Modeling Dynamic Multi-Topic Discussions in Online Forums

Hao Wu, Jiajun Bu, Chun Chen*, Can Wang, Guang Qiu, Lijun Zhang and **Jianfeng Shen†**
College of Computer Science, Zhejiang University, Hangzhou, China
{haowu, bjj, chenc, wcan, qiuguang, zljzju}@zju.edu.cn
*Corresponding Author
†Zhejiang Health Information Center, Hangzhou, China
sjf@zjwst.gov.cn

Abstract

In the form of topic discussions, users interact with each other to share knowledge and exchange information in online forums. Modeling the evolution of topic discussion reveals how information propagates on Internet and can thus help understand sociological phenomena and improve the performance of applications such as recommendation systems. In this paper, we argue that a user's participation in topic discussions is motivated by either her friends or her own preferences. Inspired by the theory of information flow, we propose dynamic topic discussion models by mining influential relationships between users and individual preferences. Reply relations of users are exploited to construct the fundamental influential social network. The property of discussed topics and time lapse factor are also considered in our modeling. Furthermore, we propose a novel measure called ParticipationRank to rank users according to how important they are in the social network and to what extent they prefer to participate in the discussion of a certain topic. The experiments show our model can simulate the evolution of topic discussions well and predict the tendency of user's participation accurately.

Introduction

With the flourish of Web 2.0 applications, we have witnessed a great deal of online social medias (such as forums, Weblogs, News Groups, Question-Answering Communities, etc.) emerge and thrive to become popular. Among these prevalent social medias, *online forums* (or *message boards*) are characterized as a unique type of platforms for information exchanging and knowledge sharing. In such platforms, users interact with each other primarily in the form of topic discussions. Usually, the content of a discussion in online forums is visually and structurally threaded, and thus facilitates users to write *comments* (or *posts*) in existing topics or create new topics. Discussion threads about a specific theme (e.g., *sports*) are grouped in each distinct *board* (or *community*). A discussed *topic* can be a technical question, a news event, a description of a product, or even a point of view, etc.

One important task in online forums is to model the evolution of topic discussions. The modeling results can reveal how information propagates via the underlying social network on Internet and thus can (i) help researchers solve many psychological and sociological problems such as human interactions and group forming (Backstrom et al. 2006); (ii) analyze social influences (Tang et al. 2009) to improve the performance of applications such as recommendation systems (Shi et al. 2009); (iii) track the emergence and popularity of new ideas and technologies.

However, online forums show great complexity (Gómez, Kaltenbrunner, and López 2008). Different from the explicit co-authorships or friendships in common social networks such as DBLP and Livejournal (Backstrom et al. 2006), the relationships or links between users in most online forums are hidden and dynamically developed through topic discussions. Popular online forums always have thousands to millions of active users, with a great diversity of individual preferences as well as roles they play. The users' participation behaviors of discussions exhibit relative randomness and may change over time. In a community, there are usually tens to hundreds of threads interweaving in discussions at the same time. Moreover, topic may drift over time even in an individual thread. Therefore, modeling the evolutional multi-topic discussions in online forums is challenging.

In this paper, we propose Topic Flow Models (TFM) to model the evolutional multi-topic discussions in online forums, which is based on the intuition of information flow (Song et al. 2006). We focus on the following four central questions in simulating the process of topic discussions:

1. What are the main mechanisms underlying user's participation in topic discussion?
2. From which perspective should we view the process of topic discussion, in order to model it theoretically and systematically?
3. How can we make use of knowledge such as the property of topics and temporal feature to characterize topic discussion for modeling?
4. How can we measure the importance of each user in the process of topic discussion?

Overview

In this section, we present the intuitions of our modeling algorithm and the problem formulation.

Intuitions

Why people join in a discussion and post comments? Intuitively, it is supposed that a user tends to post comments in response to comments posted by her "friends", or in response to comments that are interesting to herself. We thus reason that *peer-influence* and *self-preference* mechanisms are two of the most important factors that influence user's participation in topic discussions. This corresponds to the first question in the above section.

In online forums, given the observation that a newcomer may read some of the previous posts before posting, we find the process of user's participation in topic discussion can be modeled in the perspective of information flow (Song et al. 2006), that is, *the information is flowing from early posters (users who post) to late posters*. We here consider *topics* as a special kind of information that can "spread" through the social network in the process of discussions. A user's adoption of a topic (i.e, participation in discussion on a topic) is influenced by her "neighbors" in the social network as well as her own preferences of topics. This corresponds to the second question in the above section.

