A Social Network Based Approach to Personalized Recommendation of Participatory Media Content

Aaditeshwar Seth and **Jie Zhang**
School of Computer Science
University of Waterloo, ON, Canada

Abstract

Given the rapid growth of participatory media content such as blogs, there is a need to design personalized recommender systems to recommend only useful content to users. We believe that in addition to producing useful recommendations, certain insights from media research such as simplification and opinion diversity in recommendations should form the foundations of such recommender systems, so that the behavior of the systems can be understood more closely, and modified if necessary. We propose and evaluate such a system based on a Bayesian user-model. We use the underlying social network of blog authors and readers to model the preference features for individual users. The initial results of our proposed solution are encouraging, and set the agenda for future research.

Introduction

Consider an online personalized news service such as Google News. It helps users to manage the glut of information related to current events, by recommending to a user news stories that will be of interest to the user (Das *et al.* 2007). In the Google News system, a *news story* is considered as a category denoting a particular event, and is comprised of multiple *news articles* about the event. The system has been shown to perform well for story recommendation to identify the categories for events of interest to a user, but the methods used to rank articles (or *messages*) within stories are unclear. Although, considerable research in the area of *message effects* has identified factors that should be considered in the selection and ordering of news articles [1], it is unclear whether systems such as Google News take these factors into account. For example, the *simplification* of news to make it more easily understandable to readers, and the *opinion diversity* expressed in the news, have been shown to help people gain clarity and unbiased viewpoints about the event (Bryant & Zillman 2002; Jackson 1992). It is not evident whether the message ranking systems used by Google News and other services are able to ensure simplification and diversity in their recommendations. In fact, past studies on the search services of Google have shown that its algorithms tend to bias results towards more popular websites and reduce diversity (Hindeman, Tsioutsiouliklis, & Johnson 2003). The result was challenged subsequently (Fortunato *et al.* 2006).

Given the uncertainty of such observations, we consider it worthwhile that insights from media research should be used to form the theoretical foundations of news related recommender systems, so that the behavior of the recommender systems can be well understood, and modified if necessary. We therefore consider a recommender system to be good if it can not only produce useful recommendations, but also explain characteristics of the recommendations produced by it in terms of factors observed and studied by media theorists. We attempt to do so in this paper, to answer questions such as: given a set of messages about a current event that have already been read by a user, what minimum set of additional messages should be recommended to the user, so that she receives simple yet diverse information about the event? Note that we do not aim to recommend stories to a particular user; we assume that interesting stories for the user can be identified through systems such as Google News – we only aim to recommend messages about the story that the user will find to be most useful in terms of how much simplification and opinion diversity the messages will provide to the user.

We specifically focus on recommender systems for participatory messages such as blog-entries, that are marked by a high degree of user participation in writing and commenting. We feel that if we can solve this problem, the same insights can be applied to the similar problem for news articles as well. Other characteristics of our approach are as follows:

1. We model the features that provide simplification and diversity in a novel manner, based on the underlying social network of the message authors, participants, and readers (Seth 2007a). Social network based approaches to recommendation is an active area of research (Song *et al.* 2006; Yang *et al.* 2007; Yu & Singh 2003). Information about the social network of users was not available until recently with the trend towards social-networking websites and the extraction of social networks from email graphs.

2. We use the social-network based features that provide simplification and diversity to develop a personalized user-model represented as a Bayesian network, which indicates the preferences of its user towards these features. The model

[1] A discussion of the role of news media in society is beyond the scope of this paper. We refer the interested reader to (WorldBank 2002) as a starting point.

parameters are learned using prior history of message ratings given by the user. Our decision to use Bayesian networks is motivated because of their support for causality (Angermann, Robertson, & Strang 2005), which help us directly model personalized features of users in terms of factors considered in media studies. In addition, the Bayesian model can be implemented as part of a client-side application for each user, making it more privacy friendly than other centralized approaches which require a common database to keep track of the preferences of different users.

3. We have designed the model to recommend messages for a news story in an incremental manner, taking into account messages about the same news story that have already been read by the user in the past. This design has the advantage that as and when new messages about an event are published, the recommender system can decide whether or not to push the messages to a user, such that the messages provide diversity or simplification of information about the event for the user. This can be easily generalized to serve as a ranking mechanism for messages as well, although we have not examined it in this paper.

We next describe the rationale behind our approach, details of the user-model, and an evaluation using measurements and surveys done on a social-networking website.

Design rationale

In the previous section, we suggested that news recommender systems should be based on theoretical foundations that have been observed in media research, so that the behavior of the recommender systems can be understood more closely. To do this, we first propose two features of *context* and *completeness* of messages, that are directly related to providing simplification and opinion diversity respectively. We explain why these features can be estimated using graph theoretic properties of the underlying social network of message authors and readers. We then use these features to design a personalized Bayesian user-model that learns the preferences of its user towards contextual and complete information. The learned user-model predicts the usefulness of a new message for the user in terms of the amount of contextual and complete information this message will supply. Depending upon this prediction, a client-side user-agent can decide whether or not to recommend the message to the user.

