User Comments for News Recommendation in Social Media

Jia Wang
Southwestern Univ. of Finance and Economics
55 Guanghua Cun Road
Chengdu, China
wangjia@2008.swufe.edu.cn

Qing Li*
Southwestern Univ. of Finance and Economics
55 Guanghua Cun Road
Chengdu, China
liq_t@swufe.edu.cn

Yuanzhu Peter Chen
Memorial Univ. of Newfoundland
St. John's, A1B 3X5
NL, Canada
yzchen@mun.ca

ABSTRACT

Reading and Commenting online news is becoming a common user behavior in social media. Discussion in the form of comments following news postings can be effectively facilitated if the service provider can recommend articles based on not only the original news itself but also the thread of changing comments. This turns the traditional news recommendation to a "discussion moderator" that can intelligently assist online forums. In this work, we present a framework to recommend relevant information in the forum-based social media using user comments. When incorporating user comments, we consider structural and semantic information carried by them. Experiments indicate that our proposed solutions provide an effective recommendation service.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Web is one of the most important vehicles for "social media", e.g. Internet forums, blogs, wikis, and twitters. One form of social media of particular interest here is self-publishing. In self-publishing, a user can publish an article or post news to share with other users. Other users can read and comment on the posting and these comments can, in turn, be read and commented on. Digg (digg.com) and Yahoo!Buzz (buzz.yahoo.com) are commercial examples of self-publishing. A useful extension of this self-publishing application is to add a recommendation feature to the current discussion thread. That is, based on the original posting and various levels of comments, the system can provide a set of relevant articles, which are expected to be of interest of the active users of the thread.

Here, we explore the problem of news recommendation for dynamic discussion threads. A fundamental challenge in adaptive news recommendation is to account for topic divergence, i.e. the change of gist during the process of discussion. In a forum, the original news is typically followed by other readers' opinions, in the form of comments. Concerns and intention of active users may change as the discussion continues. Therefore, news recommendation, if it were only based on the original posting, can not benefit the potentially changing interests of the users. Apparently, there is a need to consider topic evolution in adaptive news recommendation and this requires novel techniques that can help to capture topic evolution precisely to prevent wild topic shifting which returns completely irrelevant news to users. A related problem is content-based information filtering (or recommendation). Most information recommender systems select articles based solely on the contents of the original postings [1] [3] [4].

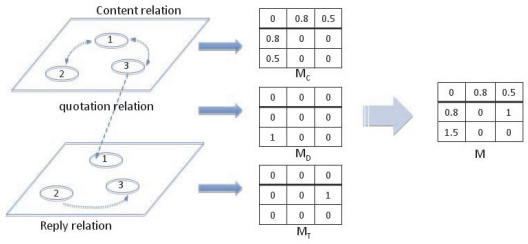

Figure 1: Multi-relation graph of comments

In this work, we propose a framework of adaptive news recommendation in social media. It has the following contributions. (1) It is the first attempt of incorporating reader comments for adaptive news recommendation. (2) We model the relationship among comments and that relative to the original posting in order to evaluate their overall impact on recommendations.

2. SYSTEM DESIGN

The proposed news recommender first constructs a topic profile for each news posting along with the comments from readers, and uses this profile to retrieve relevant news.

We first model the relationship among comments and that relative to the original posting in order to evaluate their overall impact. In our model, we treat the original posting and the comments each as a text node. This model both considers the *content similarity* between text nodes and the *logic relationship* among them. On one hand, the content similarity between two nodes can be measured by any commonly adopted metric, such as cosine similarity and Jaccard coefficient. This metric is taken over every node pair in the discussion thread. On the other hand, the logic relation between nodes takes two forms. First, a comment is always made in response to the original posting or an earlier comment. In graph theoretic terms, the hierarchy can be represented as a tree $T = (V, E_T)$, where V is the set of all text nodes and E_T is the edge set. In particular, the original posting is the root and all the comments are ordinary nodes.

*This research is supported by National Natural Science Foundation of China Grant No.60803106.

Copyright is held by the author/owner(s).
SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

There is a directed edge $e \in E_T$ from node u to node v, denoted (u, v), if the corresponding comment v is made in response to comment (or original posting) u. Second, a comment can quote from one or more earlier comments. From this perspective, the hierarchy can be modeled using a directed acyclic graph (DAG), denoted $D = (V, E_D)$. There is a directed edge $e \in E_D$ from node u to node v, denoted (u, v), if the corresponding comment v quotes from comment (or original posting) u. As shown in Figure 1, for either graph T or D, we can use a $|V| \times |V|$ adjacency matrix, denoted M_T and M_D, respectively, to record them. Inline with the adjacency matrices, we can also use a $|V| \times |V|$ matrix defined on $[0, 1]$ to record the content similarity between nodes and denote it by M_C. Thus, we can combine these three aspects linearly.

