國立交通大學

資訊管理研究所

碩士論文

結合信任網絡，社會關係及語意分析之部落格推薦機制

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

研究生：陳敬文

指導教授：李永銘 博士

中華民國九十七年六月

結合信任網絡，社會關係及語意分析之部落格推薦機制

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

研究生：陳敬文　　Student：Ching-Wen Chen

指導教授：李永銘　　Advisor：Yung-Ming Li

國立交通大學
資訊管理研究所
碩士論文

A Thesis
Submitted to Institute of Information Management
College of Management
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Institute of Information Management

June 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年六月

結合信任網絡，社會關係及語意分析之部落格推薦機制

學生：陳敬文　　　　　　　　　　　　　　　指導教授：李永銘 博士

國立交通大學資訊管理研究所碩士班

摘　　要

部落格為線上社群網路一個很好的範例，它包了以網頁為基礎並時常更新的網路日誌，其日誌依時間前後反轉排列而記載的順序來排列。而旁列的好友清單讓部落客能夠時常拜訪喜愛的部落格。本研究提出一個以結合信任模型、社會網路關係及文字語意分析的部落格推薦機制，並且描述此推薦機制如何應用到台灣－無名小站這個有名的部落格系統上。從實驗結果來看，我們從發展的部落格網路中發現許多的意涵並實證社會網路中重要的理論。實驗評估顯示我們提出的部落格推薦機制相當的可行並且具有潛力的。

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

Student：Ching-Wen Chen Advisors：Dr. Yung-Ming Li

Institute of Information Management
National Chiao Tung University

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 a blog recommendation mechanism that combines trust model, social relation, and semantic analysis and illustrates how it can be applied to a prestigious online blogging system – Wretch in Taiwan. By the results of experimental study, we found a number of implications from the Weblog network and several important theories in domain of social networking were empirically justified. The experimental evaluation reveals that the proposed recommendation mechanism is quite feasible and promising.

誌　　謝

由於你們的協助與關心，不但讓我的論文書寫得以順利，更在人生的課程中讓我學習到真誠與溫暖，我發自內心的想跟你們說聲謝謝。

首先要感謝的是我的老師 李永銘 博士，兩年內的諄諄教誨與分享讓我能夠在研究上成長與茁壯；李氏哲學讓我能夠在人生的視野上更加開闊，進而提高自己的眼界並放眼之後的人生；在論文期間老師的指正與協助讓我能夠順利的完成論文的撰寫並再研究上能夠順利的進行。其次要感謝口試委員林福仁教授、劉敦仁教授與簡宏宇教授能夠在百忙之中抽空來參加論文的口試，其建議與指教使得本論文能夠更臻完善。教授們的分享與指導，即使將來我進入職場之後也會一直銘記在心。

短短的兩年研究所期間結識了學生時期最佳的好友建邦與嘉豪, 兩年的革命情感勝於一切，因為有你們的陪伴與照顧，讓我在研究所期間有了一個很好的研究環境與歸屬感，真誠的感謝你們，讓我無論在研究上與研究生活上充滿了快樂與滿足，這份感激之心我會永遠牢記在心，謝謝你們! 我想感謝博班學長們無尾熊、Denny 與易霖：無尾熊學長是我最佳的良師益友，大方的貢獻與分享讓我在研究生生涯更為充實，放鬆時間的最佳良伴；Denny 學長無私的貢獻來幫我潤稿與論文的指點讓我的論文更加完整，還要感謝 Denny 學長的熱心與分享，讓身為學弟的我充滿感激；易霖學長是我們的大哥哥，你的指導與分享讓我更清楚事物的運作怎樣才會更有效率，真的非常崇拜。各位，沒有你們我該怎麼辦，真的是充滿感激與感恩。

學弟妹們連乃、涵文、小球、宗穎與正乾有你們的陪伴與照顧，讓我在這個階段的回憶更添光彩，謝謝你們，感謝連乃這位同學兼學弟的大方支持與鼓勵、感謝涵文為實驗室帶來更多的歡笑與快樂、感謝小球論文口試上的幫助與運作、感謝宗穎大力的貢獻與協助實驗室的運作、感謝正乾的技術協助與建議，因為有你們讓我的研究所生涯更為多采多姿，充滿回憶。

此外，更有賴許多朋友的精神支持，甚或實質協助，我才能順利完成學業。包含資管所的老師們、淑惠、欣欣、OR 實驗室的亞梅學姊，昱紹、癸堂、網路實驗室的鼎元學長、DB 實驗室的振東，挂一漏萬，可能尚有一些未提及的朋友們，在此均一併致謝。

要感謝的人太多了，感謝每位曾經幫助過我的貴人，因為有你們才有今天的我，在此，衷心的謝謝你們。最後要感謝的是家裡與女朋友千千的支持，讓我在研究上無後顧之憂，謝謝你們。

僅以此論文獻給我最親愛的家人,支持我的女朋友與關心我的師長朋友，感謝你們的包容與愛護，願與你們分享這份榮耀。

陳敬文

2008年6月　謹誌於　新竹　國立交通大學光復校區

Table of Contents

List of Figures and Tables

CHAPTER 1 Introduction

1.1 Background and Motivation

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 provide a more open channel of communication for people in the blogosphere to read, commentate, cite, socialize and even reach out beyond their social networks, make new connections, and form communities [10]. A blog social network has emerged as a powerful and potentially services-valued form of computer-mediated communication (CMC). More and more interactions take place in the blogosphere, combining the benefits of the accessibility of the web, the ease-of-use of interface and the incentive of blogging (i.e. share, recommend, comment…etc.). Blog becomes a viral marketing site based on peer-production and it is promoted yet induced by online person to person interactions. Moreover, there exists a large number of information in the blogosphere, including text-based blog entries (articles) and profile, pictures or figures, and multimedia resources. This becomes problematic for users. How do they deal with information overload problems and how do they effectively retrieve information they consider important? This gives us an incentive to develop a blog recommender approach and design an information filtering mechanism.

