground truth communication reveals that this pattern is due to the occurrence of external events e.g. 'Superbowl', 'Black Friday deals' etc.

From the results, we observe that intent to communicate shows less fluctuations across time and contacts when compared to delay. This might reveal that factors such as external events and personal contextual changes affect delay more than intent. While context can reasonably predict intent, delay seems to be more dependent on the personal habits and other latent factors, like 'age' of information and presence of cliques among people. This also indicates that the intent and delay are orthogonal parameters in characterizing communication flow. The impact of information roles and strength of ties are also emphasized in the results.

6. CONCLUSIONS

In this paper we developed a temporal dynamic representation framework for context that can help predict communication flow in social networks efficiently between a given pair of individuals. We described communication flow using two parameters based on our prior work: *intent to communicate* and *delay*. We designed features to model communication and social context and deployed them for predicting the intent to communicate and delay in transmission in a Support Vector Regression framework. For capturing temporal dynamics, a set of optimal features were selected at a given time instant using feature selection techniques. We observed qualitatively that intent is more affected due to contextual dynamics than delay. Delay seemed to be more dependent on other latent factors characterizing communication, e.g. 'age' of information, cliquey-ness among communicators etc.

There are several interesting directions for future work. There are situations where both the intent and delay are found be affected by causes beyond the context of the communicators. There might be latent factors involved which cannot be accounted for by our model of context: e.g, moods, sentiments, changes in social relationships, location, work habits etc. To handle these dynamics, we would like to explore modeling context as a Markov decision process and predict the intent and the delay more effectively.

7. REFERENCES

[1] C. BIRD, A. GOURLEY, P. DEVANBU, et al. (2006). *Mining email social networks*, Proceedings of the 2006 international workshop on Mining software repositories, Shanghai, China, 137-143.

[2] M. D. CHOUDHURY, H. SUNDARAM, A. JOHN, et al. (2007). *Contextual Prediction of Communication Flow in Social Networks*. Proceedings of the IEEE / ACM / WIC International Conference on Web Intelligence (WI '07). San Jose, CA.

[3] S. A. DELRE, W. JAGER and M. A. JANSSEN (2007). *Diffusion dynamics in small-world networks with heterogeneous consumers* Comput. Math. Organ. Theory **13** (2): 185-202

[4] P. DOURISH (2004). *What we talk about when we talk about context.* Personal and Ubiquitous Computing **8**(1): 19-30.

[5] M. GRANOVETTER (1973). *The strength of weak ties.* American Journal of Sociology **78** (6): 1360-1380.

[6] D. GRUHL, R. GUHA, D. LIBEN-NOWELL, et al. (2004). *Information Diffusion through Blogspace*, Proceedings of the 13th international conference on World Wide Web.

[7] C. HAYTHORNTHWAITE (1996). *Social Network Analysis: An Approach and Technique for the Study of Information Exchange.* Library and Information Science Research **18** 323-342.

[8] A. MANI and H. SUNDARAM (2007). *Modeling user context with applications to media retrieval.* Multimedia Systems **12**(4): 339-353.

[9] C. MOORE and M. E. J. NEWMAN (2000). *Epidemics and percolation in small-world networks.* Physical Review E **61**: 5678.

[10] X. SONG, B. L. TSENG, C.-Y. LIN, et al. (2006). *Personalized recommendation driven by information flow*, Proc. 29th ACM SIGIR Conference, Seattle, Washington, USA, ACM Press, 509-516.