Context and completeness

Information scientists have explored the notion of useful information in a message, and characterized it through features such as comprehensibility of the message, its scope, freshness, accuracy, credibility, and topic (Maglaughlin & Sonnenwald 2002; Rieh 2002). Similarly, media researchers have explored the effects of information, and use terms such as resonance, simplification, repetition, and opinion diversity to describe effects a message may produce (Bryant & Zillman 2002; Jackson 1992). Different message recipients are likely to attach different priorities to each of these features. We draw from these insights and define the following:

Context of a message relates to its *comprehensibility* (Maglaughlin & Sonnenwald 2002) or the *simplification* it provides (Bryant & Zillman 2002), based on how well the message content explains the relationship of the message to its recipient. Thus, *comprehensibility* and *simplification* can be considered as outcomes of the amount of context in the message; messages that are more contextual will be more comprehensible and improve understanding for recipients.

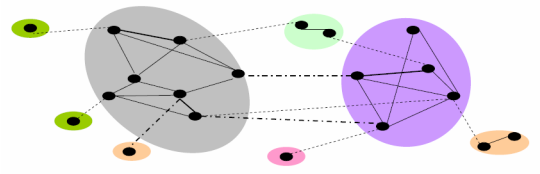


Figure 1: Strong and Weak Links

Completeness of a message denotes the depth and breadth of topics covered by the message. A concrete definition of *depth* and *breadth* is proposed by (Zhu & Gauch 2000), as the depth and breadth of the topic ontology graph covered by the message. The *scope* of the message (Maglaughlin & Sonnenwald 2002), or the *opinion diversity* provided by the message (Bryant & Zillman 2002), can be considered as outcomes of the amount of *completeness* in the message.

Note that context and completeness of messages are always observed from the perspective of a recipient user (or *ego*), and are hence features of messages personalized for the recipient user. Unless mentioned otherwise, context and completeness of messages will henceforth always be assumed to be stated with reference to some recipient user. Clearly, context and completeness of messages cannot be derived in a straightforward manner through semantic content analysis alone. Although completeness may appear to be similar to sentiment analysis in text (Kale *et al.* 2007), we have considered a broader working definition for completeness in this paper. We however do plan to accommodate sentiment analysis in our framework in the future. In related work (Seth 2007a), we took a simpler approach of using insights from the *strength-of-weak-ties* hypothesis (Granovetter 1973) to develop measures for context and completeness based on the social network of message authors and readers. We take the same approach in this paper, as explained next.

Role of social ties

The *strength-of-weak-ties* hypothesis states that social networks consist of clusters of people with *strong* ties among members of each cluster, and *weak* ties linking people across clusters, shown in Fig. 1. Whereas strong ties are typically constituted of close friends, weak ties are constituted of remote acquaintances or colleagues. The hypothesis claims that weak ties are useful for the diffusion of influence and economic mobility, because they connect diverse people with each other.

In the context of participatory messages, this hypothesis can be interpreted as follows. A participatory message written by a strong tie of a recipient, or having a large number of comments written by users strongly tied to the recipient,

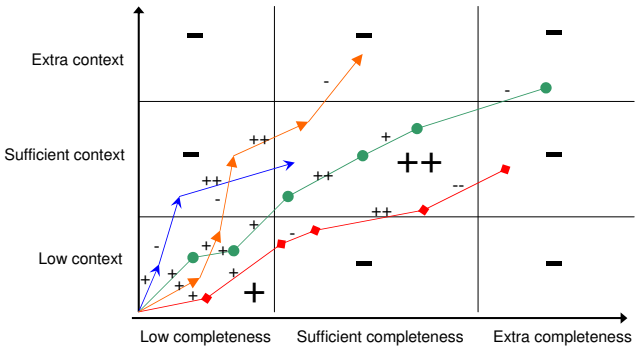


Figure 2: Knowledge state of a user

is likely to be more simple and understandable to the recipient. Similarly, a message having participation from users weakly tied to a recipient, is likely to carry more diverse information. The intuition behind this is simple. Close friends linked through strong ties will typically share the same environment and circumstances with each other, and hence message contributions made by them will be more *contextual* for other users sharing the same environment. Similarly, message contributions made by an acquaintance linked through weak ties to a recipient will bring in *completeness* of information for the recipient. Participatory messages can therefore be expected to gain context and completeness with time as different users write comments about them. Hence, the amount of contextual and complete information provided by a message at a particular time instant, will depend on the state of the message at that time. This hypothesis was verified by (Seth 2007a), and we use it to build the user-model.

User-model

The user-model is based on the assumption that users have some inherent preferences towards the marginal utility they will gain from new message recommendations, where the utility gain will depend upon the additional amount of context or completeness provided by the message. This is shown schematically in Fig. 2 to explain the intuition. The 2D space represents the *state of knowledge* of a user with respect to some news event, quantified by the amount of contextual and complete information about the event the user has received at any point in time. A particular event or news story is represented as a broken line, and each kink on the line represents a new message about the story read by the user. For each story, whenever a new message is read, it changes the state of knowledge of the user. This change will occur because the message will either provide more understanding about the story (context), or more diverse viewpoints (completeness); and the change might be rated positively or negatively by the user (+/- link annotations).