A recommender system of weblog differs from others in several ways. First, a recommendation target varies dramatically from product, movie, music, news, webpage, travel and tourism to all kinds of service, online auction seller, or even virtual community [12]. It is important for us to find the characteristics of recommendation targets because inappropriate use of recommendation may have a totally opposite effect by resulting unfavorable attitudes towards the recommendation target. Second, the blog recommender is also a provider. Unlike other contexts, blogs or bloggers in the entire blog network are highly dynamic and the recommendation environment changes fast. The blog recommendation mechanism must be more flexible and adaptable than the others. Third, it is more human-oriented. In other words, blog content itself is highly subjective and textual-sensitive for recommenders.

Blog search engine and blog recommender system serve similar function but differ to some extent. What is the difference between blog search engine and blog recommender system? This question emerges as the blog filtering approach such as search engine can also alleviate the mentioned problem. There are three folds of differences between them. First, information needs: real-time versus long-run. Some weblog aggregators, such as Technorati, provides tag-based search engine platform; moreover, Blogpulse and Daypop supply common keyword-based search engines just like Google and Yahoo but are applied in weblog domain. It allows users to find potential interesting postings, which many bloggers are talking and concerning about recently, with ease. In contrast to search engine technology, the proposed blog recommendation mechanism is long-run oriented. In other words, the former focuses on popularization however the latter is more personalized. Second, pull versus push information: the former is a paradigm of technology for pulling information. A search result is obtained after the query is submitted. As for the latter, either pull or push technology could be employed to induce the recommendation results. Third, diversity of recommendation process: the former only considers the content and term comparison. As for the latter, it considers multidimensional approaches and factors to implement the recommendation mechanism. In this study, the proposed mechanism takes all these three factors into consideration.

Moreover, recommender system is a useful alternative to search algorithms, since they help users discover items they might not have found by themselves. Interestingly enough, recommender systems are often implemented using search engines indexing data. That is, some recommender systems are proposed based on the results of search engine. Since search engine could not provide personalized results according to user's preference, a recommender mechanism will do by integrating more methodologies to make a personalized resource-provided mechanism.

1.2 Research Problems

In blog recommendation context, it is important that how we introduce interesting, personalized and socially related weblogs of this peer-produced information to bloggers through recommendation mechanism. The objective of blog recommendation mechanism in this study is bloggers or blog posts (articles). The problem is that what kinds of blog posts do we recommend? Is it most popular? Is it most trustworthy? Or recommend most similar in links or in blog content? These approaches and related researches inspire us to combine them to propose a synthetical recommendation mechanism in this study. We believe that trust model, social relation and semantic similarity play an important role in trust recommender system, social networking analysis, and information retrieval/textual comparison respectively. They are three crucial factors to help prepare the ground for the development of personalized and trustworthy recommendation mechanism.

1.3 Research Objectives

In this research, we propose a personalized, trustworthy, and adaptive blog recommendation mechanism which integrates the trust model, social relation and, semantic analysis to construct a comprehensive model in recommending bloggers and blog posts. With this mechanism, a blogger could have better opportunities to locate more interested, trustworthy, and related blogging information with greater satisfactions than other existing recommendation approaches. More specifically speaking, we want to provide bloggers with more precise and more desirable blogging information with less efforts and greater satisfactions.

The main objective of this research is to apply the proposed recommendation mechanism to the real-world blog platform and investigate the recommendation performances with an empirical validation. We take a famous BSP (Blog Service Provider), Wretch [26], as our target of experiments, and compare the recommendation performances with existing approaches, to examine if our proposed mechanism outperforms the existing ones.

1.4 Thesis Outline

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 describes the methodologies of trust model, social relation and semantic analysis. Section 5 proposes an experimental study to discover some characteristics of blog network and demonstrates the effectiveness of the proposed recommendation mechanism. Section 6 concludes the paper, discusses the potential problems and some limitations, and describes the future works.

CHPATER 2 Literature Reviews

This chapter reviews literatures related to this research, including social network-based analysis, trust model, textual and semantic-based analysis, and back-propagation neural network.

2.1 Social Relation Analysis and Ranking Mechanism

A fast-growing number of blog studies have shown that blog as social network can help researchers in understanding and analyzing certain implications and insights. It generated several issues and received lots of attention. The works [5][11][1][13] considered social relation-based dimension to measure the importance and relationships of webpage or blog. The concept of blog ranking is similar to that of blog recommendation to some extent. [5] assigns scores to each blog entry by weighting the hub and authority scores of the bloggers based on eigenvector calculations, which has similarities to PageRank [4] and HITS [9] in that all are based on eigenvector calculation of the adjacency matrix of the links. However, the work in [11] 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 researchers. Works in [1][11] have coped with this problem 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. In our data set, only 57.22% of blog posts are isolated and without any comment and citation. The recommendation pool is large enough to perform our mechanism. Equally, in order to solve the sparsity problem, the extracted communities in [13] only cover a portion of the entire blogosphere and the ranking method extract dense subgraphs from highly-ranked blogs.

2.2 Trust Model in Recommendation System

Previous researchers suggested trust as another dimension to strengthen the reliability and robustness of a recommender system. The works [19][6][20] applied trust to reinforce the ability of a recommender system. Recommenders in blog network may have social relationships or contents similar to a target user (i.e. recommendation service requester) but they may not be a reliable predictor for inducing the recommendation. Using trust in a recommender system will improve the ability of making an accurate recommendation [19], which can solve partial weaknesses of traditional content-based or collaborative filtering (CF)-based recommendation approaches. In addition, [6] use trust in a recommender system to create predictive rating recommendations for movies. The accuracy of the trust-based predicted ratings for movies, compared with other approaches, is significantly better. Moreover, [20] proposes a trust-based method based on trust inferences to deal with the sparsity and the cold-start problems. Studies above have shown trust is critical for implementing a recommender system. Accordingly, our approach constructs a trust network by friend relationships.

