A Unified Fusion Framework for Time-Related Rank in Threaded Discussion Communities

Qiang You(✉), Weiming Hu, Ou Wu, and Haiqiang Zuo

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
{qyou,wmhu,wuou,hqzuo}@nlpr.ia.ac.cn

Abstract. We propose a unified fusion framework for time-related rank, applied to find valuable posts or recommend answers in threaded discussion communities. In our model, we simultaneously consider the special structure and semantics of threaded discussion communities. As for the structure, we construct a time-related rank model with respect to reply posts analysis and attain an initial rank result. Concurrently, we reconstruct semantic trees from raw statistical features (e.g. term frequency and document length) to latent semantics and topics. With a more robust similarity computation, we produce several semantic trees. For each tree, we again compute the time-related rank score and get a series of rank results. Finally, we fuse our results in the unified fusion framework incorporating quality measures to make a final decision. Our model can be easily extended when new features or models are added. Experimental results show that our model contributes satisfactory results.

Keywords: Fusion · Quality measure · Time-related rank · Information retrieval

1 Introduction

With the Internet growing prosperously, increasing web users would like to share their hobbies, experiences etc. on the Internet. As a result, many kinds of threaded discussion communities are booming and play a more and more important role in content contribution for the web. Benefit from the openness of many threaded discussion communities, we can access the content and get the reply structure of each thread without much difficulty. Unlike the World Wide Web with huge and heterogeneous information, a threaded discussion community always keeps its eye on one or a few domains, which provides a concentrated source to us to access information for the specific domain. Providing the hotness and the character for a stable information dispatcher, mining the threaded discussion communities and gaining valuable posts or users become a urgent task. Unlike traditional mining tasks which conduct knowledge discovery on normalized data in database, the threaded discussion mining aims to find valuable information in non-normalized posts which are proposed by the web users constantly.

W.-C. Peng et al. (Eds.): PAKDD 2014 Workshops, LNAI 8643, pp. 1–12, 2014.
DOI: 10.1007/978-3-319-13186-3_46

Researchers have heavily counted on the vector space model such as term frequency (TF) to represent a document, which is based on the hypothesis the document is represented as an unordered collection of words, neglecting grammar and word order. To summarize and extract the main idea of a corpus with many related documents (always, the corpus is modeled as a matrix called the term document matrix), a series models are adopted to choose and weight the terms in the corpus. The latent semantic indexing (LSI) [1] is a widespread method in information retrieval to find the relations between terms and concepts by transforming the vector space to a new orthogonal space, which behaves effective in many applications (e.g. [2,3]). The latent Dirichlet allocation (LDA) [4] is a generative probabilistic model for collections of discrete data, which assumes that each document is a mixture of several topics and that each word's creation is attributable to one of the document's topics. LDA is a three-level hierarchical Bayesian model, where a topic is draw from the multinomial distribution conjugated with a Dirichlet distribution prior, and each word is draw from a multinomial probability conditioned on the topic.

Those models we mentioned above try to understand the meaning of the text corpora only from one perspective of the document content. While on the web, especially user generated content (UGC) web, there exists rich meta data besides the content, such as time stamp when the user posts a message or even reply structure that shows who replies whom. The threaded discussion community is a typical kind of those webs with rich structure information. Providing the more extra structure information then a bag of discrete text data, we can rank posts according to their value or recommend answer to the given question.

Classical structure methods like PageRank [5] or HITS [6] have achieved great success in information retrieval. However, they are not quite suitable for the threaded discussions without explicit link structures. What is more, the threaded discussion community always varies instantly, where the users may produce many new posts even at one minute, which is not suitable to PageRank because it is liable to the stability of the whole web.

