Winter 12-2-2007

A Synthetical Approach for Blog Recommendation Mechanism: Trust, Social Relation, and Semantic Analysis

Yung-Ming Li

Ching-Wen Chen

Proceedings of The Seventh International Conference on Electronic Business, Taipei, Taiwan, December 2-6, 2007, pp. 29-34.

A SYNTHETICAL APPROACH FOR BLOG RECOMMENDATION MECHANISM: TRUST, SOCIAL RELATION, AND SEMANTIC ANALYSIS

Yung-Ming Li, National Chiao Tung University, Taiwan, yml@mail.nctu.edu.tw
Ching-Wen Chen, National Chiao Tung University, Taiwan, kisd0955@hotmail.com

ABSTRACT

Weblog is a good paradigm of online social network which constitutes web-based regularly updated journals with reverse chronological sequences of dated entries, usually with blogrolls on the sidebars, allowing bloggers link to favorite site which they are frequently visited. In this study we propose an elaborate blog recommendation mechanism that combines trust model, social relation and semantic analysis and illustrate how it can be applied to a prestigious online blogging system – Wretch in Taiwan. By preliminary results of experimental study, we found some implications and empirically prove some theories in domain of social networking, and the example reveals that the proposed recommendation mechanism is quite feasible and promising.

Keywords: Blogosphere, Trust model, Social networking, IR, Back-propagation neural network.

INTRODUCTION

Online social networking systems and peer-produced services have gained much attention as a social medium of viral marketing, which exploits existing social networks by inspiring bloggers to share their own posts or personal information with the other bloggers. The weblogs indeed offers a more open channel of communication, people in the blogosphere read, commentate, cite, socialize and even reach out beyond their social networks, make new connections, and form communities [8]. A blog social network has emerged as a powerful and potentially services-valued form of computer-mediated communication (CMC). There exists a large number of information in the blogosphere, including text-based blog entries (articles) and profile, pictures or figures and multimedia resources. However, the problems of information overload still bother users, which can be tackled with by recommender system and information filtering approach (such as search engine technology).

A recommender system of weblog differs from with the others in nature. First, recommend target varies dramatically from product, movie, music, news, webpage, travel and tourism to all kinds of service, online auction seller and even virtual community. It is important for us to figure out the characteristics of recommend targets because the inappropriate use of recommendation may have a totally opposite effect by resulting unfavorable attitudes towards the recommend target.

Under blog recommendation context, it is particularly important that how we introduce some interesting, personalized and socially related weblogs of these peer-produced information to bloggers. The objective of blog recommendation mechanism in this study is bloggers or blog posts (articles). Then what kinds of blog posts do we recommend? Most popular, most similar in links or in semantic of blog network and content, or most trustworthy? These approaches and related researches will be described later, which inspire us to combine them to propose a recommendation mechanism in this study. We consider that trust model, social relation and semantic similarity play important roles in trust recommender system, social networking analysis and information retrieval/textual comparison, respectively and they are three crucial factors to help prepare the ground for the development of personalized and trustworthy recommendation mechanism.

The rest of paper is organized as follows. Section 2 presents related works. Section 3 designs a system framework of neural network based recommendation mechanism. Section 4 elaborates on methodologies of trust model, social relation and semantic analysis. Section 5 concludes the paper.

LITERATURE REVIEW

A fast-growing number of blog studies have showed that blog as social network can help researchers in understanding and analyzing certain implications and insights. It generated several issues which received lots of attention in several aspects. The concept of blog ranking is similar to the concept of blog recommend to some extent and like the process of search engine. [4] score each blog entry by weighting the hub and authority scores of the bloggers based on eigenvector calculations, which has similarities to PageRank [3] and HITS [7] in that all are based on eigenvector calculation of the adjacency matrix of the links. They contribute some dimensions to calculate the importance of webpage or blog. However, the work in [9] ranks blogs according to their similarity in social behaviors by graph-based link analysis, which demonstrates an excellent paradigm of link analysis. Note that there is an inherent problem of sparseness in the blogosphere which has already been noticed by many researchers, works in [1][9] have coped with it by extending and increasing explicit and implicit links based on various blog aspects where a denser graph will result in a better performance of ranking and recommending. Equally, in order to solve the sparsity problem, the extracted communities in [10] only cover a portion of the entire blogosphere, the ranking method extract dense subgraphs from highly-ranked blogs.