2.3 Textual and Semantic Analysis

Semantic analysis is also an important dimension to be taken into consideration. The works [3][1] indicated that applying semantic or textual-based analysis in blog domain is suitable and fruitful. Since the blog posts are strongly representative, we can discover the preferences and writing pattern of bloggers whom 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

[14] helps us parse and stem the crawled post contents. Index terms are highlighted through IR/NLP approaches. Many syntax-based and semantics-based approaches are used to analyze the textual relationships among blogs [22]. In [3], two methods are proposed for semantics-enhanced blog analysis that allows the analyst to integrate domain-specific as well as general background knowledge. And the iRank in [1] acts on implicit link structure to find blogs that initiate epidemics, which denote similarity between nodes in content and out-links. Undoubtedly, the content of blog post is an important source to induce recommendation.

2.4 Back-Propagation Neural Network

The researches [23][21] show that with learning ability, applying back propagation neural network to conduct forecast and prediction is appropriate in all kinds of domain. Under a multi-agent or peer-to-peer distributed environment, network is consisted of heterogeneous peers whose trust evaluation or rating standards may differ [23]. The issue is how to accurately and effectively predict trust value of an unknown party from multiple recommendations [21] by BPNN.

The major advantage of neural networks is their flexible nonlinear modeling capability. With ANNs, there is no need to specify a particular model form. Rather, the model is adaptively formed based on the features presented from the data. This data-driven approach is suitable for many empirical data sets where no theoretical guidance is available to suggest an appropriate data generating process [28]. Hence, in blog context, it may also help in deriving final recommendation score for each of the blog post which is expected to satisfy the user's preferences.

2.4.1 Feed-forward networks

A neural network is constructed from a number of interconnected neurons arranged in layers. The outputs of one layer of neurons are connected to the inputs of the following layer. The first layer of neurons is called the “input layer”, since its inputs are connected to external data, for example, sensors to the outside world. The last layer of neurons is called the “output layer”, accordingly, since its outputs are the result of the total neural network and are made available to the outside. All neuron layers between the input layer and the output layer are called “hidden layers” since their actions cannot be observed directly from the outside. If all connections go from the outputs of one layer to the input of the next layer, and there are no connections within the same layer or connections from a later layer back to an earlier layer, then this type of network is called a “feed-forward network”. Feed-forward networks (Figure 2.1) are used for the simplest types of ANNs and differ significantly from feedback networks, which we will be described next.

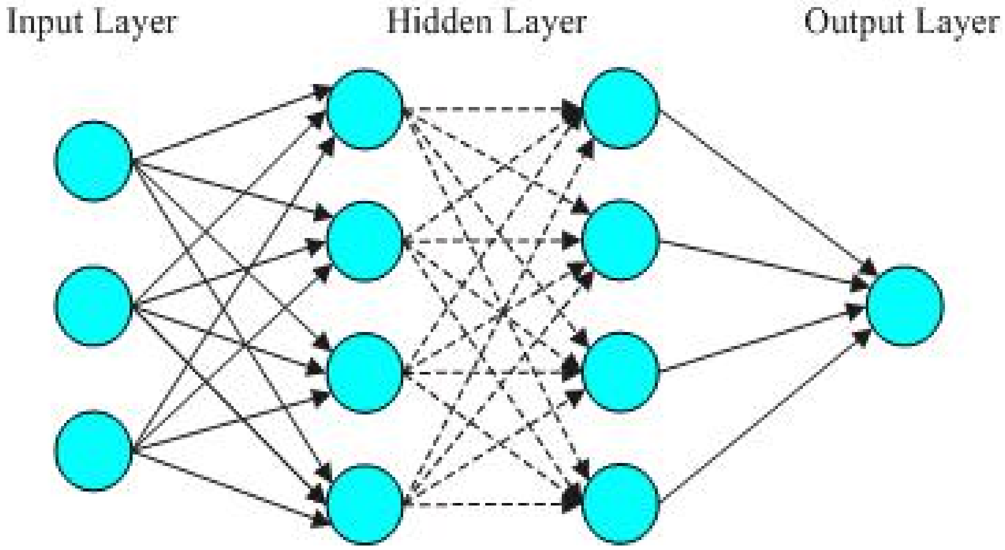


Figure 2.1 Fully connected feed-forward network

2.4.2 BPNN algorithm

BPNN is a famous supervised machine learning artificial intelligence technique. From the practical perspective, BPNN is a non-linear statistical data modeling or decision making tool. It can be used to model complex relationships between inputs and outputs or to find patterns in data [24]. It is essentially a black-box user and rules cannot be easily extracted from it. Still, it has many empirical applications includes financial prediction, decision making, machine learning, etc.

The network computes its output pattern, and examines if there is an error - or in other words a difference between actual and desired output patterns - the weights are adjusted to reduce this error. In a back-propagation neural network, the learning algorithm has two phases. First, a training input pattern is presented to the network input layer. The network propagates the input pattern from layer to layer until the output pattern is generated by the output layer. If this pattern is different from the desired output, an error is calculated and then propagated backwards through the network from the output layer to the input layer. The weights are modified as the error is propagated. Second, once the network is converged, a pattern between input and desired output data is learned. A testing data could be fed into the network to generate the predicted outputs. Then a set of evaluation data could be performed to assess the performance of the model. The following is the process of back-propagation learning algorithm [18]:

Step1: Initialization

Set all the weights and threshold levels of the network to random numbers uniformly distributed inside a small range:

$$(-\frac{2.4}{F_i},+\frac{2.4}{F_i}) \qquad (2.1),$$

where F_i is the total number of inputs of neuron i in the network. The weight initialization is done on a neuron-by-neuron basis.