Our assumption is that the message ratings given by the user on this 2D space will be consistent across stories because the user will have some inherent rating criteria based on the trajectory she prefers to follow in gaining contextual and complete information. For example, the user may prefer to gain complete information only if it is accompanied by messages that help contextualize this new information for her. Or, the user may prefer to follow a different trajectory of reading only complete information. In addition, the user may prefer to read more and more information only to some extent, and the marginal utility she gains may become negative after a certain threshold is reached when a large amount of information has already been read by the user. Our goal therefore is to learn this preference function for each user over the 2D space of the current state of knowledge of the user. Once the function has been learned, it can be used to recommend messages to the user based on the prediction of how useful the user will find the messages to be.

The actual user-model we propose is more comprehensive. We take into account the topic of the news story, and learn a preference function for each topic. We also take into account features such as the freshness and credibility of messages. Freshness is required to model traits of a user's behavior such as whether the user prefers timely recommendation of messages about some topic, and whether the user gains higher utility from the first message about a story followed by lower utility from subsequent messages. Credibility helps differentiate between messages supplying reliable versus unreliable information. Details are explained later.

Related work

Most traditional recommender systems are either based on content mining approaches, or CF (collaborative-filtering) approaches, or a combination of both through various models (Adomavicius & Tuzhilin 2005). Our method is clearly different from such traditional methods because we develop a user-model based on features of simplification and diversity examined by research in news media. Furthermore, the traditional approaches do not use information about the underlying social network of users. In addition, we consider our model based method as a filtering mechanism applied to topic-specific recommendations produced by these approaches. Therefore, our method and the traditional approaches actually complement each other.

More closely related work includes (Song *et al.* 2006; Yang *et al.* 2007; Yu & Singh 2003). (Song *et al.* 2006) make recommendations based on stochastic simulations that replicate the observed patterns of information flow on social networks. (Yang *et al.* 2007) operate in a P2P setting, and use decentralized CF algorithms executed within local social network neighborhoods of users. (Yu & Singh 2003) learn content-based gradients on links between users; this can be used to route messages along desired gradients to users who will be interested in these messages. However, unlike our method which is based on the real-world social network of users, these methods consider an artificial network formed by linking users observed to be similar to each other. Furthermore, these methods do not model message features such as diversity and simplification.

Another related body of work focuses on decentralized routing algorithms on social networks (Kleinberg 2006; Adamic & Adar 2005). These methods do not learn link annotations based on observed message flow patterns, but assume that sufficient local information for decentralized rout-

ing is embedded in the link structure of the network itself. (Kleinberg 2006) show that social networks of people that emerge as a result of geographic proximity between people, are under certain conditions capable of routing messages to a given destination using only local information at the nodes. Similarly, (Adamic & Adar 2005) show that within a corporate organization, messages routed based on organizational hierarchy can find their way to a desired destination. Similar to our approach, these models also operate on the real-world social network of users. However, these methods have not been considered to build recommender systems.

Problem definition

Message: A news article, or a blog entry, along with all its comments, is considered as a single message.

Message contributions: The main component of the message (blog-entry or news article), and all the comments in the message, are individually referred as contributions to the message by different users.

Message participants: This includes all users who have made contributions to the message.

Message environment: The underlying social network connecting the message participants is referred to as the message environment.

Message collection: A message collection is a set of similar messages, ie. a news story which is a collection of related news articles.

Knowledge state of a user: This represents all messages in a collection that have been read by the user. We quantify the current knowledge state of a user as the contextual and complete information the user has read so far.

Message freshness: This represents the timeliness of the message with respect to the period of relevance of the event to which it refers.

Message usefulness: Message usefulness is the rating given to a message by a recipient, for example, on a 5-point scale (1..5). In the proposed user-model, we assume that the usefulness rating given by a user is based on how much additional context and completeness is provided by the message, conditional on the current knowledge state of the user.

Broad topic: Each message or message collection will belong to a broad topic to which it is relevant. For example, a broad topic could be *books*, which will include messages such as book reviews, or prize announcements, etc. Similarly, *climate change* could represent another broad topic. We are presently uncertain as to what criteria we should use to automatically infer a suitable granularity to classify topics as broad or narrow. We will explain later that we instead rely on the users to choose their own levels of granularity.

Problem definition: We can now define the problem precisely as being able to predict the usefulness of a new message not seen by the user so far, based on the current knowledge state of the user. We attempt to do this by learning the parameters of a Bayesian user-model for each broad topic that is of interest to a user. Since we directly model context and completeness, we are able to meet our goal of designing a recommender system that can explain the characteristics of the recommendations produced by it, based on factors studied by media theorists.

Knowledge requirements: We have made a number of assumptions about the recipient user-agent having information about messages, message environments, etc, that are required to learn the user-model. These are as follows.

- All message participants who have been involved in the message so far.
- The message environment, that is, the social network of the recipient user and the message participants.
- Messages from the same message collection (and the associated message participants and message environments) read by the recipient user in the past.
- Archived data for usefulness ratings of messages from the same broad topic seen by the user in the past.