Recommend this kind of peer-produced services or objects needs not only social network-based link analysis but also the concept of trustworthiness and reliability of weblogs must take into consideration. A recommender in blog network may have similar social relationships or contents to a target user (i.e. recommendation service requester) but they may not be a reliable

predictor for inducing the recommendation. Using trust in recommender system will improve the ability of making accurate recommendation [13], which can solve a portion of weaknesses of traditional content-based, collaborative filtering (CF)-based recommendation approaches. The work in [5], trust takes on the role of a recommender system to create predictive rating recommendations for movies. And the accuracy of the trust-based predicted ratings is significantly better than the other approaches to a movie. As to sparsity problem, [14] proposes a trust-based method that is based on trust inferences, which relaxes the sparsity and the cold-start problems. Accordingly, our approach constructs a trust network by friend relationships where trust is mean to deal with these issues.

Additionally, the recommend mechanism is applied to the weblog's graph, which is a generalization of the post's graph. The blog posts are strongly representative and we can discover the preferences and writing pattern of bloggers who we want to recommend to. Traditional information retrieval (IR) technology is applied to handle the semantic of blog content. In examining the semantic similarity among weblogs, CKIP Chinese word segmentation system [11] helps us to parse and stem the crawled blog posts in this study. Index terms are highlighted through IR/NLP approaches. Many syntax-based and semantics-based approaches exist to analyze the textual relationships among blogs [15]. In [2], they proposed two methods for semantics-enhanced blogs analysis that allow the analyst to integrate domain-specific as well as general background knowledge. And the iRank in [1] acts on implicit link structure to find those blogs that initiate these epidemics, which denote similarity between nodes in content and out-links. Undoubtedly, the content of blog post is also an important source that we must take into consideration for inducing recommendation seriously.

In this paper, we focus on the issues of combined trust model, social relation analysis and semantic similarity as a means of recommending bloggers or blog posts. And the neural network is deployed to learn and capture the pattern of preferences of blog users and it is utilized to predict the final recommend score of each blog post in our recommendation network.

BLOG RECOMMENDATION MECHANISM

In this study, we propose an innovative recommendation mechanism on the blogosphere which employs the trust model, social relation and semantic analysis to construct a more comprehensive and more personalized framework for each bloggers on the entire blogspace. There are various important factors and dimensions we must take into consideration under blog recommendation context. Especially we conclude three underlying critical aspects of blogosphere- Trustworthiness and Reliability, social intimacy and popularity and semantic similarity, which contribute to TR, SIP and SS scores respectively. We present a neural network-based approach to learn and predict user's preference and affinity i.e. Final Recommend Score (FRS) of each blogger and blog post, by feeding these standardized scores into neural model. Figure 1 is the architecture of the proposed NN-based recommendation mechanism.

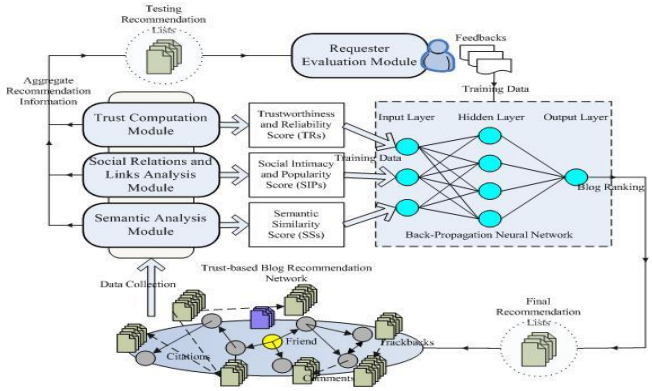

Figure 1. The architecture of the proposed NN-based recommendation mechanism

RESEARCH METHODOLOGIES

This study proposes a neural network-based blog recommendation mechanism in which we apply the concepts of trust model, social relation and semantic analysis and they contain the information of the blog network about trustworthiness and reliability, Social intimacy and popularity and Semantic similarity respectively. This information is then integrated as an initial recommend score for each objects over the recommendation network, then the initial recommend list was induced for requesters. Meanwhile, a back-propagation neural network (BPNN) is proposed to forecast the FRS according to these scores. Finally, a recommend list of blog posts or bloggers is generated for the recommendation service requester. The whole process of recommendation mechanism is divided into several steps as the figure 2 and is described as the following sub-sections.

