ABSTRACT

Personalized recommender systems are required to help people cope with the information glut of participatory media content such as blogs. We focus on news related content and propose the design of a social network based recommender system for this purpose. We show that the use of sociological theory helps to simplify the system design, improve scalability, and understand the behavior of the system. In this paper, we present an overview of our work in sociological theory and user modeling, outline the system design for a recommender system that makes use of this work, describe some open problems, and focus on one component of the system that is strongly grounded in social network theory.

1. INTRODUCTION

Given the rapid growth of participatory media content such as blogs, videos, and podcasts, there is a need to build personalized recommender systems to recommend only useful content to users [1]. However, the design of such systems presents many challenges.

1. Scalability: Systems should be able to scale to millions of messages published daily, and offer recommendations to hundreds of millions of users.

2. Personalization: Due to the varied interests of users, personalized recommendations should be provided to them.

3. Credibility: Mechanisms should be designed to validate the credibility of participatory media content to ensure that credible messages are recommended to users.

4. Grounded in media theory: Much of the participatory media content is related to political opinions that tend to be subjective. Considerable sociological research in the area of media and message effects has identified factors that should be considered in the presentation of such content to help people gain clarity and unbiased viewpoints [4, 5]. We hypothesize that participatory media improves both simplification and diversity of information [10]. Consider the following example.

BBC News published an article on November 4th 2007 about the 2007 Emergency declared in Pakistan. The article described some aspects of the event, such as President Musharraf’s justification of his decision, reaction by the judiciary, and condemnation by other political leaders. Following are two comments by readers:

1. “I recently graduated in electrical engineering from Comsats Islamabad and got a job after a long struggle in one of the telecom companies here in Islamabad. I am hired on the basis that they are starting a new project in NWFP and FATA areas. After this emergency declaration company is now thinking to cancel the project in that area for which I was hired for, as NWFP and FATA areas are prime hiding places for Taliban... Now my job is in jeopardy and don’t know what my future holds for me...”

2. “I have family in Karachi and we are leading normal lives going about our daily work, parties, schools and all, a few changes like more uniformed men and barriers not a big problem, in fact most of us are glad that Musharraf took this action, he should have done this earlier... If any Pakistani leader is to be trusted with leadership it is Musharraf, not traitors and looters...”

Clearly, both these comments explored aspects of the event that had not been considered in the original article, and thus

1 http://news.bbc.co.uk/2/hi/south_asia/7077310.stm
add extra information to the article. What is more interesting to notice is that each comment may have been useful to different people for different reasons. The first comment may have been particularly useful for fellow-graduates of the author because it could have spurred some corrective actions on their part, or explained the relevance of the article for them. In other words, the comment would have simplified the article for them. The second comment may have also been useful, but for a different reason of having increased the diversity of the article. Furthermore, it is reasonable to expect that different people may have different preferences of seeking simplified information and diverse information, and this can influence their perceived usefulness of an article and its comments. We next explain how social network information can be used to model these user preferences.

2.1 Context and completeness

Rather than use the words simplification and diversity, we resort to a broader working terminology using the words context and completeness respectively. Contextualization of a message for a recipient helps to “situate” the message better in the recipient’s context, i.e. explains the relationship of the message to the recipient, and hence leads to simplification. Completeness of a message denotes the depth and breadth of topics covered by it [8], i.e. the scope or diversity of viewpoints it contains. Restating the example in this terminology, the first comment in the BBC article would have provided context to fellow-graduates of the author of the comment, and both the comments would have provided additional completeness to the readers by exploring new aspects of the topic. Although these features of context and completeness may appear to have arisen randomly in the comments, sociological theory helps explain away this randomness. Our insight is that instead of considering the comments as independent entities, the social network relationships between the comment authors and recipients should be taken into account, as explained next.

The strength-of-weak-ties hypothesis in social networks [6] states that social networks consist of clusters of people with “strong” ties among members of each cluster, and “weak” ties linking people across clusters. Whereas strong ties are typically constituted of close friends, weak ties are constituted of acquaintances or remote colleagues. The hypothesis claims that weak ties are useful for the diffusion of information, influence, and economic mobility, because weak ties help connect diverse clusters of people with each other. Also implicit to the hypothesis is the phenomenon of homophily noticed in multiple studies, which states that people similar to one another in terms of age, ethnicity, geographical location, income status, etc., tend to cluster together [7, 9, 12]. Homophily leads to strong ties between people similar to each another, and weak ties linking diverse clusters together.