Step2: Activation

Activate the back-propagation neural network by applying inputs $x_1(p)$, $x_2(p)$,… $x_n(p)$ and desired outputs $y_{d,1}(p)$, $y_{d,2}(p)$,…, $y_{d,n}(p)$.

2.1 Calculate the actual outputs of the neurons in the hidden layer:

$$y_j(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p)\times w_{ij}(p)-\theta_j\right] \qquad (2.2),$$

where n is the number of inputs of neuron j in the hidden layer, and *sigmoid* is the *sigmoid* activation function.

2.2 Calculate the actual outputs of the neurons in the output layer:

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{jk}(p)\times w_{jk}(p)-\theta_k\right] \qquad (2.3),$$

where m is the number of inputs of neuron k in the output layer.

Step3: Weight training

Update the weights in the back-propagation network propagating backward the errors associated with output neurons.

3.1 Calculate the error gradient for the neurons in the output layer:

$$\delta_k(p) = y_k(p)\times[1-y_k(p)]\times[y_{d,k}(p)-y_k(p)] \qquad (2.4),$$

Calculate the weight corrections:

$$\Delta w_{jk}(p) = \alpha \times y_j(p) \times \delta_k(p) \qquad (2.5),$$

Update the weights at the output neurons:

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \qquad (2.5),$$

3.2 Calculate the error gradient for the neurons in the hidden layer:

$$\delta_j(p) = y_j(p) \times [1 - y_i(p)] \times \sum_{k=1}^{l} \delta_k(p) \times w_{jk}(p) \qquad (2.7),$$

Calculate the weight corrections:

$$\Delta w_{ij}(p) = \alpha \times x_i(p) \times \delta_j(p) \qquad (2.8),$$

Update the weights at the hidden neurons:

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \qquad (2.9),$$

Step4: Iteration

Increase iteration p by one, go back to Step 2 and repeat the process until the selected error criterion is satisfied.

2.5 Comparisons with Prior Literatures

In this paper, we combine trust model, social relation analysis, and semantic similarity to recommend bloggers or blog posts. As neural network is able to learn and capture the pattern of preferences of blog users, it is utilized to predict the final recommendation score of each blog post in our recommendation network. As summarized in Table 2.1, we compare the existing literatures and this research with respect to the factors included. We believe this research takes a more comprehensive methodology into consideration.

Table 2.1 Related literatures and methodologies after the year 2004.

Authors	**Domains and Issues**	**Trust Model**	**Social relation (link analysis)**	**Content (textual analysis)**	**BPNN**
Kritikopoulos. A., et al. (2006)	Ranking blogs		●		
Adar, E., et al. (2004)	Analysis of structural and links of blogs		●		
Glbeck, J., et al. (2006)	Using trust in SN-based recommender system	●			
O'Donovan, J., et al. (2005)	Trust in recommender system	●			
Berendt, B., et al. (2006)	Analysis of Semantics in blogs			●	
Weihua, S., et al. (2004)	Trust model + NN learning	●			●
Song, W., et al. (2005)	Trust recommendation in p2p network	●			●
Tsai, T.-M., et al. (2006)	Blog recommendation		●	●	
This research (2008)	Blog recommendation: Trust model, social relation and semantics	●	●	●	●

CHPATER 3 The Framework

3.1 Blog Recommendation Mechanisms

In this study, we propose an innovative weblog recommendation mechanism in the blogosphere which employs the trust model, social relation, and semantic analysis to construct a more comprehensive and personalized framework for each blogger on the entire blogspace. There are various important factors and dimensions we must take into consideration in blog recommendation context. We employ three underlying critical aspects of blogosphere: Trustworthiness and Reliability (TR), Social Intimacy and Popularity (SIP) and Semantic Similarity (SS). Moreover, we present a neural network-based approach to learn and predict user's preference and affinity. By feeding these standardized scores into neural model, the Final Recommendation Score (FRS) of each blogger and blog post can be learned. Figure 3.1 depicts the architecture of the proposed NN-based recommendation mechanism.