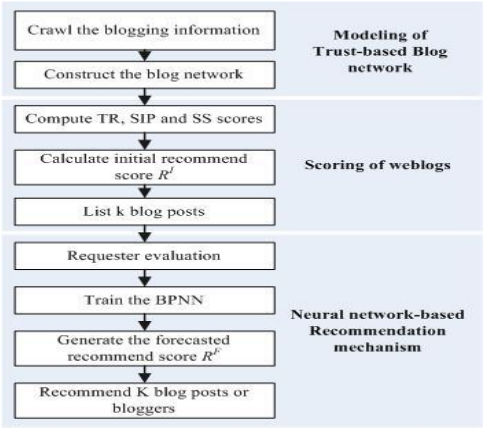

Figure 2. The whole process of recommendation mechanism and it's sub-sections

Modeling of Trust-based Blog Network

Crawl the Blogging Information. First of all, we take blogsite of requester as a starting point to search available and social-reachable agents i.e. recommenders, by performing search algorithm according to blogrolls on the side bar in the blogsite of each agent. These agents are connected level-by-level by friend or friend-of relationships in the blog network. Once the agents are decided and specified or the maximum number of searching level is reach, the members of the recommender are confirmed. Then we crawl blogging information (such as blog posts, hyperlinks, comments…etc) associate with each agent on the recommendation network.

Construct the Blog Network. To implement and evaluate the proposed model, we simulate a trust-based blog network in which apply the concepts of agent and object in [4]. In this graph-based representation blog network, m agents (bloggers) and n objects (blog posts) are denoted as nodes and document-like icons, respectively. The relation edges in the network denote heterogeneous and multiplicity of links (whether explicit or implicit links), that is, it depends on the directions and entities involved here. Note that the constructed blog network forms and extends from the requester (node in yellow), then the trust information could be propagated and inferred in the agent layer. Then the scope of object layer will be determined by these objects which can be reached by these agents in the agent layer.

Scoring of Weblogs

Calculate Initial Recommend Score R^I and List K Blog Posts. We compute a recommend score (either posts or bloggers) according to their scores of trustworthy, social relation and semantic similarity after a min-max standardization approach, which is applied to each scores (showed by upper case in eq(1)). An initial recommend list is generated with a sequence of recommend score from high to low. Recommend scores $R(i, j)$ for each post *j* of blogger *i* for given the requester *r* is defined as following,

$$R^I(r, o_{ij}) = \alpha TR^s(r, i) + \beta SIP^s(r, o_{ij}) + \gamma SS^s(r, o_{ij}) \qquad (1)$$

where uppercase *I* of recommend score R^I stands for initial recommend score and uppercase *s* of TR, SIP and SS scores mean scores after the process of standardization. Parameters α, β and γ are the self-set weights of trust score, social relation score and semantic score of objects in the blog network respectively and the values are between 0 and 1.

Then the initial recommend list was induced, which contains top k R^I score and ranges from highest R^I score to lowest one for requester for further evaluation process. Each scoring approach is presented in the following three sub-sections.

Compute TR, SIP and SS Scores:

Trust Scores (TR)