The strength-of-weak-ties hypothesis can help explain the BBC news example. Although we do not have knowledge of the underlying social network of the various comment authors, the phenomenon of homophily makes it likely that the two comment authors would belong to different social network clusters, the first author being strongly connected to recent graduates from the same university, and the second author being in a higher bourgeoisie class cluster, possibly with weak links across the two clusters. The comments will therefore be more contextual for other people in the same clusters, and provide more completeness across the clusters.

Thus, if the social relationships between the comment authors and recipients are taken into account, it can explain the reasons why people may find an article to be useful.

This can be converted into a qualitative model, as shown in Fig. 1. Part (a) shows that people in different parts of the social network may read different messages about the same event, referred to as spatial dissemination of the messages. Now, when people comment on these messages locally within their clusters, it leads to contextualization of the messages for other members of the same clusters. This is because close friends linked through strong ties would typically share the same environment and circumstances with each other, and hence message contributions made by them will be more contextual for other people sharing the same environment. Part (b) shows that when people read messages and comments written by people in adjacent clusters linked through weak ties, it provides more completeness to them. This can also happen when people cross-post across different clusters, and provide “new” viewpoints or address different aspects of the event described by the message. Part (c) shows that these new viewpoints also circulate within external clusters to gain context. This process is termed as the temporal evolution of a message, to indicate that the context and completeness of the message increases with time as it circulates among different participants. In practice, all the above processes are likely to occur concurrently and there may not be any time-sequencing between the stages. Thus, a message can be visualized as spreading on the social network; depending upon the pathways that it takes, it gains context and completeness for a recipient.

Note that the model assumes that clusters of strong ties of people define contextual boundaries between them, where people within the same cluster share a common context with each other. However, people have different interests, and clusters of common context should therefore be topic specific. Hence, we introduce the notion of a topic specific social network, which consists of only those people who are interested in a particular topic. The model is specified precisely in [10] as operating within and across clusters in topic specific social networks only, rather than the generic social network over all topics.

Such insights from sociological theory allow us to model participatory messages in mathematical terms by making use of the social network information of message authors and recipients. In [10], we developed methods to classify social network ties as strong or weak. We then formulated graph-theoretic measures to determine the degree of context or completeness in a message for a recipient, based on the contributions it gained from strongly or weakly connected portions of the social network of a recipient. These formulations were validated by correlating measurements on a web-crawl of a social networking website, orkut.com, with surveys of real users subscribed to the website.

Usefulness model: As stated earlier, different users will have different preferences for contextual or complete information they prefer to receive. To give a visual representation, Fig. 2 illustrates the mass-distribution of the number of strong and weak links of users over the population of users interested in a certain topic, obtained from our Orkut dataset. Clearly, there is a large variation in the characteristics of users for associating with other users who are strongly or weakly connected to them. Some users are “self-referential”, having a large number of strong ties but few
weak ties. Some other users are “hub users”, having few strong ties but many weak ties. We noticed that these behavioral characteristics of users also translate into how much contextual or complete information the users prefer to receive from their social network. Measures for context and completeness were used to build a Bayesian network model that learns user preferences for the usefulness they derive from contextual and complete information [11]. The Orkut dataset was used to validate the usefulness model, showing that social network information can be used to predict the usefulness that people may associate with each message.

2.2 Credibility

The phenomenon of homophily also enables us to develop a model for credibility. Evaluating the credibility of a blogger or an individual message is a complicated problem. Bloggers tend to be opinionated and different people judge the credibilities of bloggers differently according to their own contexts. Furthermore, the credibilities are likely to be topic specific; an expert in some area may not be an expert in another area. As a norm, we can expect that whenever users annotate a message positively or negatively as in websites like digg.com, they will always do so based on their inherent tastes and biases. However, consideration of the underlying social network of the users can help make sense of seemingly random ratings: strongly connected users are likely to share the same viewpoints and hence rate messages similarly, but weakly connected users may cause the diffusion of diverse opinions and introduce heterogeneity in the message ratings, especially for political messages [13].