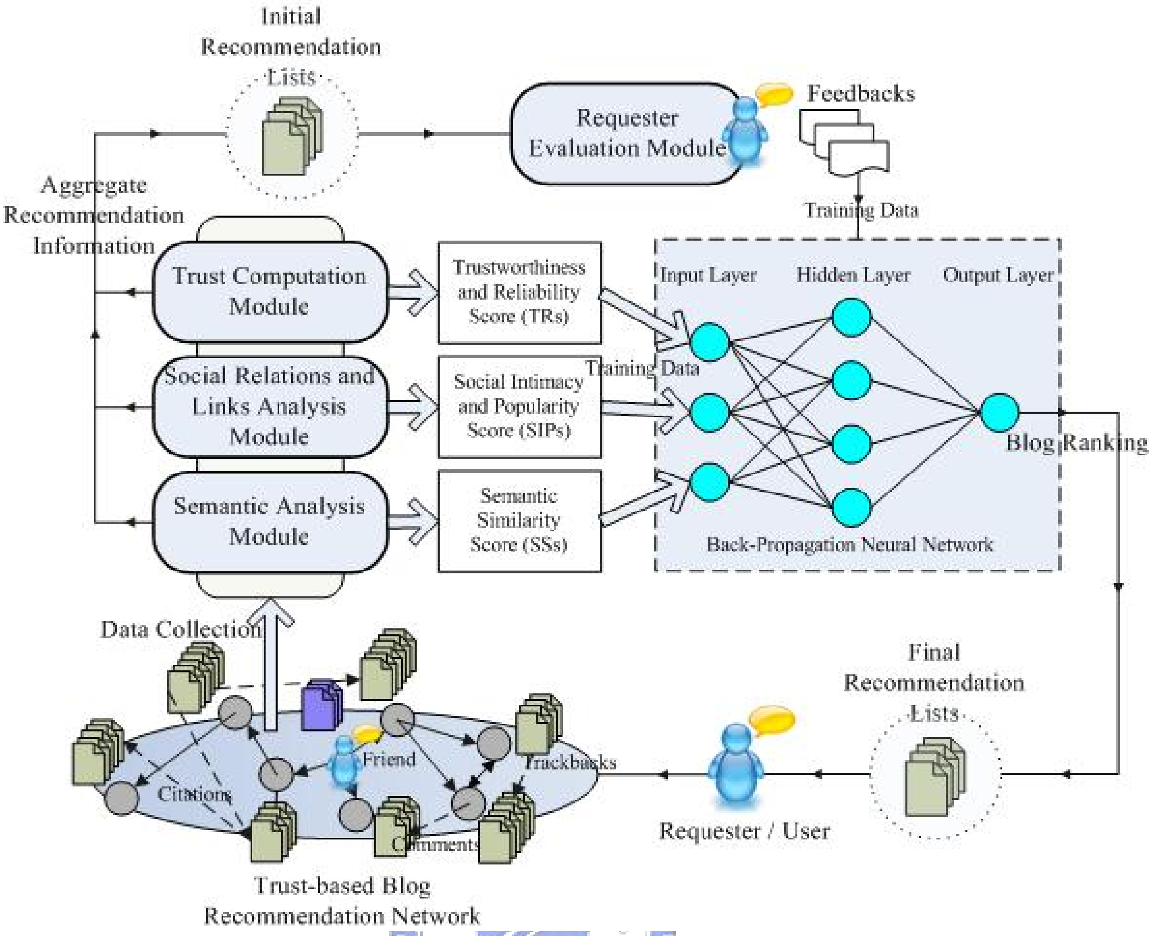

Initial Recommendation Lists
Requester Evaluation Module
Feedbacks
Training Data
Aggregate Recommendation Information
Trust Computation Module
Trustworthiness and Reliability Score (TRs)
Social Relations and Links Analysis Module
Social Intimacy and Popularity Score (SIPs)
Semantic Analysis Module
Semantic Similarity Score (SSs)
Training Data
Input Layer
Hidden Layer
Output Layer
Blog Ranking
Back-Propagation Neural Network
Data Collection
Friend
Trackbacks
Citations
Comments
Trust-based Blog Recommendation Network
Final Recommendation Lists
Requester / User

Figure 3.1 The architecture of the proposed NN-based recommendation mechanism

3.2 Trustworthiness and Reliability (TR)

Little existing literature studies trust issue among bloggers while it is widely used in social networking and distributed computing environment. By definition, trust degree in the social network connotes: belief and commitment. That is, as we said “A trusts B”, it stands for that A haves a belief in B who will provide good opinions or behaviors in the future, and A is willing to accept it. In this study, we use the similar definition. A directed trust degree between bloggers A and B is a hybrid of *referral trust* and *content-provision trust,* a suitable representation of social relation between these bloggers in the blog network.

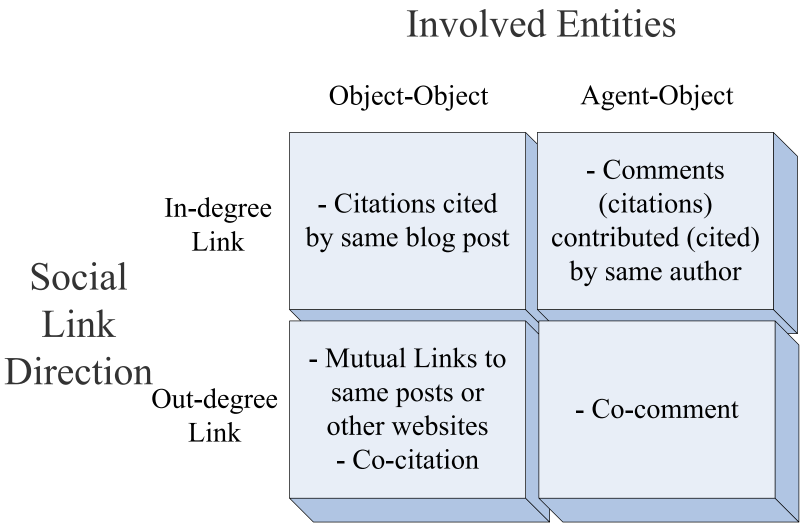
In the context of blog recommendation, every blogger is a potential provider. But in other domains, a specific query could not be satisfied by all agents but by resource providers.

However, we denote it as the concept of "resource specificity" for the reason that every blogger in blog network could be a recommender or recommendee for a given query and every available blogger could provide the contents (objects). In addition, users rely on the familiars' recommendation to retrieve certain information or services, and recommendation is more trustworthy and reliable because they have similar affinities and preference.

3.3 Social Intimacy and Popularity (SIP)

SIP involves either explicit or implicit links which are used to represent the relationships and interconnections of the social network-based graphs. In the context of blog linkage analysis: nodes stand for bloggers or article posts, and edges represent social or similarity relations between bloggers (article posts), which generally contain four social behaviors: comment, blogroll, citation, and trackback (see [2] for more detail definitions). Moreover, similarity relationship is also a key factor. They reflect the social intimacy and interest relatedness between bloggers (posts) in the blog network.

