An Infrastructure for Participatory Media *

Aaditeshwar Seth – a3seth@uwaterloo.ca

School of Computer Science
University of Waterloo, ON, Canada

Introduction

Purchasing decisions can be hard to make. For example, the decision to purchase a house by a person can be based on many parameters such as the price of the house, distance from her workplace, quietness of the neighborhood, etc. It is hard to capture all such parameters in a recommender system or a knowledge-sharing website. Further, many such parameters are highly subjective and difficult to evaluate or quantitatively compare with respect to other parameters. Therefore, people very often refer to their friends for advice, or search for relevant blogs on the Internet that may reveal some new insights. However, how can the person be sure that opinions from friends represent fairly diverse viewpoints so that she can make a wise choice? How can she judge the reliability of information obtained from blogs? An attempt has been made to address such questions in this work through the use of social networks. Specifically, the questions that are answered are as follows:

1. What characteristics of a person's social network help her get information that is diverse, relevant, and reliable?
2. Assuming most such information to be in the form of discussions in which various people participate, what characteristics of the social network formed by the participants, help determine diversity in the discussion, its relevance for the person, and its reliability?

Participatory media

It is important to notice that the same concerns for relevant, diverse, and reliable information also surface in reference to political news, or news that guide personal and professional economic decisions. Such decisions are evaluated subjectively and partially, because an objective and complete analysis is difficult. Further, information sources that supply crucial information to the decision making process can be motivated by commercial or political incentives to purposely create information asymmetry or bias.

Traditional mass-media such as newspapers, television, and radio, that are responsible for providing this information, do not support adequate diversity in information sharing, and can often be unreliable as well. Although these limitations can be overcome to a large extent by the new forms of *self-regulated* participatory media, including blogs, wikis, and podcasts, it is argued that the current tools and services (such as Google, Slashdot, Youtube, Amazon, etc) to search and disseminate this information are inefficient and have many missing features (Seth 2007). This affects the ability of these new forms of media to provide the necessary system of checks and balances to ensure relevant, diverse, and reliable information delivery.

Therefore, rather than considering only e-commerce related recommender systems, the problem is considered in a much broader context of recommender systems for participatory media.

Good information

To answer these questions, it is necessary to first develop a notion of 'good' information. Reconsidering the house purchase example:

Simply notification about a new property sale may not be as valuable to the person as when presented to her along with an analysis of relevant factors. However, because such analysis can be highly subjective, different views should be presented to the person to help her discriminate between multiple choices. In addition, information presented to the person should be produced by reliable sources that are accountable for the information they provide.

This motivates three essential features of 'good' information.

1. *Context*: Information is easier to understand when it is presented in reference to conditions or circumstances relevant to its recipients.
2. *Completeness*: People should be presented with a diversity of views, to prevent bias due to ignorance, or politics or commerce.
3. *Reliability*: This refers to the degree of trust that is placed in a source of information by an information consumer. Note that different notions of 'reliability' can be attached to information depending upon the circumstances and viewpoints of the entities receiving the information. Thus,

*This research was supported by grants from the National Science and Engineering Council of Canada, the Canada Research Chair Program, Nortel Networks, Intel Corporation, and Sprint Corporation.

the trust placed in an information source depends on the context of its recipients. This is especially true for subjective information, where each person may have her own notion of reliability for an information source. Individual notions of reliability can be aggregated across entities to form collective notions as well.

In this paper, the concept of *information capital* is introduced that specifies graph theoretic measures for context, completeness, and reliability, based on the structure of social networks. The *information capital of a person* denotes her ability to receive 'good' information from members in her social network, ie. information which is contextually relevant, complete, and reliable. This concept is then extended to *information capital of discussions*, which denotes the 'goodness' of information that the discussion carries. This extension is then applied to develop a measure of the *fractional information capital* gained by a person from viewing or participating in the discussion.

The concept of fractional information capital is then used to outline the design of an infrastructure for information recommendation, that recommends information to explicitly maximize the information capital that people gain from the information. The infrastructure will thus be able to provide guarantees on the relevance, diversity, and reliability of the information presented to different people.