Credibility model: We use this insight to hypothesize that users will have different propensities to agree with other members of their cluster, or with the general public opinion, or be completely self-reliant in their own beliefs [19]. This was used to develop an Eigenvector based method to differentiate between different kinds of credibilities: contextual credibility based on ratings given by users in the same topic specific cluster; experienced credibility based only on ratings given by a specific user; and public credibility based on all available ratings. These Eigenvectors were used to learn the characteristics of a user towards judging the credibilities of contextual and complete information, again by using a Bayesian network. The credibility model was validated on a dataset obtained from a web-crawl of digg.com.

3. RECOMMENDATION PROCESS

The sociological models can be used to develop a process for generating message recommendations. We provide a detailed description later in Section 4.

1. We assume that the social network structure is given 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 topics $T$ is given in advance, and a boolean value for each (user $u \in U$, topic $t \in T$) pair indicates whether or not the user is interested in the topic. The induced subgraph $G_t$ formed by users who are interested in some topic $t$ is called the topic specific social network for that topic.

3. We assume that a clustering algorithm exists to find clusters in 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.

4. Now, when a new event occurs related to topic $t$, many messages $M$ about the event will be created. For a message $m \in M$, the active environment of the message is defined as the set of participant users who write comments or replies to the message. $U_m$ refers to the participant users, and $C_m$ refers to their contributions. $R_m$ are the ratings given to $m$ by other users.

5. For a recipient user $u$ who has already read messages $M' \subset M$, the additional amount of context and completeness a new message $m$ will provide to $u$ can be computed as functions $Context(G_t, u, U_m, C_m, M', t)$ and $Completeness(G_t, u, U_m, C_m, M', t)$, using the Bayesian usefulness model learned for $u$ [10,11]. The credibility of $m$ can
also be evaluated based on ratings by other users through a function \( Credibility(G_t, u, U_m, R_m, t) \) using the Bayesian credibility model learned for \( u \) [19]. Together, these models eventually produce a probabilistic answer of whether or not to recommend \( m \) to the user. This can be used to rank-order or drop new messages to be pushed to the user.

This recommendation process has many useful points, chiefly attributed to its foundations in sociological theory. The theory helps us develop models to personalize message recommendation for users, because it allows us to identify features behind the root causes for user preferences. The model driven approach also allows us to separate the learning process individually for each user, and hence learn the user preferences in a privacy-friendly manner without relying upon a centralized agency to aggregate personal user data. This gives us a number of advantages over other styles of producing recommendations, such as collaborative filtering (CF) [21]. First, CF requires all data to be reported to a centralized agency that tries to compute correlations between different pairs of users. This is not privacy-friendly and presents scalability challenges. Second, CF only observes correlations between users without trying to understand the reasons behind the correlations. This does not capture contextual information which could, for example, explain why a user may seem similar to another user at certain times and not at others. Some CF systems do try to get around these problems: [23] proposes a decentralized system, and [22] takes higher order correlations between users into account. However, our process can produce just as good recommendations in more desirable ways [11], ensuring scalability, credibility, privacy-friendliness, and a strong grounding in media theory which helps us explain the behavior of the system in sociological terms. We next describe the architecture to implement this recommendation process.

4. SYSTEM ARCHITECTURE

Primarily inspired by Cobra, the blog and RSS-feed aggregation system proposed in [14], Fig. 3 shows the proposed architecture of the system. We next describe the various components and their associated open problems. Later, we focus on one specific component and present a few results to illustrate the plausibility of our approach.

1. Data sources: These are news or blogging websites that host content. They typically provide RSS feeds, or can be customized to “ping” aggregation services, such as our system, whenever new content is published.

2. I/O servers: These are responsible to download new content from the data sources. Since the I/O servers will be bandwidth bound, they should be geographically distributed to download content from sources close to them.

3. Topic categorizers: The downloaded content needs to be analyzed to categorize it into various topics. Document classification schemes such as Latent Semantic Indexing, or fast matching of tags and keywords with pre-defined topics may prove useful here [15]. The topic categorization servers are likely to be CPU bound; they can be collocated with the I/O servers to minimize any further download of content. Metadata about new messages, including information about the message author and topic, is then pushed to the data center described next.

4. Distributed data structure: Inspired by the centralized approach followed by current Internet systems, we expect most data to be stored in large clusters of servers in data centers. This includes the topic specific social network graphs \( G_t \), messages \( M \) in circulation in the system, their current active environments \( U_m \), with metadata about various components \( C_m \), and the credibility ratings \( R_m \) given to the messages by various users. Other than storage, the second task is that of computation. We are able to split the calculation for the Context\((..)\), Completeness\((..)\), and Credibility\((..)\) functions mentioned earlier into generic components that are computable in the data centers, and user-specific components that are computable in the client application described later. The generic components computed in the data center are appended to the metadata for each message.