In this paper, we propose a time-related rank model which both considers the time stamps of each posts and the reply-to structure of the discussion thread. With semantic reconstruction, we easily fuse the content to the proposed rank model. The main contributions of this paper are summarized as follows:

- We construct a unified framework to fuse different models incorporating quality measures. Our framework is carried out in two steps. First we construct the quality measure model, then we use the result as a priori to the fusion model. Posit that the threaded discussions in one community are in the same knowledge domain, we choose a subset of threaded discussions to evaluate the quality of the models. Then we popularize the result to the whole threaded discussions in the same community.
- We propose a time-related rank model to alleviate the influence of the time factor when rank different posts with different time stamps, which is difficult for classical models such as PageRank to handle.

– We propose a method to reconstruct the structure of the posts in a thread according to their semantics. Thus, the structure and semantics of a thread with many posts can be easily adopted in our unified fusion framework.

The rest paper is organized as follows. Section 2 introduces the related work. Section 3 briefly introduces the characteristics of threaded discussion communities. Section 4 prepares the needed work including the time-related rank model and the semantic reconstruction. Section 5 gives a careful description of our unified framework which combines several semantic models together based on quality measures. Section 6 provides a thorough set of experiments on two real data sets collected from the apple discussion forum[1] and *Slashdot.org*[2]. We conclude the paper in Sect. 7.

2 Related Work

To the best of our knowledge, little previous research studies the time-related rank in threaded discussion communities. However, there are still a lot of work related to the threaded discussions mining, which can be mainly categorized into semantic models and structure models. As for the semantic models, with information extraction, [7] aimed at ranking answers for given questions in web forums. References [8,9] reconstruct the relationship among posts and threads based on the similarity of topics and semantics. Previous structure models such as PageRank [5] and HITS [6] are under the assumption that the whole web is stable in general, while the threaded discussion communities are not. FGrank [10] modifies PageRank to suitable to the forum pages by constructing page level link graph based on the topic hierarchy without considering the reply-to graph of the posts. Other than the separated models, [11] proposes a sparse coding approach to simultaneously modeling semantics and structure of threaded discussions. It uses the reply-to graph as ground truth and justifies the reply reconstruction of the post by content similarity. However, we believe that the reply-to relationships of the posts should fuse with the content to get a better result.

3 The Problem Setting

Recent years, with the development of the World Wide Web, a lot of UGC web communities have been arising. The threaded discussion community is a typical UGC web community that doesn't emphasize the function of social communication like Facebook or Twitter but supply a place to solve problems or share profound insights, such as mailing list, BBS or Q&A forums. The typical structure of a threaded discussion community is shown in Fig. 1. A threaded discussion community is constituted by a lot of threads in the same knowledge domain. The process of the development of a threaded discussion community can be treated

[1] https://discussions.apple.com/community/ipad

[2] https://slashdot.org

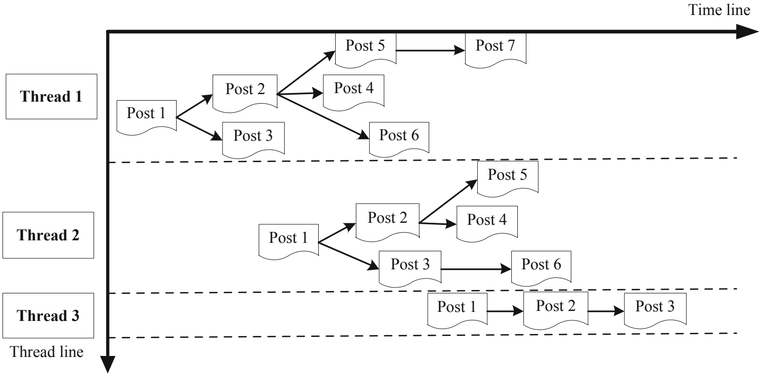


Fig. 1. A typical structure of a threaded discussion community

as the threads' creation and development, which is described as follows. First, one user releases the first post, which attracts a few users to discuss. Then, they propose posts one by one until a consensus is reached. With the creation of more and more threads, the threaded discussion community becomes mature. There are so many posts in a thread and so many threads in a threaded discussion community. Which posts are more valuable or have more insight then others and should be recommended to the other users? That is the main problem we should solve in this paper.