The intrinsic sparsity problem of blog network should be addressed as there are quite a few links among the entire blogspace and the majority of blogs are isolated. Hence, many approaches are introduced to tackle the problems by adding implicit links between blog entries, such as the similarity of counting the number of common tags/categories, the number of coupling URLs to news article, and the number of authors posted in both weblogs [11]. We use the following two-dimension table to demonstrate crucial factors in inducing SIP score. Accordingly, the table enables us to clarify the distribution of factors in computing the score of social intimacy similarity and helps us to reason and build our models based on these implicit links.



Social Link Direction	Involved Entities: Object-Object	Involved Entities: Agent-Object
In-degree Link	- Citations cited by same blog post	- Comments (citations) contributed (cited) by same author
Out-degree Link	- Mutual Links to same posts or other websites - Co-citation	- Co-comment

Figure 3.2 Distribution of related social links among blogs

In this study, we utilize the multiplicity of links which takes social intimacy and popularity as a basis to calculate our score of social relation in a more comprehensive and exquisite way.

3.4. Semantic Similarity (SS)

The main purpose of semantic similarity analysis is to induce the most similar posts and to discover potentially interesting posts for the bloggers. While blog search technology provide similar services, it actually focuses on rating blog entries based on their similarity of posts contents or topics, once a set of keywords or tags is given. Compared with it, SS analysis provides full-text based content matching approach to compute textual relatedness of a blog post pair. Therefore, information retrieval, text mining, or social tagging methodology is proposed to handle these issues. As to this study, a traditional IR approach is applied to compute textual similarity between given weblog posts in the entire blogspace.

Integrating the above concepts which rates the weblog posts according to its own trustworthiness and reliability, social intimacy, popularity, and semantic similarity scores in a combinatorial manner, is able to induce a comprehensive and exquisite blog recommendation score. Then a recommendation list of weblog posts is generated by ranking scores from high to low.

3.5. Neural Network-based User Evaluation Process

In this study, a user is defined as a blogger who has his/her own weblog, interacting with other bloggers and has the needs to discover familiars in the blog social network.

In this step, a three-layer back-propagation neural network (BPNN) is employed to forecast the FRS (Final Recommendation Score) for each weblog post after we get feedbacks from the results of user evaluation. The number of input neurons in the input layer is three (that is TR, SIP and SS score respectively). For the sake of output of training data of neural network, we design a web-based evaluation interface to collect the feedbacks from users according to the initial recommendation list. This process aims to generate the output data of back-propagation neural network for training, that is, the user feedback is deemed as the actual output of output layer in the back-propagation neural network.

To train the network, we set a threshold value as a performance target and train the network until the network reaches convergence. Then, the FRS can be derived for each bloggers/posts by employing the trained network. In other words, feeding another TR, SIP and SS scores into trained network and FRS is generated finally. Due to the ability of learning and forecasting of BPNN, the proposed model could capture the social behaviors and preference patterns of users in which a truth-revealing recommendation result will be produced.

CHAPTER 4 Research Methodologies

This study proposes a neural network-based blog recommendation mechanism combined with the concepts of trust, social relation, and semantic analysis. This mechanism contains the information of the blog network about trustworthiness and reliability, social intimacy and popularity, and semantic similarity respectively. The whole process of recommendation mechanism is divided into several steps as shown in figure 3.3 and is described in the following sub-sections.

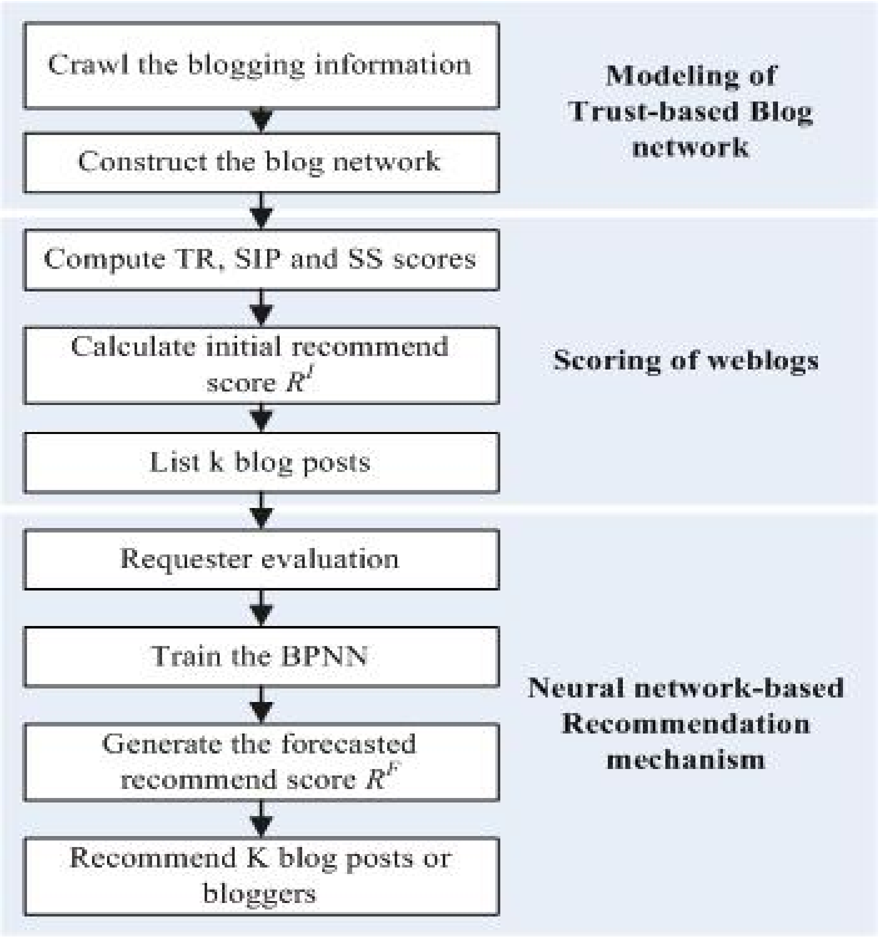

Crawl the blogging information
Construct the blog network
Modeling of Trust-based Blog network
Compute TR, SIP and SS scores
Calculate initial recommend score R^I
List k blog posts
Scoring of weblogs
Requester evaluation
Train the BPNN
Generate the forecasted recommend score R^F
Recommend K blog posts or bloggers
Neural network-based Recommendation mechanism