To the best of my knowledge, I am not aware of any other recommender system that considers context, completeness, and reliability. Most recommender systems address the need for contextual information through personalization, or reliability through reputation systems, but completeness is often absent in these systems.

Outline

The concept of information capital is described next. Due to space constraints, only methods to calculate the *information capital of people* are described. Details of how to extend it to *information capital of discussions* and *fractional information capital* gained from viewing or participating in a discussion, are explained in (Seth 2007). This is followed by an outline of the system model for an infrastructure for participatory media.

Information capital

Referring to the *strength-of-weak-ties* hypothesis (Granovetter 1973), the terminology of 'strong' and 'weak' ties was introduced to state that not only do clusters of people exist with 'strong' ties among members of each cluster, but these clusters are linked to each other through 'weak' ties themselves. 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 particularly important for the diffusion of information, influence, and economic mobility, because weak ties help connect diverse clusters of people with each other.

The hypothesis has been found to be true in a number of situations, and has formed the basis of a theory of *social capital*. Social capital is defined as a network characteristic emerging from the *structure* of social networks, or the *resources* embedded in networks, which can be used as a measure of the social assets available to people or communities from their social network (Lin 1999). For example, the number of weak acquaintances of a person that can help her get a new job, or the density of strong ties in a community that can help in reaching consensus on various economic and political issues, can all be considered as measures of social capital.

The same concept of strong and weak ties can be applied in relation to the system described in this paper. The immediate social network of a person comprised of strong ties can be considered as enhancing the context of information that circulates among these ties. For example, greater clustering in a person's neighborhood is likely to lead to easier contextualization of information by its circulation in the neighborhood of the person. Similarly, the remote social network of a person comprised of weak ties can be considered as providing completeness, ie. capable of enhancing the diversity of information that a person is likely to come across. Completeness is likely to increase with an increase in the size of the network spanned by the person's weak links.

The concept of social capital can thus be extended to *information capital*, which measures the ability of people to receive 'good' information based on their 'strong' and 'weak' ties. Context is enhanced through 'strong' ties, completeness is enhanced through 'weak' ties, and the reliability of neighbors constituting these ties gives an indication of the quality of context or completeness of information that is provided by them. Thus, information capital can be considered as a tuple of ($context$, $completeness$), calculated based on the structural properties of the person's social network. Reliability is included as a part of the functions for calculating context and completeness.

It is important to realize that information capital is a topic-specific characteristic. For example, a person's information capital related to understanding climate change is not enhanced by members of her social network who are not interested in the topic. Therefore, the social networks used for calculating information capital should be topic-specific.

The following approach can be used to calculate context and completeness:

1. Extract the topic specific social network constituted only of those people interested in a specific topic.
2. Cluster the topic specific social network such that people within each cluster have strong links between them, and people in different clusters are connected with weak links.
3. For each cluster of strong ties V, calculate its clustering coefficient (Newman 2003).

   $C_V = \frac{1}{|V|} \sum_i \frac{|\{\triangle' s \ centered \ on \ u_i\}|}{\binom{d_i}{2}}$

   Here, d_i is the indegree of person u_i. Thus, $\binom{d_i}{2}$ is the maximum number of $\triangle$'s (triangle's) that can be centered around u_i. The fraction within the summation is called the clustering coefficient of person u_i, and the clustering coefficient of the strong cluster is calculated as the average of the clustering coefficients of its people. A high clustering coefficient indicates a high probability that people of this cluster will participate in discussions relevant to the topic.