We assume that a threaded discussion community $\mathcal{C}$ is constituted by N threads, and the i-th thread D_i is represented as a directed graph $G_i(V, E, \mathbf{ts})$. The node $v \in V$ is associated with a post which can be modeled by TFIDF, LSI or LDA. There is a time stamp $ts_v \in \mathbf{ts}$ which stands for the moment when the post v is proposed. The edge $(u \rightarrow v) \in E$ exists between two node u, v if post u replies post v. Supposing that there is a metric function f mapping the post v to its value $f(v)$, we get the rank result just according to this value. PageRank is one of this metric function in ranking web pages according to their reputation. In threaded discussion communities, inspired by PageRank, we propose a time-related rank model which is more suitable for our problem.

4 The Preparation Work

Before the unified fusion process, we introduce the time-related rank model. Through semantic reconstruction with several existing vector space models, we easily fuse the content analysis in the rank model.

4.1 The Time-Related Rank Model

The rank model should consider three important factors in a threaded discussion with many posts if the post ranks high. (1) The post should be released *timely* in a thread. (2) The post should attract discussion posts *as many as possible.* With large amount of discussion, the post becomes focused and should also be recommended to the other users. (3) The post with many replied posts which should reply *immediately.* A post that attracts many users to discuss immediately shows the post is active in a short-term response. In conclusion, a post that is timely released and with large posts replying immediately should rank high.

Given a thread D represented by a directed graph $G(V, E, \mathbf{ts})$, we construct the model as follows. The weighted matrix W is calculated with the element $w(u, v) = K(ts_u, ts_v)$. We define a function $h(v) = H(ts_v)$ to depict the timeliness of post v in the thread. We treat the time-related rank (trr) score calculation of each node as an iterative procedure. In step t, the trr score of node v

$$trr^{(t)}(v) = h(v) \sum_{(u \to v) \in E} \frac{trr^{(t-1)}(u)}{w(u, v)} \tag{1}$$

We repeat the iterative procedure until divergence, and rank the posts in a thread with respect to the trr score.

4.2 Semantic Reconstruction

It is hard to analyze the semantics of each post individually because the post released by the users is short and sparse, which means the post itself has incomplete semantics and misses a large part of background knowledge. We reconstruct a semantic tree based on one vector space model from a thread with many posts where each node represents a post and near neighbors have similar semantics. Thus, the post is not individual in semantics with the help that the neighbors provide the context information in the semantic tree.

Given a thread D with m posts $\{L_i\}_{i=1}^m$, their time stamps $\{ts_i\}_{i=1}^m$ where $ts_i < ts_j$ if $i < j$ and the similarity measure function $S(L_i, L_j)$, we reconstruct the semantic tree through the following method. In our similarity computation, we define the similarity measure function as the weighted sum of two parts. The first part is a cosine similarity, and the second part is a similarity between two posts with respect to the post length. The parameter λ here weights the two parts.

$$S(L_i, L_j) = \lambda \frac{L_i L_j + \|L_i\| \, \|L_j\|}{2\|L_i\| \, \|L_j\|} + (1 - \lambda) \frac{2 \, \|L_i\| \, \|L_j\|}{\|L_i\|^2 + \|L_j\|^2} \tag{2}$$

As for post L_j, we choose one post as its predecessor from the ahead posts. The predecessor should have the most similarity with L_j.

$$L_* = \arg_{L_i} \max_{1 \le i \le j-1} S(L_i, L_j) \tag{3}$$

Let j decrement from m to 2, then the semantic tree is reconstructed.

After semantic reconstruction, we again calculate the *trr* score just as the reply structure analysis in a thread. As for the semantic tree and the reply structure graph, it is easy to tackle whatever “combine then rank” or “rank then combine”.