The interpersonal trust values derive directly from blogrolls is the TR scores in this study. All agents assign trust value to his/her friends listed in the blogroll on homepage of blog site. The computation of TR scores is divided into two steps: First, for a given requester (also blogger) *r*, we collect and aggregate trust information then form the trust-based blog network of him/her for further inference and filtering. Second, a search algorithm is applied to the constructed blog network in the former step, and set a maximum search layer as stopping criteria. The aim of this step is to find out social-reachable and available agents from the given requester who is the root of the blog network. These agents form the recommender set $RC(r)$ of requester *r*. The TR score of agent *s* is computed by trust inference mechanism which is the most widely used one in trust-based social networking computing approach [6]. $TR(r, s) = t_{rs}$ and,

$$t_{rs} = \frac{\sum_{j \in adj(r)} t_{rj} \times t_{js}}{\sum_{j \in adj(r)} t_{rj}} \qquad (2)$$

where

r is the requester of blog recommendation, *s* stands for these social-reachable and available agents, and $s \in RC(r)$. t_{rs} is the value of trust degree from agent *r* to *s*, and $t_{is} \in [0,1]$. $adj(r)$ means adjacent agents of agent *r*, i.e. friends of blogger *r*.

Social Relation Scores (SIP)

This section measures social intimacy and population (SIP) score of each agent in the blog network via their interrelationships and shared properties. To combine a complete view in recommendation process, SIP score is separated into SI and Popularity scores: SI addresses the social similarity strength or the degree of familiar on agent-agent aspect. However, Popularity emphasizes global reputation on object aspect. SIP score is introduced in the following formula,

$$SIP(r, o_{ij}) = SI(r, i) + Popularity(o_{ij}) \quad (3)$$

where $SIP(r, o_{ij})$ measures the scores of every object or agent in blog network given a requester agent *r* as a basis for comparison and computation. $SI(r, i)$ and $Popularity(o_{ij})$ represents social intimacy relation and popularity scores respectively.

Social intimacy captures the idea of social similarity by examining the degree of interaction between agents and mutual behaviors (links) toward certain blogs or websites.

$$SI(r, i) = sim(iL(r, A), iL(i, A)) + sim(oL(r, A), oL(i, A)) \quad (4)$$

where *r, i* stands for the requester of blog recommendation (source agent) and certain agent respectively, and $r, i \in A$. A denotes a set of agents (or websites) which are social-reachable and available agents, i.e. agents (websites) which can be reached by links (hyperlinks) or inferences mechanism. $iL(r, A)$ is a vector which simply counts the number of social links from r to each of the agents in set A, where social links in here denote out-degree link which actually includes the situations of co-citation, co-comment and mutual link between the agents. $sim(\cdot)$ is the function to compute the similarity between two agents by inner product calculation. Contrast to out-degree aspect, the latter part of formula measures the in-degree link which includes the situations of comments (citations) contributed (cited) by same author (blog post). However, $oL(r, A)$ counts the number of social links from agent set A to agent *r*.

Popularity measures social importance of an agent in blog network. We measure the in-degree (the number of incoming links) in this case as a rough substitute for popularity for the ease of computing. Since an object *u* belonging to an agent *s*, we compute the aggregate value of *u* as a weight sum of the relative number of comments and citations as following,

$$Popularity(o_{ij}) = w_{co} \times \frac{Comment(o_{ij})}{\max Comment(A)} + w_{ci} \times \frac{Citation(o_{ij})}{\max Citation(A)} \quad (5)$$

Where $Comment(o_{ij})$ ($Citation(o_{ij})$) counts the number of comments (citations) in object *j* of agent *i*. And $\max Comment(A)$ ($\max Citation(A)$) is the maximum number of comments (citations) in our dataset. Obviously, the popularity score of an agent *i*, $Popularity(i)$, is the sum of popularity score of objects belonging to *i*. The parameters w_{co} and w_{ci} are the weights of in-degree links from comment and citation behaviors respectively.

Semantic Scores (SS)

Once the blogging data is crawled, we apply CKIP (Chinese Knowledge and Information Processing) Chinese word segmentation system [11] to parse the content of blog post after the HTML tags are removed. We extract the syntactical functions which we need (normally nouns and besides we select several kinds of verbs) for the process of stop word removal. Then the remaining words are the index terms. After that, a basic cosine similarity metric of term vectors with standard TFIDF [12] weighting scheme is deployed to represent each index term of each blog article. Semantic score measures textual similarity between blog posts of requester and the posts of the other bloggers in the given blog network (once the blog network is constructed). Suppose there are n agents (bloggers) in the blog network, semantic score is an agent-to-object score which is defined as blow,

$$SS(r, o_{ij}) = sim(q, d_{ij}),$$

$$\text{Where } i \in [1, n], \quad j > 0 \text{ and } 0 \le SS(r, o_{ij}) \le 1 \quad (6)$$

where *q* stands for index terms of blog postings which were published by requester r and we deem it as a query. Note that *q* could be generated by selecting any subset of objects of agent *r*. The variable d_{ij} is a vector of the TFIDF scores of index terms of blog post *j* of agent *i*.