Intuitively, the important comments are those whose topics are discussed by a large number of other important comments. Therefore, we propose to apply the PageRank algorithm [2] to rank the comments as

$$s_j = \lambda/|V| + (1 - \lambda) \times \sum_{c_i} r(c_i, c_j) \times s_i,$$

where λ is the damping factor as in PageRank and this value is recommended to be 0.85, i and j are node indices, and $|V|$ denotes the number of text nodes in the thread. In addition, $r(c_i, c_j)$ is the normalized weight of comment c_i referring to c_j defined as

$$r(c_i, c_j) = \frac{M_{i,j}}{\sum_{c_j} M_{i,j} + \epsilon},$$

where $M_{i,j}$ is an entry in the graph adjacency matrix and ϵ is a constant to avoid division by zero.

Once the importance of comments on one news posting is quantified by our model, this information along with the news itself are fed into a synthesizer to construct a topic profile of this news discussion thread. The profile is a weight vector of terms to model the language used in the thread. Consider a news posting d_0 and its comment sequence $\{d_1, d_2, \cdots, d_m\}$. For each term t, a compound weight $W(t)$ is calculated. It is a linear combination of the contribution by the news posting itself, $W_1(t)$, and that by the comments, $W_2(t)$. The weight contributed by the news itself, d_0, is:

$$W_1(t) = w(t, d_0)/\max_{t'} w(t', d_0)$$

The weight contribution from the comments $\{d_1, d_2, \cdots, d_m\}$ incorporates not only the language features of these documents but also their importance of leading a discussion in related topics. That is, the contribution of comment score is incorporated into weight calculation of the words in a text node.

$$W_2(t) = \sum_{i=1}^{m} w(t, d_i)/\max_{t'} w(t', d_i) \times s_i/\max_{i'} s_{i'}$$

Such a treatment of compounded weight $W(t)$ is essentially to recognize that readers' impact on selecting relevant news and the difference of their influence strength.

With the topic profile constructed as above, we can use it to select relevant news for recommendations. That is, the retriever returns an order list of news with decreasing relevance to the topic. Our model to differentiate the importance of each comment can be easily incorporated into any good retrieval model. In this work, our retrieval model is derived from [4].

3. EXPERIMENTAL EVALUATION

To gauge how well the proposed recommendation approach performs, we carry out a series of experiments on a synthetic data set collected from Digg and Reuters news website. We randomly select 20 news articles with corresponding reader comments from Digg website. These news articles with different topics are treated as the original news postings, recommended news are selected from a corpus of articles collected from Reuters news website. This simulates the scenario of recommending relevant news from traditional media to social media readers for their further reading. We compared the proposed approach to three other retrieval approaches as the baseline: one is a simple content filter (CF) which treats news and comments as a single topic profile, the other two are well-known news recommendation methods [1], Okapi and LM.

Table 1: Overall performance

	The Proposed	CF	Okapi	LM
$P@10$	0.94	0.789	0.827	0.804
MAP	0.932	0.8	0.833	0.833

To observe the impact of readers' concerns on original news posting in social media, we investigate the effect of the three forms of relationship among comments, i.e. content similarity, reply, and quotation. We carry out a series of experiments for this purpose. we find that replies are slightly more effective than quotations and both of these outperform pure content similarity. In other words, the importance of comments can be well evaluated by the logic organization of these comments. We also notice that the incorporation of content similarity decreases the system effectiveness. This may seem to contradict our intuition that the textual information should complement the logic-based models. By further investigating our results, we find that content similarity sometimes misleads the decision on the importance of the comments. Besides, the computation cost of calculating the content similarity matrix M_C is very high. Therefore, we only apply the structural information to determine the importance of each comment.

We have t-tests using $P@10$ and MAP as performance measures, respectively, and the p values of these tests are all less than 0.05, which means that the results of experiments are statistically significant. We conduct a series of preliminary experiments to find the optimal performance obtained when the topic file word number is 60 and combination coefficient α is 0.7. As shown in Table 1, the overall performance of the proposed approach performed significantly better than the best baseline methods.

4. CONCLUSION

In this work, we present a framework for adaptive news recommendation that incorporates information from the entire discussion thread. This study can be extended in a few interesting ways. For example, we can use this technique to process personal Web blogs and email archives. The technique itself can also be extended by incorporating such information as reader scores on comments, chronological information of comments, and reputation of users. Indeed, its power is yet to be further improved and investigated.

5. REFERENCES

[1] T. Bogers and A. Bosch. Comparing and evaluating information retrieval algorithms for news recommendation. In *Proc. of ACM Recommender systems*, 2007.

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

[3] J.-H. Chiang and Y.-C. Chen. An intelligent news recommender agent for filtering and categorizing large volumes of text corpus. *International Journal of Intelligent Systems*, 19(3):201–216, 2004.

[4] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In *Proc. of CIKM*, 2000.