In real data, there exists various latent topics in discussions of a community in online forums. Regarding to different topics, users' participation patterns of discussion and social influences between them are different (Tang et al. 2009). We thus explore users' different hidden social networks associated to different topics. Moreover, user's participation behaviors change over time. Hence, the *time lapse factor* should be incorporated into our modeling. This corresponds to the third question in the above section.

Based on the above discussion, we hence naturally model the evolution of topic discussions as *Random Walks* (Lovasz 1993) of *topics* on graphs that corresponds to the underlying social networks, where users are represented as nodes and their relationships are represented as directed weighted edges. The stationary-state probability for each node of the random walk represents how likely a topic flowing in the network will arrive at a certain user, or the importance and willingness of a user in participation to the discussion of a certain topic. This corresponds to the fourth question in the above section.

Problem Formulation

The core of our dynamic multi-topic discussion modeling is to mine the underlying social network associated to users' adoptions of a certain topic. We formally define:

Data Input

- **Thread Document:** We use D to denote the set of discussion thread documents, where $d \in D$ is a *thread document* that contains the text of all posts (comments posted by users) in a thread of a community.
- **Reply Link:** The most explicit links between users in online forums are the reply links. We use R_{ij}^d denote the frequency of a user u_i replied by u_j in a thread document d. Besides, let C_i^d be the number of comments posted by u_i that exclude the comments in response to any other users in d, that is, the number of comments in response to the *thread root* (the initial post of the thread).

Data Output

- **Influential Network:** We use a directed graph $G = (V, E)$ to model the underlying influential network (social network), where V is the set of nodes with a size $|V| = n$, and E is the set of $n \times n$ edges. Each node $v_i \in V$ represents a user u_i, and each directed weighted edge $e_{ij} \in E$ represents the influential relationships of u_i to u_j in adoption of topics. Let $\mathbf{W}$ be the $n \times n$ affinity matrix where each entry w_{ij} denotes the edge weight of e_{ij}, i.e., the strength of the *peer-influence* of the two users.
- **Topic-level Influential Network:** Considering the latent topics in discussions, we define a graph $G^z = (V^z, E^z)$ to represent the influential network associated to a latent topic z. Topic modeling methods such as pLSI (Hofmann 1999), LDA (Blei, Ng, and Jordan 2003) and LTM (Cai, Wang, and He 2009) can be used to analysis the latent topics. Correspondingly, we use v_i^z, e_{ij}^z, w_{ij}^z, $\mathbf{W}^z$ to denote a node, an edge, its weight and the affinity matrix associated to a latent topic z.
- **User Preference:** We use a vector $\mathbf{y} = [y_1, \cdots, y_n]^\top$ to denote the users' preferences of topic discussion, where each element y_i represents how likely a user u_i prefers to join in a discussion by *self-preference*, and this factor is independent of the *peer-influence* of her neighbors in the network. Correspondingly, let $\mathbf{y}^z$ denote the preferences of users in discussion associated to a latent topic z.
- **ParticipationRank:** To measure the importance and willingness of a user in participation to discussion of a certain topic, we define a ranking called ParticipationRank, which is denoted as a vector $\mathbf{p} = [p_1, \cdots, p_n]^\top$, where p_i corresponds to the ranking score of u_i.

Topic Flow Models

In this section, we go into full exposition of our Topic Flow Models. First, we describe Basic Topic Flow Model (B-TFM) which does not consider the latent topics. Then we extend to Topic-specific Topic Flow Model (T-TFM). Furthermore, we consider time lapse factor and introduce Time-sensitive Topic-specific Topic Flow Model (TT-TFM).

Basic Topic Flow Model

Since discussion threads in a board are more or less related to a specific theme (e.g., *sports*), we first describe Basic Topic Flow Model (B-TFM) without multiple latent topics.

We model the topic discussions by mining the users' adoption behaviors of a topic. By *one adopts a topic*, we mean *one participates in discussion of a topic*. Participation typically consists of browsing and posting comments, though, only posts are visible. We thus view *a user's joining in a discussion* as *posting at least one post in the discussion*.