4. For each person, calculate her integration coefficient into her strong cluster (Valente 1995).
   $t_i = \frac{1}{(|V|-1)D_V} \sum_{u_j \in V}(D_V - d(i,j))$
   Here, $d(i,j)$ denotes the distance from person u_i to u_j, calculated as shortest path between the two people. D_V denotes the diameter of the cluster V = maximum distance between any two people $\in V$. Thus, the integration coefficient will be close to 1 for people who are well integrated in their cluster, ie. they are close to many other people. It will be close to 0 for people who are present along the boundaries of the cluster and are not well integrated. A high integration coefficient for a person indicates a high probability that other people will help contextualize information for her.
5. Now, calculate context as follows:
   $Context_i = C_V \; t_i \; |V|$
   Thus, context denotes the ability of people to receive contextual information from their strong links, and is enhanced by improved clustering and integration.
6. Let w_j denote the number of weak links of person u_i into cluster V_j, where V_j is not the same as the cluster of strong ties of person u_i. Then:
   $Completeness_i = \frac{1}{\sum w_j} \sum |V_j| \; w_j$
   Thus, completeness denotes the ability of people to receive information from diverse parts of the social network through their weak links.

The method outlined above can be enhanced with reliability measures as well. Individual or collective notions of reliability can be calculated using approaches similar to those outlined in (R. Guha & Tomkins 2004; Langville & Meyer 2004). Various other examples of functions for calculating context, completeness, and reliability as Eigenvector feedback centralities are also given in (Seth 2007).

Information capital of discussions

Similar definitions of information capital can be used for each discussion or a collection of discussions as well. The density of strong links among people participating in a discussion indicate context, and the network-span covered by weak links between people indicate completeness. Context and completeness can be combined into a single measure depending upon the discussion topic. For example, the amount context in a discussion is likely to be sufficient to measure the 'goodness' of personal discussions, but both context and completeness will be important for subjective discussions such as political opinions. Initial user-studies indicate that information capital of people and discussions indeed reflect user perceptions about 'good' information in different scenarios.

Fractional information capital

Although information capital of discussions denotes a generic measure of the 'goodness' of information carried by a discussion, it does not say anything about the relevance of the discussion for a person. However, similar intuitions can be used to develop a measure for the amount of additional information capital that a discussion will provide to a person. This is termed as the *fractional information capital* provided by the discussion to person u_i, and is denoted by functions $Context(X, u_i)$ and $Completeness(X, u_i)$. Here, X refers to the active environment of a discussion comprised to people currently participating in the discussion.

System outline

It is proposed to overcome the short-comings of current mass-media and new-media tools by building a distributed infrastructure for participatory media based on social networks of people. The propagation of information units (such as news articles, blogs, podcasts, and videos, referred to as *discussions* or *messages* henceforth) in the social network is modeled as an information diffusion process. Here, users are considered as content-based message routers who give comments on messages, link to other messages, and forward messages to their friends based on the interests of their friends. The information sharing system is meant to assist users through automation of the diffusion process, along with ensuring guarantees on the 'goodness' of information provided to users.

In addition, fairness guarantees are provided in the information diffusion process, so that all messages have an equal opportunity to popularize themselves. This is done to mitigate the preferential-attachment behavior inherent to complex evolving systems, so that even relatively unknown resources are not undermined, and new viewpoints can be brought to light.

The 'goodness' properties of the information sharing system will allow for:

1. *Information push*: The system will push relevant information to interested users in a semi-automated fashion. The mode of delivery will be similar to RSS feeds or 'Google-alerts', except that the system will help identify sets of messages that together provide contextual, complete, and reliable information for users.
2. *Information pull*: Users will be able to search for information in the network. The set of messages returned will together be contextual, complete, and reliable. Additional tools will be provided such that if a user visits an external resource, such as a website, then related information will be shown alongside to convey a more complete picture.
3. *Query information*: Users will be able to dispatch queries into the network, and get response from a diverse set of participants who share a common context with them. Queries are more suitable than information-pull for gathering tacit information that has not been explicitly stated as yet by other users.

Only the push system is described in this paper. Essential building blocks of the system are identified next.

System model

Information diffusion is modeled as taking place in a multi-agent interaction environment with two kinds of agents.

1. *User agent*: A user agent corresponding to each user models user-behavior by predicting the actions that a user might perform when she is presented with a message.
2. *Meme* [1]*agent*: Each message is carried by a corresponding meme agent that tries to maximize its objective function in the presence of other meme and user agents. This objective function can be designed to increase the information capital of users, or the information capital of the message, or other variations.