All of these are reasonable assumptions to make. Knowledge about the social network of users is becoming available through APIs provided by social networking websites such as Facebook (*www.facebook.com*). Information about message participants and ratings given by users is also made publicly available on most blogging websites (eg. *www.livejournal.com*). If we assume that the user-agent will be implemented as part of a client-side application (Seth 2007b), then the application can even keep track of messages read (or clicked) by the user. Of course, an obvious problem is to manage the identities of users across multiple websites, and ensure that the identities are authentic. This issue will also likely be resolved through consortiums such as the OpenSocial initiative (*code.google.com/apis/opensocial/*), and other solutions for identity management. We do not consider various implementation specifics in this paper.

User-model

We represent the user-model as a Bayesian dependency graph shown in Fig. 3. Directed edges indicate a dependency from the originating variable to the target variable. Shaded ovals represent hidden variables and unshaded ovals represent evidence variables. The partially shaded oval for message usefulness is a variable denoting the rating given by the user, and is available as an evidence variables during the training phase only. The goal is to infer this variable for a new message, given the evidence variables and the parameters of the learned model.

The message usefulness is assumed to depend upon the two hidden variables for contextual and complete usefulness respectively, provided by the message. The hidden variables for contextual usefulness depend on evidence variables for the new amount of context provided by the message, the current state of knowledge of the user (quantified as the current amounts of contextual and complete information read by the user so far), and the freshness of the contextual component of the message. The dependency relationships for the completeness hidden variable are exactly similar. We next describe methods to measure the evidence variables.

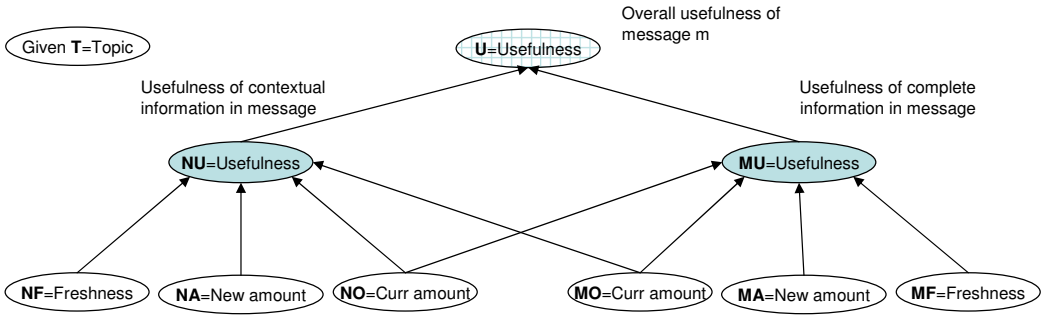


Figure 3: User model

Calculating evidence variables

We use the methodologies proposed by (Seth 2007a) to estimate the context and completeness provided by messages. These may not be the only ways to measure context and completeness based on graph-theoretic features of the message environment (Milo *et al.* 2002), but we found them to perform well in previous work.

1. We assume that the social network structure is known in advance as a directed graph $G(U, E)$, where users are represented as nodes U with edges E between users who are friends or know each other.

2. We assume that a list of broad topics T is known in advance, and a boolean value for each (user u_i, topic $t_k \in T$) pair is also known that indicates whether or not the user is interested in the broad topic. The induced subgraphs formed by users who are interested in the same broad topics are referred to as topic specific social networks (TSNs).

3. We assume that a clustering algorithm is known, that can be used to cluster topic specific social networks such that users within each cluster have strong links between them, and users in different clusters are connected with weak links. This will produce a network schematically similar to the one shown in Fig. 1. Each cluster denotes contextual boundaries such that users within each cluster share a common context with each other. Finding such a clustering algorithm for social networks is an active area of research (Tantipathananandh, Berger-Wolf, & Kempe 2007). In earlier work (Seth 2007a), we used an algorithm based on stochastic flow simulation (Dongen 2000), which gave us good results. We will use the same algorithm in this paper. Note that a shortcoming of this representation is that it restricts a user to be a member of only one cluster. In the future, we will extend the representation to allow users to be members of multiple clusters as well.

4. For each cluster of strong ties V, calculate its clustering coefficient C_V (Newman 2003). We will use the clustering coefficient as a proxy for the cohesivity of the cluster, to denote the degree of shared context among members of the cluster. We sometimes refer to C_{V_i} as the clustering coefficient of the cluster to which user u_i belongs. Note that these calculations are done separately within each TSN.

5. For each user u_i, calculate her integration coefficient γ_i into her cluster (Valente 1995). We will use the integration coefficient of a user as a proxy for amount of contextualization provided by messages written by the user.

$$\gamma_i = \frac{1}{(|V|-1)D_V} \sum_{u_j \in V} (D_V - d(i,j)) \quad (1)$$

Here, $d(i, j)$ is the distance from user u_i to u_j, calculated as the shortest path between the two users. D_V is the diameter of the cluster V = maximum distance between any two users $\in V$. Thus, the integration coefficient $\gamma_i \in [0, 1)$ of user u_i into her cluster V, is close to 1 if she is well integrated into her cluster, ie. they are close to many other users. Similarly, γ_i is close to 0 if she is present along the boundaries of the cluster and is not well integrated.