Topics are considered as a special kind of information that can diffuse or flow among nodes of the influential network through edges until a stationary state is established. It is analogous to a random surfing on the web graph along web

links in PageRank (Brin and Page 1998). Intuitively, u_j follows u_i to join in discussion as u_j replies u_i, and it is supposed information flows from u_i to u_j. We thus exploit the frequency of each user u_i replied by u_j, which is R_{ij}^d in a thread document d, and define each element w_{ij} of the affinity matrix $\mathbf{W}$ associated to the influential network G as

$$w_{ij} = \sum\nolimits_{d \in D} R_{ij}^d \tag{1}$$

Here w_{ij} represents the strength of the influence of u_i to u_j in adoption of a topic. The transition probability matrix $\mathbf{S}$ determining the random walk on G can be defined as

$$\mathbf{S} = \alpha \mathbf{D}^{-1}\mathbf{W} + (1-\alpha)\mathbf{N} \tag{2}$$

where $\mathbf{D}$ is the diagonal matrix with (i,i)-element equals to the sum of the i-th row of $\mathbf{W}$. Note here to deal with the rows of $\mathbf{W}$ that are summed to zero, we replace each element of these rows with 1/n. $\mathbf{N}$ is the matrix with all elements equal to 1/n. Eqn. (2) can be interpreted as a probability α of transition to an adjacent node, and a probability $1-\alpha$ of jumping to any node on the graph uniformly at random. By introducing the term $(1-\alpha)\mathbf{N}$, we ensure the transition matrix $\mathbf{S}$ irreducible and the graph G strongly connected.

The transition probabilities represent the *peer-influence* that how likely users influence each other in adoptions of a topic. We then consider the *self-preference* factor and exploit comments posted by a user u_i that exclude the comments in response to any other users, the number of which is C_i^d in a thread document d. We define each element of the user preference vector $\mathbf{y} = [y_1, \cdots, y_n]^\top$ as

$$y_i = \sum\nolimits_{d \in D} C_i^d \tag{3}$$

The normalized vector $\mathbf{q} = [q_1, \cdots, q_n]^\top$ is given by

$$q_i = y_i / \sum\nolimits_{i=1}^{n} y_i \tag{4}$$

The stationary-state probability distribution of the random walk (i.e., ParticipationRank $\mathbf{p}$) over all nodes can be obtained by repeatedly iterating the following equation

$$\mathbf{p}_{(t+1)} = \beta \mathbf{S}^\top \mathbf{p}_{(t)} + (1-\beta)\mathbf{q} \tag{5}$$

where $\mathbf{p}_{(t)}$ is the ParticipationRank vector in t-th iteration, and β controls the balance of *peer-influence* and *self-preference* mechanisms. This is analogous to personalized PageRank (Langville and Meyer 2004) and corresponds to a problem of Random Walks with Restarts (Lovasz 1993), where $\mathbf{p}_{(t)}$ will converge to $\mathbf{p}^*$. Substituting $\mathbf{p}^*$ for $\mathbf{p}_{(t+1)}$ and $\mathbf{p}_{(t)}$, we have

$$\mathbf{p}^* = \beta \mathbf{S}^\top \mathbf{p}^* + (1-\beta)\mathbf{q} \tag{6}$$

Following some algebraic steps, we can finally obtain

$$\mathbf{p}^* = (1-\beta)(\mathbf{I} - \beta \mathbf{S}^\top)^{-1}\mathbf{q} \tag{7}$$

where $\mathbf{I}$ is the identity matrix. We can use this closed form to compute the ParticipationRank, which measures the importance and willingness of each user in participation to the discussion of a certain topic, or reflects how likely a topic will arrive at a node (user) in the network in the perspective of information flow.

Topic-specific Topic Flow Model

In this subsection, we extend to Topic-specific Topic Flow Model (T-TFM) for topic discussions.

In real data, there exists various latent topics in discussions of a community of online forum. Regarding to different topics, user's participation patterns of discussion are different. An active participator of topics about politics may not be interested in sports. We thus explore users' different influential networks associated to different latent topics. Here, we view each discussion thread document as a probabilistic mixture over T latent topics, that is, a thread document d can be clustered into a topical class $z \in Z = \{z_1, \cdots, z_T\}$ with the probability $P(z|d)$. We can use $P(z|d)$ to represent the strength of topic flow regarding to z for those users who join in discussion of d. Here for each latent topic z, an independent corresponding influential network needs to be generated. We adapt the affinity matrix $\mathbf{W}$ for influential relations between users in Eqn. (1) and obtain $\mathbf{W}^z$ corresponding to z as follows:

$$w_{ij}^z = \sum\nolimits_{d \in D} P(z|d) R_{ij}^d \tag{8}$$

Correspondingly, we adapt user preference vector $\mathbf{y}$ in Eqn. (3) to obtain $\mathbf{y}^z$ associated to z as follows:

$$y_i^z = \sum\nolimits_{d \in D} P(z|d) C_i^d \tag{9}$$

The computations of $\mathbf{S}^z$, $\mathbf{q}^z$ and $\mathbf{p}_z^*$ can be easily obtained by still applying Eqn. (2), (4) and (7) respectively, where the only thing we need to modify is substituting vectors or matrices with ones that are subscripted by z. Here we construct a set of ParticipationRank $\{\mathbf{p}_z^*\}_{z \in Z}$ to measure the willingness of each user to participate in discussion of each latent topic z.

Note that in two extreme cases: when $T = 1$, and when the probabilities of latent topics in each document $\{P(z|d)\}_{z \in Z}$ have a uniform distribution, Topic-specific Topic Flow Model reduces to Basic Topic Flow Model.

To analyze the latent topics, we adopt Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), which is a well-defined generative model for topic modeling. Moreover, it has no overfitting problem to which Probabilistic Latent Semantic Indexing (pLSI) (Hofmann 1999) is susceptible.

Time-sensitive Topic-specific Topic Flow Model

In this subsection, we propose Time-sensitive Topic-specific Topic Flow Model (TT-TFM) by incorporating *time lapse factor* for topic discussions.

User's preferences of topics change over time in real data. For example, an enthusiast of indoor sports may adopt interests in travel gradually, which may in turn diminish his or her passion in indoor sports. Thus user's participation patterns of topics during different time have various contributions to the characterization of the user's current or future behaviors. It is intuitive that one's most recent activity is relatively good indicator for the tendency of one's participation behavior, rather than that of long time ago. Hence, we introduce time lapse factor and reformulate the affinity matrix associated to latent topic z, i.e., $\mathbf{W}^z$ in Eqn. (8) as:

$$w_{ij}^z = \sum\nolimits_{d \in D} \exp(-\tau \cdot \Delta t_d) P(z|d) R_{ij}^d \tag{10}$$

where Δt_d represents how much time lapse from the indicating time of the thread d (e.g., the disclosing time) to current time (or time for prediction), and τ is a forgetting parameter that reflects how quickly user's adoption behaviors change over time. We deal with the user preference vector $\mathbf{y}^z$ in Eqn. (9) likewise:

$$y_i^z = \sum\nolimits_{d \in D} \exp(-\tau \cdot \Delta t_d) P(z|d) C_i^d \qquad (11)$$

The computations of $\mathbf{S}^z$, $\mathbf{q}^z$ and $\mathbf{p}_z^*$ can be generated just as we discussed in the above subsection.

Note that when $\tau = 0$, this time-sensitive model reduces to the Topic-specific Topic Flow Model.

Experimental Results

In this section, we first describe the details of our datasets. Then we present the experiments and the evaluation results of our proposed topic discussion models.

Description of Data Sets

For data preparation, we selected the popular online forum Honda-tech[1], which provides a platform for users to exchange information about Honda products. Here we used the data in two of the most representative and largest communities "Drag Racing" and "Honda/Acura". The community "Drag Racing" is designed for the fans of car racing to share their hobbies, while "Honda/Acura" is a community for customers to exchange information of "Acura" cars.

We manually wrote a wrapper to crawl the metadata (includes *Timestamp*, *User Name*, *User ID*, *Replied User Name* and *Message Text*) of all threads in the two communities across one year period. The time window ranges from Sept. 1st, 2008 to Aug. 31st, 2009. Finally, we generated two datasets with information of time, texts, users and explicit reply relations preserved. We filtered out threads having only one post. The statistics are listed in Table 1. We selected users who have posted more than the average number of posts over *all users* as *active users*, who are supposed to show more regularity of participation than ordinary users.

Evaluations and Experiment Setup

In order to measure the performance of our models, we investigate the task of predicting user's participation of discussions in an unseen period. The ParticipationRank measures to what extent each user prefers to join in discussion of a topic, and thus can be used as an indicator.