Figure 4.1 The whole process of recommendation mechanism and it's sub-sections

Note that a recommendation score for an object (agent) in this study represents combined degree of trustworthy, potentially alike in social interaction, and semantically related in blog contents with respect to the recommendation service requester. In other words, the object (agent) with the higher score has more value and utility to be recommended to requester, and he/she will have a greater preference and likeness toward the object (agent).

4.1 Trust-based Blog Network Model

Crawl the blogging information

First of all, we take the blogsite of the 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 reached, the members of the recommender are confirmed. Then we crawl blogging information (such as blog posts, hyperlinks, comments, etc) associated 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 which applies the concepts of agent and object in [5]. In this graph-based representation blog network (shown in figure 4.2), *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. After that, the scope of object layer will be determined by these objects which can be reached by these agents in the agent layer. First of all, the problems of clarifying the existence of links and of classifying and annotating known links for both explicit and implicit ones are the first steps toward identifying potential relationships in this incomplete graph. In this study, the relations are classified into following three aspects: Agent-to-Agent relation, Agent-to-Object relation, and Object-to-Object relation.

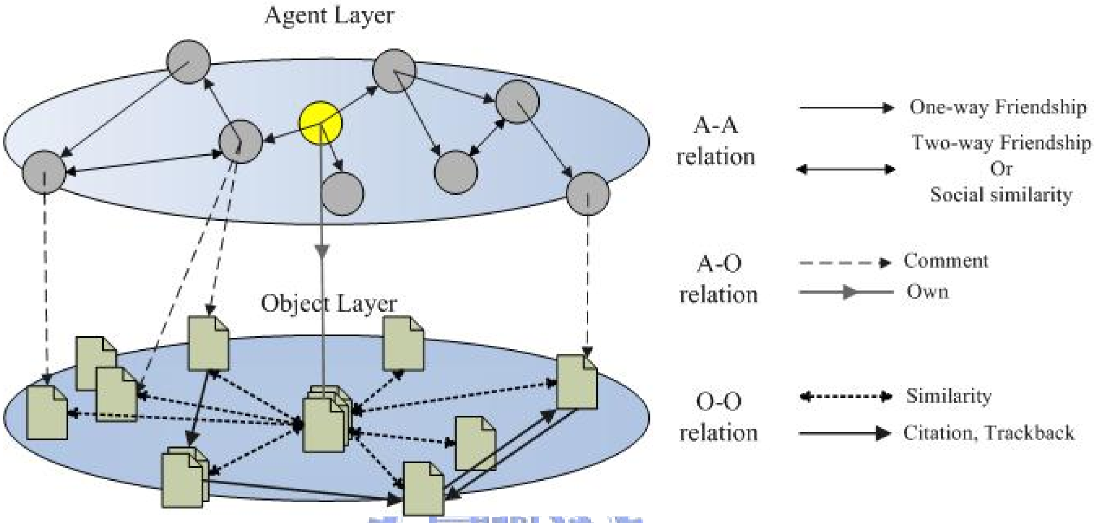


Figure 4.2 The definitions and classifications of links among blog network

4.1.1 Agent to Agent Relation (A-A relation)

A-A relation contains two kinds of relations. Firstly, a friend or friend-of relation, reflected in the blogroll, is a hyperlink from agent to agent. We quantify the relation as a degree of trustworthiness and reliability toward an agent who is worthy to be conducted a belief and commitment that the agent will have good referral or recommendation behaviors, i.e. trust degree. As a result, It forms (trust degree) the TR score.

Second sort of relation is about social similarity level which measures the strength of social intimacy and interaction in common between agents. In this section, not only real links in physical but also implicit similarity relations of social behaviors are taken into account, i.e. links in common, topic similarity, number of hyperlink in common, the number of same tags or comments contributed by same author in a post, etc. By aggregating these relations, we could derive a social similarity score. In this study, we take behaviors of comment and citation for constructing a part of SIP score.

4.1.2 Agent to Object Relation (A-O relation)

In a blog social networking environment, much of the interesting interactions occur in comment behaviors and it is the most interactive and conversational way, compared with other interactions. This kind of agent to object relation not only reveals the interests and social intimacy of blogger (commentator) toward specific blog post but also shows the popularity of bloggers. It is intuitive that a certain object will get a higher popularity score when it has more comments and citations (in-degree links) from other agents in spite of the community type, semantic of blog posts, and recency /freshness factors. In examining the SIP score associated with popularity degree, comment is a crucial social behavior to express the social importance in blog network.

Another relation between agent and objects is possession relation and it implies that objects are submitted by an agent. Here is the entrance to connect agent with object layer for the purpose of inducing a personalized and requester-oriented social networking and computing mechanism.

4.1.3 Object to Object Relation (O-O relation)

In addition to computation of SIP score, citation and trackback behaviors should be brought into model to improve the recommendation completeness Thus, we especially emphasize on similarity between objects. Previous studies reckons similarity as an important perspective in recommendation domain [7][16][17]. In blog context, similarity plays the same role in recommending blog articles and bloggers. The proposed approach divides the concept of similarity into two sorts: social intimacy similarity and semantic similarity of blog posts which associate with SIP and SS score respectively.