Neural Network-based Recommendation Mechanism

In this study, neural network is used to capture the non-linear relationships between these factors, and requester's preferences and social status in blog social network accurately in a comprehensive view to forecast the FRS for each object or agent.

Requester Evaluation. Once the initial recommend list of *k* blog posts (bloggers) is delivered to requester, which accompany with a detail description of philosophy behind the recommendation processes by a web-based interface. For the requester, all he/she has to do is review these posts (bloggers) and make a unbiased evaluation by scoring each posts (bloggers) selected according to his/her own preference based on the degree of perceptibly relatedness and similarity with respect to himself/herself.

Train the BPNN. The characteristics, preference and social behaviors vary dramatically among human beings. Neural network-based recommendation mechanism is special for its leaning and forecasting abilities to imply the implicit relationships behind these factors and requester's pattern of preference. Notably, a forecasted score for each object will be obtained and the weights of initial recommend score with respect to three scores will be learned (i.e. weighting variables α, β and γ for TR, SIP and SS scores, respectively) through the neural network. To train the back-propagation neural network, we combine three scores i.e. TR, SIP and SS, and the results from requester evaluation process as testing data for BPNN. Once the network is trained, it can be used to *Generate the Forecasted Recommend Score* R^F and then generate *Recommend List of K Blog Posts or Bloggers* to the requester.

EXPERIMENT STUDY

So far we have introduced that combine trust model, social relation and semantic analysis and they will be crucial factors to guarantee high-quality recommendations in blog network. In this section, we want to highlight how promising this recommendation framework is and how much it will satisfy the users by utilizing the blogging data from Wretch, which is a famous blog system in Taiwan, to show the entire recommendation processes. We begin by explaining how the dataset was collected. Then some statistical data will be presented such as the number of bloggers in the recommendation network, average number of friends of bloggers and of blog posts for each blogger.

Data Descriptions

We describe our proposed mechanism by using a dataset collected from the Wretch [17] which is a Taiwanese community website. It is the most famous weblog community in Taiwan with millions of users registered now where users can upload photos to album, write the blog and interact with others by these services [16].

In early July 2007, we start crawl related blogging information including blogger account, friend relations, article id, article content (object), citations, comments and publish datetime for each blogger by using the crawler we designed, once the recommendation network is constructed. Note that, the objects are crawled according to the agents which have been crawled.

The detail statistics information of this experimental recommendation network is presented in table 1 and 2. We can observe that the network size is drastically increasing, and we can predict that the network will achieve saturated situation when the network spread up to 5~6 layer. That is, the network will be close to the entire blog network of Wretch (i.e. about 2.5 millions+ users).

Table 1. Statistics of recommendation network (up to 3rd layer).

Characteristics of recommendation network	Statistics
# of agent (blogger) in the network	22,336
# of object (blog post) in the network	338,614
Average # of friend of an agent	29.722
Average # of objects of an agent	15.160
Average # of comments of an object	2.382
Average # of citations of an object	0.084

Table 2. The # of agent and friend relationship in each layer according to the root: "chiang1000".

# / layer	root	1st layer	2nd layer	3rd layer	4th layer
The # of agent	1	23	927	21,384	299,539
The # of friend relationship	23	972	30,299	632,389	NA

So, an experimental small recommendation network about 20,000+ agents and 330,000+ objects will be construct and limited the layer to 3rd layer, due to the reasons that the network size grows up exponentially with the layer increased. And accompany with the results that computability of trust and semantic similarity will greatly decreased.

For entire network, about 57.22% of objects are isolated and without any comment and citation. From statistics, we found that 99% of the objects have comments range from 0 to 15, 80% range from 0 to 2, however 57.4% of objects do not have any comments. Moreover, 99% of the objects do not have any citations. Because of the sparse nature of blogosphere we have mentioned before, our approach seek to increase the density of the implicit links between bloggers and between blog posts. This enhances the reliability and comprehensiveness of recommendation mechanism.