6. For each user, calculate a local credibility score within her cluster, and a global credibility score across the entire topic specific social network. In this paper, we assume local and global scores to be equal, and we estimate them using a naive heuristic as the number of contributions made by a user. We will explore credibility computation in future work, based on popular mechanisms for trust-propagation such as pagerank and HITS (Langville & Meyer 2004). The credibility score for user u_i is denoted as δ_i. Note that the score is topic specific: the same user can have different credibility scores for different topics.

7. We assume that for each event and its corresponding message collection M, the set of messages $m_j \in M$ seen by the user in the past is known. The message participants of the j^{th} message are denoted as $U(m_j)$, and their individual contributions are denoted as $R(m_j)$. For participants linked through strong or weak ties with a message recipient, the participants are denoted as $U_s(m_j)$ or $U_w(m_j)$ respectively. The same convention is followed for their respective contributions: $R_s(m_j)$ or $R_w(m_j)$.

8. NA = Evidence variable for the *new amount of context provided by a message*: Referring to Fig. 3, the amount of contextual information provided to user u_i by a new message m_n is considered to be composed of three components: a social network based estimate of the context provided, the length of the message, and the credibility of the message.

Each component is calculated separately for every contribution to the message.

8a. The social network based component of the context provided to user u_i by the j^{th} contribution of message m_n belonging to a broad topic t_k, is denoted as $SNContext_{ijnk}$. As before, we skip the subscript k for simplicity, and model the social network based component as follows:

$$SNContext_{ijn} = C_{V_i}.\gamma_j \tag{2}$$

Here, C_{V_i} is the clustering coefficient of the cluster of the message recipient u_i. γ_j is the integration coefficient of the j^{th} message participant (eqn. 1). Note that this is valid only for message participants in the same cluster as the recipient user (γ_j is 0 otherwise). Thus, more context will be provided by a contribution from a user closely integrated into the cluster of the recipient user.

8b. The component based on message length is used as a proxy for the amount of information conveyed by the message. We will extend this simple heuristic in the future by using language models and other information-retrieval techniques for length normalization (Singhal 2001). In this paper, we assume that a global constant L_k for topic t_k is known, such that the information content of the j^{th} contribution of message m_n is measured as:

$$info_{jn} = \text{min(length}(r_j), L_k) \tag{3}$$

Here, r_j is a contribution made by the j^{th} participant in message m_n, considered only for message participants who are in the same cluster as the recipient user. L_k denotes a maximum threshold length for contributions.

8c. The credibility of each message participant is considered as a proxy for the credibility of the contribution made by the participant. In the future, we plan to extend credibility computation of messages by taking the ratings given by other users into account. However, in the present scenario, the component for credibility can be simply expressed as:

$$Cred_{jn} = \delta_j \tag{4}$$

Here, δ_j denotes the credibility of the j^{th} participant for the topic to which the message belongs. As before, this is valid only for those users who are in the same cluster as the recipient user.

8d. We now need a method to combine the three components to calculate the contextual amount of information provided to user u_i by message m_n. However, currently we do not have any theoretical basis for combining these components except that they should be positively correlated with NA. We do believe that we need to identify a global definition for the function to calculate NA in a uniform way for all users, because NA estimates the contextual amount of information in any message, and it should therefore have a uniform information theoretic formulation. We represent this as follows:

$$(na)_{in} = \sum_{j \in U_s(m_n), R_s(m_n)} f(SNContext_{ijn}, info_{jn}, Cred_{jn})$$

In our evaluation, we experiment with different functions f to combine the three components in a product form, and choose the function that gives us the best performance. However, the actual function can be inferred statistically when data on a large number of users is available. Finally, the sum of the contextual information provided by each contribution is then considered as the overall contextual information provided by the message.

9. NO = Evidence variable for the *current amount of context already provided to the user*: Referring to Fig. 3, this is essentially the sum of the context provided by individual messages from the same message collection that have been seen by the user, and can be expressed as:

$$(no)_i = \sum_{m_n \in M} (na)_{in} \tag{6}$$

10. MO = Evidence variable for the *current amount of completeness already provided to the user*: Referring to Fig. 3, the completeness provided to user u_i by a set of messages M seen by the user, is also expressed as being composed of three components: a social network based estimate of the completeness provided, the message lengths, and the message credibilities.

10a. Before computing the social network based component of completeness, we introduce a metric called the second-degree integration of user u_i's cluster V_i into an adjacent cluster V_j. Let W denote a subset of the destination nodes of weak links from the cluster of user u_i into cluster V_j, where V_j is not the same as V_i. Calculate the second-degree integration as follows:

$$\gamma_{ij}(W) = \frac{1}{(|V_j| - 1)D_{V_j}} \sum_{u_k \in V_j} (D_{V_j} - d_j(W, k)) \tag{7}$$

Here, $d_j(W, k)$ is the minimum distance to user u_k in cluster V_j from any user $\in W$. Thus, γ_{ij} will be high if the subset of the destination nodes of weak links into the neighboring cluster are well distributed across the cluster, such that the minimum distances from the nodes $\in W$ to every other user in V_j is small. We will use the second-degree integration coefficient as a proxy for the degree of completeness provided by the adjacent cluster, to capture the intuition that a larger subset of weak ties into an adjacent cluster will provide more completeness. Note that $\gamma_{ii}(W)$ can be calculated in the same manner, except that W will now include members from the same cluster as the recipient user.