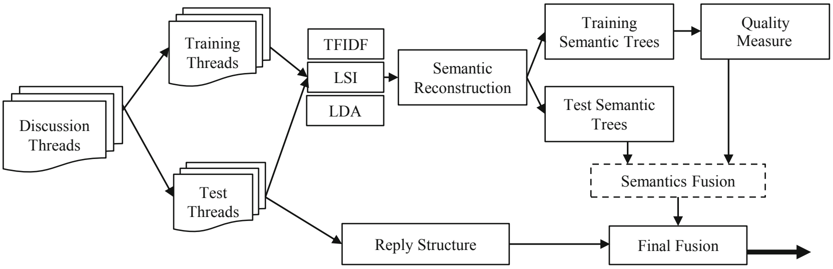


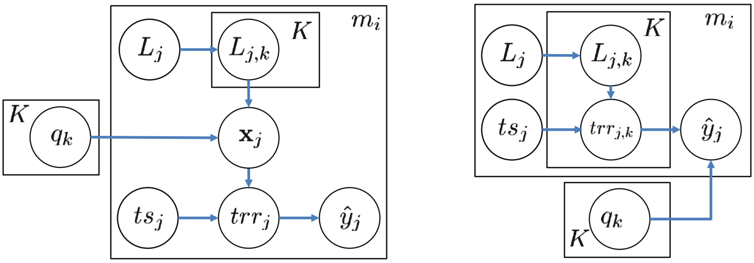
Fig. 2. The unified fusion framework based on the quality measures

5 The Unified Fusion Framework

In the introduction section, we have simply listed three vector space models: TFIDF, LSI and LDA. The first model is directed and contains many details of a document, while it may also include junks. The second model is a way to get the concepts or latent classes of a document, which is totally from the matrix decomposition of TFIDF, but may remove noise. LDA is a generative topic model which is widely used to cluster words with similar semantics. No one model can handle all the data sets. In our unified framework, we first construct a quality measure model to test how much the model is suitable for subset threads of the threaded discussion. The quality measure model is based on the assumption that in the same threaded discussion community, the same knowledge domain is adopted. For example, in the apple discussion forum, people are talking about the apple products. Second, we populate the quality of each model to the other threads. As Fig. 2 shows, our framework consists three main parts: semantic reconstruction, quality measure and fusion procedure. The first part has been described. Now we study the next two parts.

5.1 The Quality Measure for the Semantic Models

Our quality measure model is based on the assumption that all the discussion threads are in the same knowledge domain, which makes our model very suitable for the web communities that concentrate on a few central issues. Central discussions can produce profound insights easier than talking too many issues simultaneously. We randomly choose N_t threads from the whole threads. For the i-th thread D_i, there are m_i posts $\{L_j, ts_j, y_j\}_{j=1}^{m_i}$ in it where L_j is the content of the j-th post in thread D_i, ts_j is the time stamp of the post and y_j the label which stands for the rank or the score that the other users give. As for each model, there is a quality factor measuring the contribution to our result. Suppose that there are K semantic models. The quality vector is $\mathbf{q} = (q_1, .., q_k, .., q_K)$. With post L_j and its time stamp ts_j, after combining several semantic models, we get the output $\hat{y}_j = f(L_j, ts_j, \mathbf{q})$. Our quality measure becomes to solve the following optimization problem.



(a) Combine then rank (b) Rank then combine

Fig. 3. The two strategies of quality measure for semantic models

$$\mathbf{q}_* = \arg\min_{\mathbf{q}} \sum_{i=1}^{N_t} \sum_{j=1}^{m_i} (\hat{y}_j - y_j)^2 \tag{4}$$

There are two strategies to handle the fusion of semantic models. They are *combine then rank* and *rank then combine.*

Combine then Rank. The strategy is a pre-combination which combines the models of the post with the quality factors, reconstructs one semantic tree and calculates trr score at last. As Fig. 3a shows, we obtain the fusion representation $\mathbf{x}_j$ of the post L_j by merging different representations into a new vector.

$$\mathbf{x}_j = (q_1 L_{j,1}, .., q_k L_{j,k}, .., q_K L_{j,K}) \tag{5}$$

Our semantic reconstruction is based on the new fusion representation. Given the reconstructed semantic tree, we calculate the trr score $trr(\{\mathbf{x}_j, ts_j\})$ and then translate it into output $\hat{y}_j = T\left(trr(\{\mathbf{x}_j, ts_j\})\right)$. The translation function $T(.)$ maps the trr score to rank or the mark the other users give.