User agent The user agent for each user u_i tries to learn $P_i(A_i = a|X)$, ie. the probability of the user taking action $A_i = a$ given the active environment X. Here, actions include {*Like, Dislike, Forward, Comment, Junk*}, and the environment is defined using features such as the number of 'strong' ties of a user who $like$ a message, the number of weak ties, and features based on other information cascade patterns (J. Leskovec & Kleinberg 2006). The user behavior model can be derived using Conditional Random Fields (CRFs) (Wallach 2004), that seems to be most suitable for the purpose.

Influence and reputation The probability distribution $P_i(A_i = a|X)$ for user agents in fact defines the transition probabilities among different environment states connected in a Markov chain. This can be used to create a model of influence and reputation propagation; for example, the expected number of users who $like$ a message recommended by some user indicate the reputation of the user. Influence can be measured as the difference between the expected number of users who like a message with and without a recommendation by the user.

A Markov chain spanning all possible states is intractable to solve. Statistical approximations can be made where influence and reputation are calculated in an incremental manner directly from observed samples of the environment.

Meme agent Meme agents try to maximize their objective function, which can be defined in different ways. An example based on the fractional information capital is as follows.

Depending upon the message topic and user preferences, a parameter λ_i can be used to combine context and completeness of the message to denote the fractional amount of information capital gained by a user from the message.

$$IC_i = Context(X, u_i) + \lambda_i Completeness(X, u_i)$$

Here, $Context(X, u_i)$ and $Completeness(X, u_i)$ are the same as the functions described earlier. The objective of the meme agent is then to maximize the overall utility by selecting a set of users in discrete stages to whom the message is pushed in succession.

Fairness in meme propagation can be ensured through constrained maximization, where each meme agent is given a fixed amount of credit at the beginning. Credit is gained if users make a positive action upon receiving the message, and is proportional to the reputation of the users. Credit is lost whenever the message is pushed to any user, and is proportional to the influence of the user. Thus, each message is given an equal opportunity to popularize itself.

The meme objective can now be expressed as follows: At each stage, select the best set of users F, yielding:

$$argmax_F \sum_{u_i \in F} P_i(A_i = a|X) IC_i$$

Here, the actions $A_i = a$ denote actions having a positive response. Maximization is done under constraints that the meme-credit for maintaining fairness never drops below 0.

Other parameters can also be added to the objective function, such as the rate of increase of information capital of messages, or the information capital of message aggregations. Similarly, the credit scheme for fairness can be extended to give extra credit to meme agents that help increase the information capital of users.

[1]The term 'meme' was introduced in *The Selfish Gene* by Richard Dawkins. Memes are units of cultural information that propagate and sustain themselves in societies.

Conclusions

In this paper, the importance of contextual, complete, and reliable information for better decision making was stated. These measures of information quality can help improve the design of e-commerce recommender systems, and are equally applicable in broader scenarios of recommender systems for participatory media. A novel infrastructure for participatory media based on social networks was then proposed, and essential components required to build this system were identified.

It seems plausible to build the envisioned system as a distributed infrastructure, with user agents implemented as part of client-side applications and meme agents implemented in distributed databases. However, there are clearly many open and interesting problems, both in terms of theoretical formulation of the system as well as its implementation. I am actively exploring these issues in current work.

References

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

J. Leskovec, A. S., and Kleinberg, J. 2006. Patterns of influence in a recommendation network. *PAKDD*.

Langville, A., and Meyer, C. 2004. A survey of eigenvector methods for web information retrieval. Technical Report http://meyer.math.ncsu.edu/Meyer/PS_Files/Survey.pdf.

Lin, N. 1999. Building a network theory of social capital. *International Sunbelt Social Network Conference*.

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

R. Guha, R. Kumar, P. R., and Tomkins, A. 2004. Propagation of trust and distrust. *WWW*.

Seth, A. 2007. An infrastructure for participatory media. Technical Report http://www.cs.uwaterloo.ca/ a3seth/pmtech.pdf.

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

Wallach, H. 2004. Conditional random fields: An introduction. Technical Report http://www.inference.phy.cam.ac.uk/hmw26/crf/.