10b. Let $V(m_j)$ denote the set of clusters spanned by participants in a message m_j, and $V(M)$ denote the set of clusters spanned by all messages $m_j \in M$. The social network based component of completeness provided to user u_i by messages M belonging to a broad topic t_k, is denoted as $SNCompleteness_{ik}$. As before, we drop the subscript k:

$$SNCompleteness_i = \sum_{j \in V(M)} |V_j|.\gamma_{ij}(W_j) \tag{8}$$

Here, W_j includes those ties that lead from V_i to V_j, among the participants in M. The summation therefore denotes the sum of the completeness contributed by individual clusters to which participants of M belong. This can be intu-

itively understood as the "area" of the social network graph spanned by the messages.

10c. The completeness components based on lengths of individual contributions and the credibility of the components are calculated in the same manner as that for context (eqn. 3, 4), and will be included as weights in the summation for $SNCompleteness_i$ (eqn. 8) in function g described next.

$$info_j = \sum_{r \in W_j} info_r, Cred_j = \sum_{r \in W_j} Cred_r \quad (9)$$

10d. We can now express the overall completeness provided by the set of messages M seen by the user as $(mo)_i$:

$$(mo)_i = \sum_{j \in V(M)} g(SNCompleteness_{ij}, info_j, Cred_j) \quad (10)$$

Similar to the function f for $(na)_i$, an appropriate function g for $(mo)_i$ can now be identified, to combine the social network component with the length and credibility components. The difference is that for the calculation of completeness, these components are combined collectively for all messages $\in M$, rather than individually for each message as in the calculation for context.

11. MA = Evidence variable for the *new amount of completeness provided by a message*: Referring to Fig. 3 and following the same method as above, the additional amount of completeness provided by a new message m_n to user u_i can be expressed as follows.

11a. The social network component $SNCompleteness_{in}$ can be expressed as:

$$SNCompleteness_{in} = \sum_{j \in V(m_n) \setminus V(M)} |V_j|.\gamma_{ij}(W_j) + \sum_{j \in V(m_n) \bigcap V(M)} |V_j|.(\gamma_{ij}(W'_j) - \gamma_{ij}(W''_j)) \quad (11)$$

Here, W_j includes those users $\in V_j$ who are linked through weak ties from V_i to V_j, from among the participants in m_n. The first summation therefore denotes the completeness contributed by new clusters to which participants of m_n belong, that did not have any participation from users in the earlier messages M seen by the user. The second summation includes clusters that are common among message m_n and the earlier messages M seen by the user, but it only considers the additional amount of completeness provided by new participants in m_n who did not participate earlier in M. This is captured by considering the difference in γ calculated on W'_j and W''_j, where W''_j includes those users $\in V_j$ who are linked through weak ties from V_i to V_j, from among participants only in $U(M)$, and W'_j includes the corresponding set of users $\in U(m_n) \bigcup U(M)$.

11b. The overall amount of completeness provided by a new message can now be expressed by combining each Right-Hand-Side component in the summation of $SNCompleteness_{in}$ above, with $info_{jn}$ and $Cred_{jn}$ using the same function g.

12. NF = Evidence variable for *freshness of the contextual information provided by a message*: In this paper, we use a naive heuristic to calculate freshness. We assume that the event becomes relevant from the time instance of the first message contribution, and consider the time elapsed for subsequent contributions made to the message as an inverse measure of freshness. The contextual freshness is then calculated as the mean of the freshness of the contributions made by participants strongly linked to the recipient.

13. MF = Evidence variable for *freshness of the completeness provided by a message*: This is calculated in the same manner as the freshness of contextual information provided by a message, except that all contributions from strong and weak ties are considered in this case.

Learning and inference

During the learning phase, we assume that our knowledge requirements stated in Section "Problem definition" will be satisfied. Therefore, we will be able to calculate the evidence variables **NA, NO, NF, MA, MO, MF**, and know the user ratings for the usefulness variable **U**. This will allow us to learn the parameters for the user-model using standard algorithms such as EM (Russel & Norvig 2003).

During the inference phase, we will use the user-model to calculate $P(U)$. This can be calculated using standard MCMC or Join-Tree belief propagation algorithms for Bayesian networks (Russel & Norvig 2003). The value of $P(U = u)$ can be used to decide whether or not to recommend the message to the user.

Cold start: In general, learning and inference based on prior history always face a problem of cold-start for new users. The standard method to solve this is to use content-based models during the initial stages when sufficient data is not available (Melville, Mooney, & Nagarajan 2002). However, we do not explore the cold-start problem in this paper.

Evaluation

We next evaluate the performance of the user-model for different users in terms of the correct prediction of message usefulness ratings given by the users.