The design of efficient data structures to store large social network graphs, such that clustering and computation operations on these graphs can be executed efficiently, is an open problem. Relevant techniques include the design of structure indices [16] using virtual coordinates and landmark methods, that can be used to optimize the computation of metrics such as shortest paths, edge betweenness, and centralities in large social network graphs.

5. Reflectors: Whenever a new message enters the system, the message and its associated metadata has to be pushed out to various client applications. This can cause significant network overhead in message exchanges if a central message registry is maintained. We therefore design a network overlay to provide scalability in message updates as follows.

A reflector instance exists for each cluster in a topic specific social network, and hosts an RSS feed for messages relevant to users of that cluster. These feeds carry the message metadata and are read by the client application. Supernodes are elected to act as reflectors from among the client applications for users belonging to different clusters. Whenever information about a new message is received and updated in the centralized data structure, a few reflectors are notified about the message. The initial choice of reflectors may include those corresponding to the cluster of the message author, and the clusters adjacent to that. Subsequently, based on ratings given to the message, routing algorithms operating between reflectors will spread message updates to other reflectors. This helps to propagate updates about new messages to relevant reflectors without causing significant network overhead in push or pull of message updates from a central message registry. In Section 5, we describe the design of an efficient topology for the reflector overlay, such that even naive content based routing algorithms operating
on the overlay will not cause a substantial downgrading in the scalability of message updates.

6. **Client application**: A client application runs on user devices such as laptops and mobile phones. The application periodically downloads RSS feeds from the reflector corresponding to the cluster of the user, and monitors the user’s browsing history to learn the usefulness and credibility Bayesian models. When updates about new messages are received, the user-specific components of the Context(...), Completeness(...), and Credibility(...) functions are computed, and the models are used to predict the probabilities that the user may find some messages to be useful. Messages expected to be useful are then pushed to the user. Continuous connectivity with the Internet is not required, making the applications ideal to run in a delay-tolerant fashion on mobile phones. Privacy is ensured because the models are learned only at the client side. This can also make it valid to monitor user activities such as the time a user spends on a message, to tune the models.

This design addresses all the challenges enumerated earlier. Scalability to millions of new messages each day is ensured by distributing the functionality of crawling, topic categorization, and statistics maintenance across different servers. User level personalization for hundreds of millions of users is provided by partitioning the computation and personalization functions into centralized and user-specific computations. Network overhead for notifications of new messages to many users is avoided by constructing a reflector overlay for efficient routing of message updates. Lastly, privacy is ensured because user preferences do not have to be reported back to the central servers.

5. **REFLECTOR OVERLAY TOPOLOGY**

Based on the ratings given by users, we expect to eventually assemble similarity matrices for context and completeness of topic specific clusters. These matrices will tell us which clusters of users are contextually similar, and which clusters provide completeness to each other. Dimensionality reduction using Principal Component Analysis can be used to output high dimensional identity vectors for each cluster, such that users of clusters close to each other in this high dimensional space will benefit from exchanging messages with each other. Since each cluster is represented as a reflector instance, our goals are to find an overlay topology for reflector connectivity, and a routing algorithm on this overlay, such that message updates can be propagated efficiently.

We plan to develop an interest-gradient based routing algorithm, similar to [17, 18]. Client applications will advertise the preferences of their users towards contextual or complete information to the corresponding reflectors. The reflectors will advertise these preferences to other reflectors a few hops away in the reflector graph, so that interest-gradients are formed and relevant message updates can get forwarded to the appropriate reflectors.

We next present a few results from our analysis of a plausible topology for the reflector overlay. Primarily inspired by the expression of the navigability property of social networks [12], we felt that constructing an overlay based on social network relationships may turn out to be efficient. This was indeed the case, demonstrating yet another useful property of social network graphs.