We divide each of the two datasets into 12 continuous time windows. Each time window is one month long. The evaluations are conducted by examining the relation between user activity at time window t (*period for training*) and in the *first week* of time window $t+1$ (*period for prediction*). We thus have 11 *experiment groups* for each dataset. We then divide each dataset into two parts, one with 6 experiment groups for model tuning and the other with the left 5 experiment groups as held-out data for model validation.

For each time window t except the last one, the estimated ParticipationRank $\mathbf{p}^*$ is calculated for B-TFM. However, a set of ParticipationRank $\{\mathbf{p}_z^*\}_{z \in Z}$ are generated in T-TFM or TT-TFM. We synthesize $\mathbf{p}^*$ for both of them as follows:

$$\mathbf{p}^* = \sum\nolimits_{z \in Z} \sum\nolimits_{d \in D_F} P(z|d) \mathbf{p}_z^* \qquad (12)$$

where D_F is the set of thread documents disclosed in the future (during *period for prediction*). The sum of the probabilities of topic z in threads, $\sum_{d \in D_F} P(z|d)$, indicates the probability that topic z is discussed in *period for prediction*. We calculate the sum of the products, each of which is $\sum_{d \in D_F} P(z|d) \mathbf{p}_z^*$ corresponding to a latent topic z, to obtain the final values of $\mathbf{p}^*$ for ranking.

Now each entry value of $\mathbf{p}^*$ in each of the three models indicates the probability that a user prefers to join in discussion of a topic. To predict whether a user will join in discussion during an unseen time period is a binary classification task. However, in our experiments, most entry values of $\mathbf{p}^*$ are below 0.01. Hence, it is not appropriate to use an arbitrary cut-off value (e.g., 0.5) for classification. Instead, we generate a ranking list of users, which is sorted by their corresponding values in vector $\mathbf{p}^*$. Then we look up the ranking list from top to bottom, to examine whether the users appear (*post at least once*) in the first week of time window $t+1$. Ground truth of one week data is modest to both avoid randomness of user activity and keep sensitivity for evaluation.

Table 1: General Description of Datasets

Statistics	Data Sets	
	Drag Racing	Honda/Acura
# of threads	2375	4018
# of posts	68904	127913
# of users	3658	9193
# of active users	640	1674
avg. posts/thread	29.0	31.8
avg. posts/user	18.8	13.9

Evaluation Metrics

We employ the following ordering metrics to evaluate the ranking lists generated using each model, and evaluation results are averaged over the given experiment groups.

1. Precision at a cut-off rank k, where $k = 10$ (P@10).
2. Average Precision (AP).

Parameter Settings

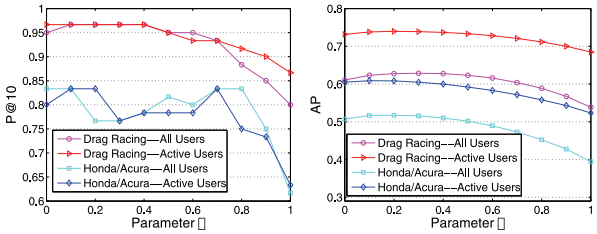
We set the damping factor α to 0.85 as usual. There are still several parameters must be fixed in our models. We tune them on the training data with 6 experiment groups.

- **Combination weight** β. The parameter β controls the balance of *peer-influence* and *self-preference* mechanisms of user's participation in topic discussion. We tune parameter β using B-TFM. Figure 1 depicts the performance of B-TFM with different values of β. It shows $\beta = 0.3$ is a good choice for "Drag Racing". However, a relatively small value $\beta = 0.1$ achieves good performance for "Honda/Acura". We reason that "Drag Racing" related to hobbies exhibits more strong ties between users

[1] http://www.honda-tech.com

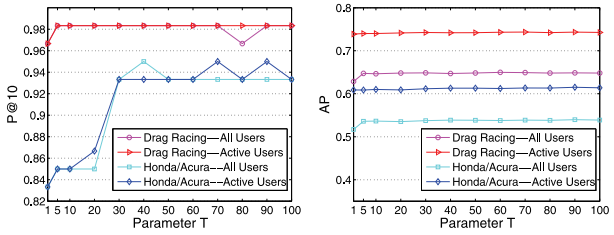
Table 2: Performance of all methods: $\beta = 0.3$ (B-TFM parameter), $T = 30$ (T-TFM parameter) for "Drag Racing" and $\beta = 0.1$, $T = 40$ for "Honda/Acura", $\tau = 0.01$ (TT-TFM parameter) for both communities.