Dataset: We chose a popular social-networking website, Orkut, to evaluate our solution. Users in Orkut can subscribe to communities of interest and participate in discussions in these communities. We consider a *community* equivalent to the granularity of a *broad topic* as defined earlier, a *discussion* equivalent to a *message collection* within the broad topic, and a *posting* in a discussion equivalent to a *message*. For example, a community on *Politics* may have a discussion about *911*, with many postings in the discussion. Users in Orkut can also identify their real-life friends and link to them, which can give us information about the underlying *social network* as well. We collected this information in prior work (Seth 2007a), where we crawled multiple communities, discussions, and social networks on Orkut. We also did a survey of users to tune a clustering algorithm for identification of strong and weak ties among users. The only piece of information we were lacking to analyze our model, was message ratings by various users. We recruited volunteers from randomly selected users in 4 communities, and

asked them to rate 10 messages each in 4 message collections from the same topic. Ratings from 5 users were obtained in our current work. We hope to collect more data in the future.

Experiment: We used an open-source package, OpenBayes, to code our model. We simplified the model by discretizing the evidence variables of **NA, NO, NF, MA, MO, MF** into 3 states, the hidden variables of **NU, MU** into 2 states, and a binary classification for the usefulness variable **U** $\in$ {useful, not useful}. We assumed that users read the messages in order, so that we could estimate the variables in an incremental manner. For each user, we then studied the performance of our classifier with different choices of functions f and g. We will infer the functions statistically when we have more data; in current work, we worked with the following functions for context f, and similar functions for completeness g:

- $f = (SNContext)^{\{0.5,1,2\}}.(info)^{\{0.5,1,2\}}.(Cred)^{\{0.5,1,2\}}$: We studied different permutations of the exponents to examine the relative effects of the social-network component, the length, and the credibility. Within each experiment, the same permutation was used for all of **NA, NO, MA, MO**.
- $f = (SNContext).log_2(2+info).log_2(2+Cred)$: The logarithms were applied to reflect a subdued increase in the relative importance of different components.
- $f = \{SNContext, info, Cred\}$: Only a single component was considered in each experiment to study its impact on performance.

For each experiment, we ran *k-fold* cross validation tests to study the performance of the classifier. The model was learned using 80% randomly selected ratings with the EM algorithm. The rest of the 20% ratings were inferred using the MCMC and Join-Tree implementations in OpenBayes. Both the methods gave similar results; we only show the MCMC results here. Since we are interested in the binary classification $P(U = 0, 1)$, we use the standard ROC plot for the true-positive-rate (TPR) and false-positive-rate (FPR) to test the performance (Davis & Goadrich 2006). Our goal is to achieve high TPR with low FPR in the classification produced by the user-model.

Results: Fig. 4 shows the results for 5 kinds of functions: the product of logarithms of the three components, each component considered separately, and the polynomial function $f = (SNContext).(info).(Cred)^2$. The envelope across all polynomial functions is also shown. Each point is the (TPR, FPR) result for a single user; since we have 5 users, there are 5 points for each function. Results using the CF approach are also shown for comparison, but these are for only 4 users because the ratings of the 5^{th} user were not highly correlated (< 0.2 (Melville, Mooney, & Nagarajan 2002)) with ratings of any other user.

Although the evaluation is for only a few users, our initial results are encouraging. The polynomial and log-product functions consistently dominate functions that consider only a single component. This is evident from the (TPR, FPR) values for the polynomial and log functions

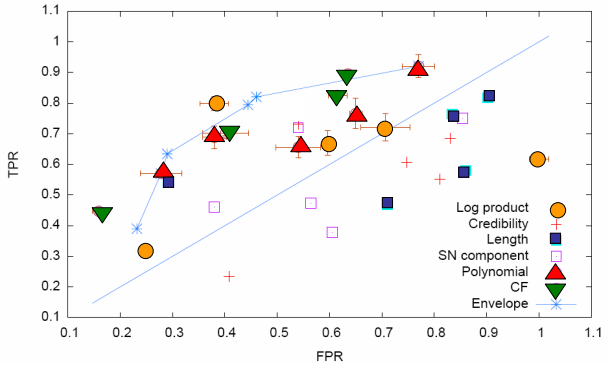
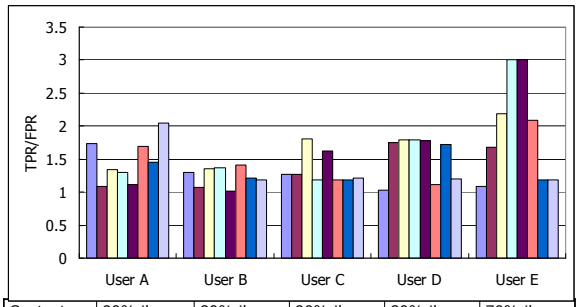


Figure 4: ROC plot for different functions f and g

that mostly remain above the random baseline for $y = x$ on the ROC plot. Although not shown here, we also note that the greater importance given to credibility in the polynomial function seems to be consistent across results with other polynomial functions. For example, $f = (SNContext)^{0.5}.(info)^{0.5}.(Cred)$ also dominates $f = (SNContext)^{0.5}.(info)^{0.5}.(Cred)^{0.5}$.

The performance of the polynomial functions in close to that of the CF approach. We consider this encouraging because: (a) Our approach can produce results even for users whose preferences are not correlated with those of other users. (b) There is much room for improvement in our results. We have used very naive heuristics for length and credibility, and un-weighted clustering and integration coefficients. We are hopeful that more sophisticated measurements will produce better results. (c) Our model can offer better explanations for behavior of the news recommender system in terms of factors considered in media studies.