Rank then Combine. The strategy first reconstructs each semantic tree on each model, then calculates trr score respectively, finally combines the trr scores in one score with the quality vector. As Fig. 3b shows, the output is

$$\hat{y}_j = T\left(\sum_{k=1}^{K} q_k trr(\{L_{j,k}, ts_k\})\right) \tag{6}$$

5.2 The Final Fusion

The reply structure of a thread and the semantics of the posts in the thread should be both considered in our problem. The posts those with much value and should be recommended to the other users must have at least two characteristics. (1) The posts should be ranked high with respect to the trr score in the reply structures, which suggests that the posts have much value in the eye of the users who have read the posts and actively participated in the discussion. (2) The posts should be ranked high in the semantic tree. The semantic tree is based on the similarity measures between posts. Those posts which are ranked high in the semantic tree suggest that they are more similar to the thoughts of the other users.

In the framework of the time-related rank model, we can easily combine the results of the two important factors:

$$trr = \alpha trr_{st} + (1 - \alpha) trr_{se} \tag{7}$$

where trr_{st} represents the trr score from the reply structure of the thread, and trr_{se} stands for the trr score based on the semantic reconstruction. The parameter α can be acquired by training the subset threads used in the quality measure. Then we rank each post in the thread according to the trr score.

6 Experiments

We collect two kinds of data sets over a period of time by a web crawler designed for the threaded discussion communities. One is from the iPad Q&A board in the apple discussion forum, the other is from the technique community *Slashdot.org*. These two data sets are chosen because of the following reasons: (1) The two data sets are from two kinds of typical threaded discussion communities. One is the Q&A forum, and the other is an open discussion forum where everyone can participate and judge the comments. Both of them have time stamps in each post, and the reply structure can be extracted without much difficulty. (2) These two data sets are all or at least partial labeled. The iPad Q&A data set can label the answers "Helpful" by other users or "Solved" by the questioner, while *Slashdot.org* can give each comment a score ranging from −1 to 5 by all the participators. The quality vector **q** and the weight α in final fusion are acquired by supervised learning, which relies on the labeled data. For each data

Table 1. The basic statistics of the data sets

Data set	iPad Q&A	Slashdot.org
Number of threads	1130	664
Number of posts	8489	146569
Number of users	2175	14241
Average thread length	7.51	220.74
Average words per post	63.09	76.33
Average posts per user	3.90	10.29
Timestamp(mins from 1970)	21075992 - 21590652	22091472 - 22633163
Number of topics	5	5

sets, we select 5 hottest topics and ignore the unqualified threads that have posts fewer then 3 or without labels or ratings. The basic statistic results are shown in Table 1, from which we know that the two kinds of threaded discussion communities are quite different in average thread length, users active degree and so on. However, proving the similarity in content and structure organization, we can get the valuable answers to the questions or recommend the popular comments in our unified fusion framework.

6.1 Evaluation for the Time-Related Rank Model

Our model is different from the previous studies largely because we consider the time stamp which represents the timeliness of the post in a thread. There are two time intervals considered in our trr model. One is how long the post has stayed on the webpage until now, the other is between the post and its reply posts. Let us take iPad Q&A data set as an example. As shown in Fig. 4, every post belongs to one thread and has a time stamp that represents the released time. Figure 4a shows that the distribution of the posts in each thread in the time line. Every post is either unlabeled or labeled with "Helpful" or "Solved". Figure 4b shows the time intervals between post and its reply posts follow the power law distribution. The most of the time intervals between reply posts are less than a few hours. When the time interval becomes large, the number quickly decreases.