Metrics	Methods	Drag Racing		Honda/Acura	
		All users	Active Users	All users	Active Users
P@10	Random	0.320	0.540	0.260	0.440
	PostNum	0.840	0.840	0.860	0.860
	B-TFM	0.920	0.920	0.920	0.920
	T-TFM	**0.960**	**0.960**	0.920	0.920
	TT-TFM	**0.960**	**0.960**	**0.940**	**0.940**
AP	Random	0.329	0.544	0.261	0.443
	PostNum	0.594	0.693	0.535	0.607
	B-TFM	0.620	0.722	0.557	0.636
	T-TFM	0.638	0.728	0.570	0.645
	TT-TFM	**0.643**	**0.736**	**0.576**	**0.652**



(a) P@10 vs. β (b) AP vs. β

Figure 1: Performance of B-TFM vs. parameter β

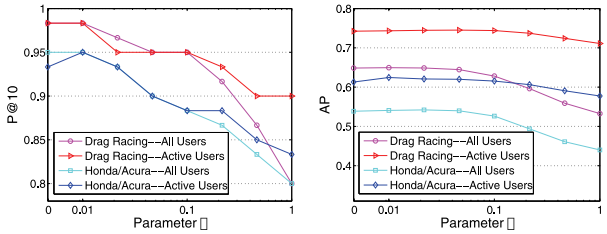


(a) P@10 vs. T (b) AP vs. T

Figure 2: Performance of T-TFM vs. parameter T.

than "Honda/Acura", where various customers gather together. Therefore, users in "Drag Racing" are more likely to be influenced by friends than that in "Honda/Acura".

- **Latent topic number** T. We then tune the number of latent topics T for T-TFM. Figure 2 illustrates the performance of T-TFM with different choices of T. We find with $T \geq 5$, the performance of T-TFM is competitively better than that with $T = 1$, where T-TFM reduces to B-TFM. This indicates topic discussions is topic-specific. Moreover, it is observed the performance of T-TFM is not sensitive to T. Finally, we empirically fix T as 30 for "Drag Racing" and 40 for "Honda/Acura".
- **Forgetting parameter** τ. We then tune τ in TT-TFM. τ reflects how quickly user's adoption behaviors change as time lapses. Figure 3 illustrates the performance curves of TT-TFM versus various values of τ, where Δt_d in Eqn. (10) is measured in days. We observe small values of τ such as 0.01 achieve good performance for



(a) P@10 vs. τ (b) AP vs. τ

Figure 3: Performance of TT-TFM vs. parameter τ.

"Honda/Acura". Similar observation is found in "Drag Racing". This indicates time lapse factor has just a slight influence on user's participation in topic discussion. It may be due to the regularity of users in discussions of somewhat unchanged topics in a short term (one month).

Comparison Results

Based on the exploration of parameter settings in the previous subsection, we have fixed $\beta = 0.3$, $T = 30$ for "Drag Racing" and $\beta = 0.1$, $T = 40$ for "Honda/Acura", and fixed $\tau = 0.01$ for both communities. We validate the models on the held-out data and present the evaluation results in Table 2. Comparison is conducted on the three proposed models (B-TFM, T-TFM and TT-TFM) and two baselines, i.e., random ranking of users (Random) and ranking of users by the total number of posts (PostNum). For Random, we repeat the procedure 100 times and record average results.

From Table 2, we have several observations: (i) Our T-TFM or TT-TFM achieves the best performance, compared to B-TFM and the baselines Random and PostNum. This indicates topic discussions in online forums is topic-specific and user's participation behaviors change over time. (ii) Generally, prediction over *active users* achieves better performance than that over *all users*, it is because *active users* show relatively more regularity. (iii) In most cases, the performance on "Drag Racing" is better than "Honda/Acura". This reflects the fact that most users of "Drag Racing" are fans of car racing and they show much regularity, whereas various customers are gathering in "Honda/Acura" and their behaviors show relative randomness.

Related Work

To the best of our knowledge, no previous study has systematically considered the problem of modeling dynamic multi-topic discussions in online forums, especially from the perspective of information flow. However, there are two lines of closely related work that we will review in this section.