In prior work [10], we crawled the Orkut website to extract a large social network graph of users, and their membership in various communities of interests. Similar communities were clustered together to yield coarse grained topic interests of users. For the purposes of this paper, we extracted three subgraphs of different sizes (10000, 17000, and 23000 users) from our original dataset, clustered the topic specific graphs in each dataset, and constructed a reflector overlay as follows. Every cluster in a topic specific graph corresponds to a node in the reflector overlay. An edge is added between two nodes if they have a user in common, or there is a pair of users, one in each cluster, having a social network tie between them. This represents the intuition that users are aware of their own topic interests (of course) and the topic interests of their friends; therefore, an edge in the reflector overlay connects nodes having common users or friends respectively. The three overlays constructed in this manner consisted of (1899, 4580, and 6593) nodes, with (92, 165, and 301) topics respectively.

With the additional assumption of homophily, [12] suggested that such topologies are likely to have the navigability property that messages can be routed to a desired destination using a simple greedy algorithm that requires routing state of size $O(\text{degree of node})$ per node, and routes messages with stringent bounds on the route length. However, we found that topic popularity in our datasets followed a power-law, whereas the model described in [12] assumed a uniform distribution of topic popularity. This resulted in disjoint subgraphs for nodes of the same topic in the overlay, which meant that greedy routing algorithms would not work on our dataset. To examine the practicality of routing algorithms requiring more routing state, we therefore examined the average node separation and maximum amount of routing state that needs to be maintained at each node.

Fig. 4 shows the distribution of average length of shortest paths between reflector nodes of the same topic in the different datasets. The CDFs to the right consider each topic specific network individually, which can be sparse and may even contain disjoint subgraphs. The separation between nodes that did not have a path to each other was assumed to be the maximum diameter in the respective dataset. The CDFs to the left consider the entire graph which allows paths to hop across different topics. As expected, the node separation improves considerably; the average separation is $\sim 3$, with 80% of the topics having a separation less than 4.
What is more interesting, however, is that the distributions remain consistent across the three datasets, indicating that node separation is independent of the number of users. A possible explanation is that of topic locality, as noticed for P2P networks [20]. This observation is encouraging because it means that nodes need not flood their interests very far into the topic node graph, and hence not cause a significant increase in the size of the routing tables of other nodes.

We next computed the lower and upper bounds for distribution of the routing state per node. We did this for the worst-case where all nodes of a topic are assumed to be contextually similar or provide completeness to each other, and hence messages should be routable between all pairs of nodes of the topic. Further, we assumed a naive routing algorithm that establishes at least one route for every pair of nodes per topic, and can hence build overlapping routes. We find the upper bound by first computing the shortest paths $d_{ij}$ between all pairs of nodes; then, for each node $x$, we compute the number of pairs of 1-hop, 2-hop, etc neighbors $(y, z)$ of the same topic such that $d_{yz} = d_{xy} + d_{xz}$ for $d_{yz} \in \{2, 6\}$. This gives us an upper bound of the maximum number of shortest paths the routing algorithm may establish between nodes of the same topic that can possibly pass through some other given node. For the lower bound, for a given ordering of choosing nodes $x_1, x_2, \ldots$, we count $(y, z)$ only the first time it occurs as a shortest path passing through some node $x_i$. After repeated iterations with different orderings of choosing nodes $x_1, x_2, \ldots$, the mean distribution of the number of shortest paths across different orderings gives the lower bound distribution.

Fig. 5 shows the distribution for the largest dataset. The routing state for $d = \{1, 5, 6\}$ was insignificant. It can be seen that the majority of nodes need state $< 500$ for $d = 2$. The state increases for $d = \{3, 4\}$ but the lower bounds remain below 1000, while the upper bounds remain below 4000 for the majority. The routing state was also found to remain unchanged with the number of users, giving more evidence of topic locality in the reflector overlay. We next analyzed the overlays consisting only of nodes for popular topics with more than 200 users, which represented a 20% cutoff from the maximum topic size. The routing state was again found to remain unchanged, which indicates that popular topics ensure connectivity, while unpopular topics “hang” off the popular topics. This is an interesting observation; in the future we plan to build mathematical models to characterize such graph properties. For now, given the small node separations and routing state, we safely conclude that the proposed topology is feasible to use.

6. CONCLUSIONS

We have proposed the design of a personalized and scalable recommender system for participatory media content using social network theory. Other methods such as CF [21–23] are only able to replicate average user behavior, whereas our approach uses sociological theory to understand the contextual reasons behind a user’s preferences, and can hence produce closer-fitted recommendations. Although many open problems remain to be solved, we are hopeful that our methodology of integrating elements of information science and sociology will improve the state-of-the-art in recommender systems for participatory media.

7. REFERENCES