	User A	User B	User C	User D	User E
Context	90% ile	80% ile	90% ile	80% ile	70% ile
Completeness	20% ile	40% ile	30% ile	70% ile	50% ile

Figure 5: Performance for different users

Fig. 5 shows the TPR:FPR ratios for the 5 users with 8 randomly selected polynomial functions. Ratios greater than 1 indicate better performance, but it is interesting to note that the performance is low for users B and C across all functions. It will be useful if we can predict in advance the users for whom our model may not perform well. To investigate this, we show in the table the percentile of the context and completeness values for the social network of

these users, calculated according to (Seth 2007a). In (Seth 2007a), it was shown that these values can predict the ability of a user to receive contextual and complete information respectively from her social network. Here, we see that the performance for users D and E is significantly higher, and these users also have high values for context and completeness of their social networks. Although we do not have sufficient data to test our hypothesis, but it seems that our model performs well for users whose social networks have high values of context and completeness. We will do more extensive analysis in future work.

Conclusions

In this paper, we proposed and evaluated an approach to personalized recommendation of participatory media content using social networks and a Bayesian user-model. Our model directly takes into account the preferences of users towards simplification and opinion diversity in recommendations. This can help understand the behavior of the recommender system in terms of factors studied by media theorists, and modify the behavior if necessary. Our initial results on the quality of recommendations seem promising.

Acknowledgements

We wish to express our sincerest thanks to Prof. S. Keshav and Prof. R. Cohen for their comments on early drafts of the paper, and to the anonymous reviewers who gave us invaluable suggestions to improve the paper.

References

Adamic, L., and Adar, E. 2005. How to search a social network. *Social Networks* 27(3).

Adomavicius, G., and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowledge and Data Engineering* 17(6).

Angermann, M.; Robertson, P.; and Strang, T. 2005. Issues and requirements for bayesian approaches in context aware systems. *LNCS*.

Bryant, J., and Zillman, D. 2002. *Media Effects: Advances in Theory and Research*. Lawrence Erlbaum Associates.

Das, A.; Datar, M.; Garg, A.; and Rajaram, S. 2007. Google news personalization: Scalable online collaborative filtering. *WWW*.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and roc curves. *ICML*.

Dongen, S. 2000. Mcl: A cluster algorithm for graphs. Technical Report PhD Thesis.

Fortunato, S.; Flammini, A.; Menczer, F.; and Vespignani, A. 2006. Topical interests and the mitigation of search engine bias. *PNAS* 103(34).

Granovetter, M. 1973. The strength of weak ties. *American Journal of Sociology* 78(6).

Hindeman, M.; Tsioutsiouliklis, K.; and Johnson, J. 2003. Googlearchy: How a few heavily linked sites dominate politics on the web. *Midwest Political Science Association*.

Jackson, S. 1992. *Message Effects Research: Principles of Design and Analysis*. Guilford Press.

Kale, A.; Karandikar, A.; Kolari, P.; Java, A.; Joshi, A.; and Finin, T. 2007. Modeling trust and influence in the blogosphere using link polarity. *ICWSM*.

Kleinberg, J. 2006. Complex networks and decentralized search algorithms. *ICM*.

Langville, A., and Meyer, C. 2004. A survey of eigenvector methods for web information retrieval. *SIAM Review* 47(1).

Maglaughlin, K., and Sonnenwald, D. 2002. User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgements. *Information Science and Technology* 53(5).

Melville, P.; Mooney, R.; and Nagarajan, R. 2002. Content-boosted collaborative filtering for improved recommendations. *AAAI*.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: Simple building blocks of complex networks. *Science* 298.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Review* 45(2).

Rieh, S. 2002. Judgement of information quality and cognitive authority on the web. *Information Science and Technology* 53(2).

Russel, S., and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Pearson Education.

Seth, A. 2007a. An infrastructure for participatory media. Technical Report CS-2007-47.

Seth, A. 2007b. An infrastructure for participatory media. *AAAI-WS Recommender Systems*.

Singhal, A. 2001. Modern information retrieval: A brief overview. *IEEE Data Engg. Bulletin* 24(4).

Song, X.; Tseng, B.; Lin, C.; and Sun, M. 2006. Personalized recommendation driven by information flow. *SIGIR*.

Tantipathananandh, C.; Berger-Wolf, T.; and Kempe, D. 2007. A framework for community identification in dynamic social networks. *SIGKDD*.

Valente, T. 1995. *Network Models of the Diffusion of Innovations*. Hampton Press.

WorldBank, T. 2002. *The Right to Tell: The Role of Mass Media in Economic Development*. The World Bank.

Yang, J.; Wang, J.; Clements, M.; Pouwelse, J.; de Vries, A. P.; and Reinders, M. 2007. An epidemic-based p2p recommender system. *SIGIR-WS Large Scale Distributed Systems*.

Yu, B., and Singh, M. 2003. Searching social networks. *AAMAS*.

Zhu, X., and Gauch, S. 2000. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. *SIGIR*.