Online Forums

Recent study about online forums focuses on applications such as question-answer services (Cong et al. 2008) and context-based search (Seo, Croft, and Smith 2009). Mining the regular user behaviors and the mechanisms underlying collective dynamics (Kaltenbrunner, Gonzalez-Bailon, and Banchs 2009) is another new trend. Shi et al. (Shi et al. 2009) observed that users' community joining behaviors display some regularities, and the weak relationships between users defined by replies have similar influence as those of real friendships or co-authorships in (Backstrom et al. 2006). We similarly reason a user's participation behavior is influenced by her linked friends as well as her preferences. On the other hand, the relationships between users in most online forums exhibit relative randomness and less commitment of structural relationships (Shi et al. 2009). In analysis of the underlying networks, researchers usually make use of the reply relationships among users that are based on shared preferences or different opinions (Goh et al. 2006; Gómez, Kaltenbrunner, and López 2008). In this paper, we also leverage the reply relationships likewise to construct the underlying influential network.

Information Propagation

Information flow in networks has received a great deal of attention in recent years. The extensive studies of diffusion of innovations by sociologists (Rogers 1995; Strang and Soule 1998) can shed insights into information diffusion in social networks, since the role of *word of mouth* is essential in both of the two processes of diffusion. Individuals in networks usually influence each other directly or indirectly. The behavior of one's adoption of innovation or information can "spread" through the network (Backstrom et al. 2006), that is, one adopts innovation or information by following her friends or neighbors that are early adopters. We attempt to model online discussion in objective way by following the idea of information flow (Song et al. 2006), with the most influential factors of user's participation considered.

Conclusions and Future Work

This paper focuses on the problem of modeling dynamic multi-topic discussions in online forums. We argue that a user's participation in topic discussion is motivated by either her friends or her own interests. We thus mine the influential social network connecting users as well as user's preferences associated to the discussions of different topics. By following the idea of information flow, we propose our Topic Flow Models for topic discussions. In order to measure how likely a user prefers to join in discussion of a topic, we propose an algorithm called ParticipationRank. Experimental results show our models can predict the tendency of user's participation in topic discussion accurately, especially when latent topics and time lapse factor are considered.

Since the users interact with each other through well-organized forum structure, the understanding of forum pages should be beneficial to users' behavior analysis. In the future, we plan to explore the page layout structure (Cai et al. 2004) for better user behavior modeling. Building recommendation systems directly based on the information flow created by discussions is another interesting future direction.

Acknowledgement

We thank the reviewers for their helpful comments. This work was supported by China National Key Technology R&D Program (2008BAH26B00 & 2007BAH11B06).

References

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *SIGKDD*, 44–54.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems* 30(1-7):107–117.

Cai, D.; He, X.; Ma, W.-Y.; Wen, J.-R.; and Zhang, H. 2004. Organizing www images based on the analysis of page layout and web link structure. In *ICME*, 113–116.

Cai, D.; Wang, X.; and He, X. 2009. Probabilistic dyadic data analysis with local and global consistency. In *ICML*, 105–112.

Cong, G.; Wang, L.; Lin, C.; Song, Y.; and Sun, Y. 2008. Finding question-answer pairs from online forums. In *SIGIR*, 467–474.

Goh, K.; Eom, Y.; Jeong, H.; Kahng, B.; and Kim, D. 2006. Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions. *Physical Review E* 73(6):66123.

Gómez, V.; Kaltenbrunner, A.; and López, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *WWW*, 645–654.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*, 50–57.

Kaltenbrunner, A.; Gonzalez-Bailon, S.; and Banchs, R. E. 2009. Communities on the web: Mechanisms underlying the emergence of online discussion networks. In *WebSci'09: Society On-Line*.

Langville, A., and Meyer, C. 2004. Deeper inside pagerank. *Internet Mathematics* 1(3):335–380.

Lovasz, L. 1993. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty* 2(1):1–46.

Rogers, E. 1995. *Diffusion of innovations*. Free Press.

Seo, J.; Croft, W.; and Smith, D. 2009. Online community search using thread structure. In *CIKM*, 1907–1910.

Shi, X.; Zhu, J.; Cai, R.; and Zhang, L. 2009. User grouping behavior in online forums. In *SIGKDD*, 777–786.

Song, X.; Tseng, B.; Lin, C.; and Sun, M. 2006. Personalized recommendation driven by information flow. In *SIGIR*, 509–516.

Strang, D., and Soule, S. 1998. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual review of sociology* 24(1):265–290.

Tang, J.; Sun, J.; Wang, C.; and Yang, Z. 2009. Social influence analysis in large-scale networks. In *SIGKDD*, 807–816.