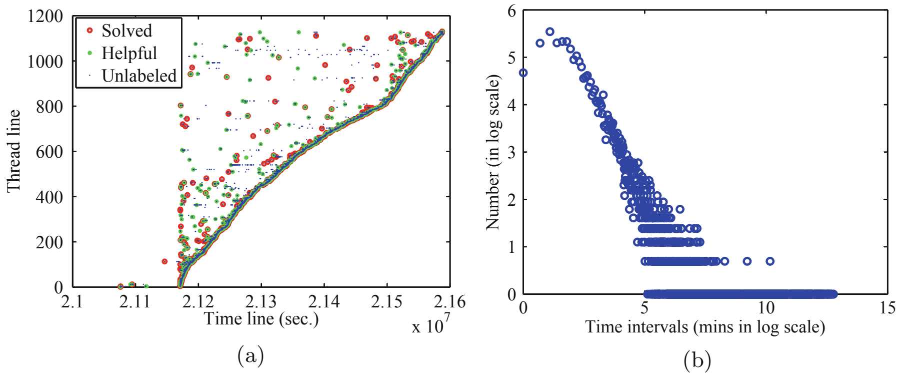


Fig. 4. Timeliness in iPad Q&A data set

In the experiment with the trr model, we define the time interval function $K(ts_u, ts_v) = \log(ts_u - ts_v)$ between post v and its reply post u. The timeliness of post v is $h(v) = H(ts_v)$, which can be calculated as

$$H(ts_c, ts_v) = \exp\left(-\frac{ts_v - ts_{min}}{ts_{max} - ts_{min}}\right) \tag{8}$$

As for each thread, ts_{max} is the time stamp of the timeliest post and ts_{min} is the latest.

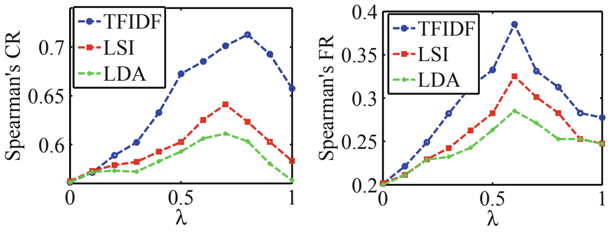
Table 2. The rank evaluation results

Data set	iPad Q&A				Slashdot.org			
Criteria	TS(ASC)	TS(DESC)	PR	TRR	TS(ASC)	TS(DESC)	PR	TRR
Spearman's RC	0.8449	0.5859	0.7856	0.8378	0.3247	0.1108	0.4648	0.5119
Spearman's FR	0.8418	0.5882	0.7824	0.8351	0.3212	0.1156	0.4664	0.5111

In the iPad Q&A data set, the post in a thread can only be one of three labels, "Unlabeled", "Helpful" or "Solved". We rank them 3, 2, 1 respectively. While in *Slashdot.org*, the posts those are marked a score from -1 to 5 are directly ranked from 7 to 1. We rank the posts in a thread according to four criteria:(1) time stamp in ascending order (2) time stamp in descending order (3) PageRank score (4) trr score, as shown in Table 2. Mallows model with Spearman's rank correlation and footrule [12] is introduced to measure the results. The iPad Q&A data set is from a forum that many senior iPad users or even service staff answer the questions timely. As a result, when the question is released by a green hand, it can be solved the first time, which is shown in the table that rank the time stamp in ascending order performs best. Our trr model performs much better then PageRank because we take the timely release and timely reply into consideration. While on the dataset from Slashdot.org, our trr model performs best.

6.2 The Semantic Reconstruction Experiments

We select three models TFIDF, LSI and LDA to conduct the semantic reconstruction. In the similarity calculation (see Eq. 2), the parameter λ balances the angle and the length of two post vectors. We carry out an experiment to choose the best λ for three models on each data set. As shown in Fig. 5, we choose λ for the three models $\lambda_{iPad} = (0.8, 0.7, 0.7)$ on iPad Q&A data set, and $\lambda_{Slashdot} = (0.6, 0.6, 0.6)$ on *Slashdot.org*.



