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 denote the frequency of a user u replied by u in a thread document d. Besides, let C be the number of comments posted by u 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 $G = (, E)$ to model the underlying influential network (social network), where is the set of nodes with a size $| | = n$, and E is the set of $n \times n$ edges. Each node $v \in$ represents a user u , and each directed weighted edge $e \in E$ represents the influential relationships of u to u in adoption of topics. Let $\mathbf{W}$ be the $n \times n$ affinity matrix where each entry w denotes the edge weight of e , 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 $G = (, E)$ 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 , e , w , $\mathbf{W}$ 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 represents how likely a user u 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 $\mathbf{y}$ denote the preferences of users in discussion associated to a latent topic z.
- **ParticipationRank:** To measure the importance and willingness of a user in participati0(t)1(o)-6(p)-6(i4(c)y)Tj32s ngn ef-1(

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

$$w = \sum_{\in} R \quad (1)$$

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

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

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 of transition to an adjacent node, and a probability $1 -$ of jumping to any node on the graph uniformly at random. By introducing the term $(1 -)\mathbf{N}$, 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 that exclude the comments in response to any other users, the number of which is C in a thread document d. We define each element of the user preference vector $\mathbf{y} = [y_1, \cdots, y_n]^\top$ as

$$y = \sum_{\in} C \quad (3)$$

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

$$q = y \Big/ \sum_{=1}^{n} y \quad (4)$$

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

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

where $\mathbf{p}_{( )}$ is the ParticipationRank vector in t-th iteration, and 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 $\mathbf{p}_{( )}$ will converge to $\mathbf{p}^*$. Substituting $\mathbf{p}^*$ for $\mathbf{p}_{(+1)}$ and $\mathbf{p}_{( )}$, we have

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

Following some algebraic steps, we can finally obtain

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

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 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 $\mathbf{W}$ corresponding to z as follows:

$$w = \sum_{\in} P(z|d)R \quad (8)$$

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

$$y = \sum_{\in} P(z|d)C \quad (9)$$

The computations of $\mathbf{S}$, $\mathbf{q}$ and $\mathbf{p}^*$ 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 $\{\mathbf{p}^*\}_{\in}$ to measure the willingness of each user to participate in discussion of each latent topic z.

Note that in two extreme cases: when $= 1$, and when the probabilities of latent topics in each document $\{P(z|d)\}_{\in}$ 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., $\mathbf{W}$ in Eqn. (8) as:

$$w = \sum_{\in} \exp(- \cdot \Delta t)P(z|d)R \quad (10)$$

where Δt 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 is a forgetting parameter that reflects how quickly user's adoption behaviors change over time. We deal with the user preference vector $\mathbf{y}$ in Eqn. (9) likewise:

$$y \ = \sum_{\in} \ \exp(- \ \cdot \Delta t \ )P(z|d)C \qquad (11)$$

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

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

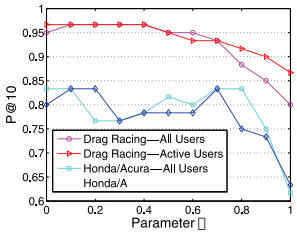
Experimental Results

In this section, we first describe the details of our datasets.

NoWeeceual5(vyr)-1(e)-291di-279(t)1(h)a-2891ow-291te

Table 2: Performance of all methods: = 0.3 (B-TFM parameter), = 30 (T-TFM parameter) for “Drag Racing” and = 0.1, = 40 for “Honda/Acura”, $= 0\ 01$ (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**



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 dis-

play80hlay80fular us4 j-(e)-2(s-6(ed)-271(S)2(m)-95e)-351(m)-5(e)-31r sateds