Notably, the recommendation network in this study is formed according to the requester's friend network i.e. trust network. In other words, we fetch the users, who are reachable walking the network of trust of starting requester, into our dataset. We conduct our experiments with our target requester and some of her friends listed in the blogroll who are all the bloggers of Wretch for providing recommendation information and evaluating the effectiveness of the proposed recommendation mechanism and BPNN in this study, i.e. trust value, blog posts and feedback data for training the neural network.

CONCLUSIONS

In this study we propose an elaborate recommendation mechanism that combines trust model, social relation and semantic analysis and illustrate how it can be applied to a prestigious online blogging system – Wretch in Taiwan. By preliminary results of experimental study, we found some implications and empirically prove some theories in domain of social networking, and the example reveals that the proposed recommendation mechanism is quite feasible and promising.

As to future work, we will conduct several experiments to evaluate the efficiency and accuracy of proposed recommendation mechanism and illustrate how it can be applied to online blogging system – Wretch in Taiwan. We expect that the proposed neural network-based mechanism will actually help in predicting a more accuracy and personalized recommendation list for the bloggers. Finally we will design an experimental blog system to implement the recommendation mechanism for further analysis and application.

ACKNOWLEDGEMENTS

This research was supported by the National Science Council of Taiwan (Republic of China) under the grant NSC 96-2416-H-009-017.

REFERENCES

[1] Adar, E., Zhang, L., Adamic, L. and Lukose, R. (2004) "Implicit structure and the dynamics of blogspace", *Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, WWW.

[2] Berendt, B. and Navigli, R. (2006) "Finding your way through blogspace: Using semantics for cross-domain blog analysis", *In AAAI Symposium on Computational Approaches to Analyzing Weblogs*.

[3] Brin, S. and Page, L. (1998) "The Anatomy of a large-scale hypertextual web search engine", *proceedings of 7th international World Wide Web Conference*.

[4] Fujimura, K., Inoue, T. and Sugisaki, M. (2005) "The EigenRumer algorithm for ranking blogs", *2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, WWW.

[5] Golbeck, J. and Hendler, J. (2006) "FilmTrust: Movie Recommendations using Trust in web-based Social Networks", *IEEE Consumer Communications and Networking Conference*.

[6] Golbeck, J. (2006) "Trust and nuanced profile similarity in online social networks", *Journal of Artificial Intelligence Research*.

[7] Keinberg, J. M. (1999) "Authoritative sources in hyperlinked environment", *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, Vol. 46, No. 5.

[8] Kolari, P., Finin, T. and Lyons, K. (2007) "On the structure, properties and utility of internal corporate blogs", *ICWSM*.

[9] Kritikopoulos, A., Sideri, M. and Varlamis, I. (2006) "BlogRank: ranking weblogs based on connectivity and similarity features", *Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, pp. 198.

[10] Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J. and Tseng, B. (2006) "Discovery of blog communities based on mutual awareness", *3rd Annual Workshop on the Weblogging Ecosystem*.

[11] Ma, W.-Y. and Chen, K.-J. (2003) "Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff", *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pp. 168-171, online demo system: http://ckipsvr.iis.sinica.edu.tw/

[12] Manning, C.D. and Schutze, H. (1999) Foundations of statistical natural language processing, *MIT Press*, Cambridge, MA.

[13] O'Donovan J. and Smyth, B. (2005) "Trust in recommender systems", *Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 167-174.

[14] Papagelis, M., Plexousakis, D. and Kutsuras, T. (2005) "Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inference", *In Proceedings of iTrust 2005*, pp. 224–239.

[15] Tsai, T.-M., Shih, C.-C. and Chou, S.-C. T. (2006) "Personalized Blog Recommendation: Using the value, semantic, and Social Model", *Innovations in Information Technology*, pp. 1-5.

[16] Wikipedia: http://en.wikipedia.org/wiki/Wretch_%28website%29

[17] Wretch blog: http://www.wretch.cc/blog/