(a) λ for iPad Q&A

(b) λ for Slashdot.org

Fig. 5. Find the λ for each data set

6.3 Evaluation for the Unified Fusion Framework

The quality measure follows the training paradigm. For each data set, we randomly choose $\gamma = 0.2$ of the whole threads to train. The similarity parameters on each data set in the semantic reconstruction are adopted in our quality measure. The topic numbers that we choose in LSI and LDA are both 5. As shown in Table 3, the quality measure of three semantic models on two strategies is consistent, the quality of TFIDF is the best because the posts are short in threaded discussion communities. With words in a post as much as possible, we can get the semantics of the post much better. While the disadvantage is also obviously, TFIDF is more time-consuming then other two models.

Table 3. The quality measure for the unfied fusion framework

Data set	iPad Q&A			Slashdot.org		
Quality	q_{tfidf}	q_{lsi}	q_{lda}	q_{tfidf}	q_{lsi}	q_{lda}
Combine then rank	0.51	0.41	0.08	0.66	0.25	0.09
Rank then combine	0.54	0.38	0.08	0.69	0.22	0.08

The training process also gets the fusion parameter α between semantics and structure incorporating quality measures. Based on all the above parameters we get from the experiments, we get the final fusion time-related rank result on the remaining test data set. As shown in Table 4, our fusion framework which combines both semantics and structure information of the thread performs much better than the model just from structure or semantics in time-related rank.

Table 4. The rank results for the unified fusion framework

Data set	iPad Q&A				Slashdot.org			
Criteria	St	Se(CR)	Se(RC)	Fusion	St	Se(CR)	Se(RC)	Fusion
Spearman's RC	0.8133	0.6228	0.6452	0.8456	0.4121	0.3373	0.3487	0.5521
Spearman's FR	0.8025	0.6182	0.6438	0.8424	0.3351	0.3256	0.3412	0.5414

7 Conclusions

We have described a time-related rank model in our paper, which takes the time stamp of the post into consideration. Based on the assumption that the post which is timely released and with large posts replying immediately should rank high, we have designed an algorithm to alleviate the influence of the time factor when rank different posts with different time stamps. We have also proposed a method to reconstruct the structure of the posts in a thread according to their semantics. Finally we have constructed a unified framework to fuse different models incorporating quality measures. In the unified fusion framework, the structure and semantics of a thread with many posts can be easily adopted. Experiments on two data sets from two kinds of typical threaded discussion communities have demonstrated that our time-related rank model works better than PageRank. The unified fusion framework is also easily extended when new features or models are added.

Acknowledgments. This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the National Science Foundation of China (Grant No. 61379098, No. 61103056), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JASIS **41**(6), 391–407 (1990)
2. Gee, K.R.: Using latent semantic indexing to filter spam. In: Proceedings of the 2003 ACM Symposium on Applied Computing, pp. 460–464. ACM (2003)
3. Baron, J.R.: Law in the age of exabytes: Some further thoughts on 'information inflation' and current issues in e-discovery search. Rich. JL Tech. **17**, 9–16 (2011)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web (1999)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM (JACM) **46**(5), 604–632 (1999)
7. Cong, G., Wang, L., Lin, C.Y., Song, Y.I., Sun, Y.: Finding question-answer pairs from online forums. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 467–474. ACM (2008)
8. Blei, D.M., Moreno, P.J.: Topic segmentation with an aspect hidden markov model. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 343–348. ACM (2001)
9. Shen, D., Yang, Q., Sun, J.T., Chen, Z.: Thread detection in dynamic text message streams. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 35–42. ACM (2006)
10. Xu, G., Ma, W.Y.: Building implicit links from content for forum search. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 300–307. ACM (2006)
11. Lin, C., Yang, J.M., Cai, R., Wang, X.J., Wang, W.: Simultaneously modeling semantics and structure of threaded discussions: a sparse coding approach and its applications. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 131–138. ACM (2009)
12. Spearman, C.: The proof and measurement of association between two things. The Am. J. Psychol. **15**(1), 72–101 (1904)