4.2. Scoring Approaches of Weblogs

Calculate initial recommendation score R^I and list K blog posts

We compute a initial recommendation score (either for post or blogger) according to their scores of trustworthy, social relation, and semantic similarity after a min-max standardization approach applied to each score (showed by upper case in eq(4.1)). An initial recommendation list is generated with a sequence of recommendation score ranking from high to low. Recommendation scores $R(i, j)$ for each post *j* of blogger *i* for given the requester *r* is defined as following:

$$R^I(r,o_{ij}) = \alpha TR^s(r,i) + \beta SIP^s(r,o_{ij}) + \gamma SS^s(r,o_{ij}), \qquad (4.1)$$

where uppercase *I* of recommendation score R^I stands for initial recommendation score and uppercase *s* of TR, SIP and SS scores mean scores after the process of standardization. Parameter 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 their values are between 0 and 1.

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

Compute TR, SIP and SS Scores

4.2.1 Trust Scores

The interpersonal trust (an agent-to-agent relation) values derive directly from blogroll relationships (i.e. 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 a 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*. As listed in the following equation, the TR score of agent *s* is computed by trust inference mechanism and it is the most widely used one in trust-based social networking and computing approach [7]:

$$TR(r,s) = t_{rs} \text{ and, } t_{rs} = \frac{\sum_{j \in adj(r)} t_{rj} \times t_{js}}{\sum_{j \in adj(r)} t_{rj}}, \qquad (4.2)$$

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 $t_{rs} \in [0,1]$,

adj(r) means adjacent agents of agent *r*, i.e. friends of blogger *r*.

4.2.2 Social Relation Scores

This section measures social intimacy and population (SIP) score of each agent in the blog network via their interrelationships and shared properties. Combining a complete view in recommendation process, SIP score is divided into SI and Popularity scores, SI addresses the social similarity strength or the degree of familiar on agent-agent aspect. While, Popularity emphasizes global reputation on object aspect (shown in figure 4.3).

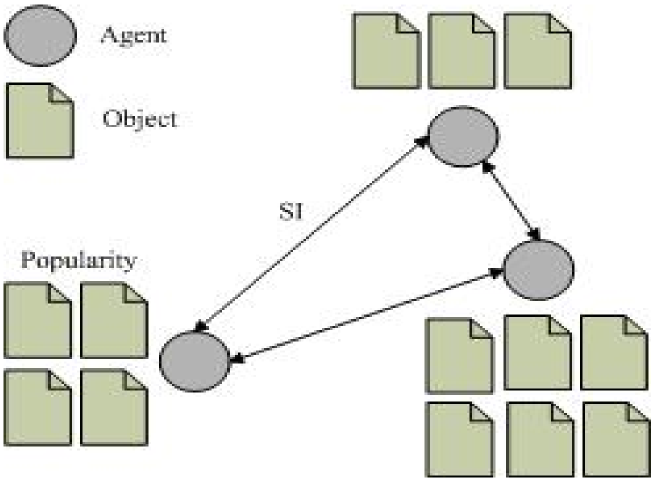


Figure 4.3 The sketch map of Social Intimacy relation and Popularity

SIP score is introduced in the following:

$$SIP(r,o_{ij}) = \alpha SI^{s}(r,i) + (1-\alpha)Popularity^{s}(o_{ij}), \qquad (4.3)$$

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, and $SI^{s}(r,i)$ and $Popularity^{s}(o_{ij})$ represents social intimacy relation and popularity scores respectively. α is the self-set weight and uppercase *s* means the score after standardization process.

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

$$SI(r,i) = sim(iL(r,A), iL(i,A)) + sim(oL(r,A), oL(i,A)), \qquad (4.4)$$

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*, which is shown in vector form.

Popularity measures social importance of an agent or object in blog network. In general, three approaches are suitable for ranking nodes in a graph-based representation network: in-degree, HITS [9] and PageRank [4]. We measure the in-degree (the number of incoming links) in our model as a rough substitute for popularity for the ease of computing. Since an object *u* belongs to an agent *s*, we compute the aggregate value of *u* as a weighted sum of the relative number of comments and citations are as follows:

$$Popularity(o_{ij}) = w_{co} \times \frac{Comment(o_{ij})}{\max Comment(A)} + w_{ci} \times \frac{Citation(o_{ij})}{\max Citation(A)}, \qquad (4.5)$$

where $Comment(o_{ij})$ ($Citation(o_{ij})$) are 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.

4.2.3 Semantic Scores

Information retrieval techniques are originally used for extracting meaningful concepts and transforming unstructured text to structured data from documents. Especially in blog context, the recommendation target, source and the nature of interaction focus on texts, including topics, article contents which imply and convey significant information about the bloggers themselves. In this section, we apply traditional IR technique to compute the textual similarity among blogs and blog posts. There are several steps needed to calculate the semantic score of each post in the network (shown in figure 4.4).

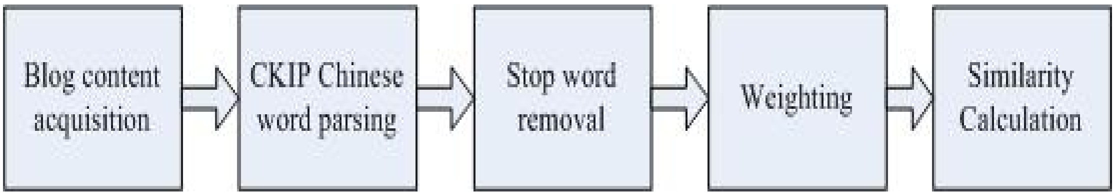


Figure 4.4 The steps of semantics similarity analysis

Once the blogging data is crawled and HTML tags are removed, we apply CKIP (Chinese Knowledge and Information Processing) [14] Chinese word segmentation system to parse the content of blog post after the HTML tags are removed. CKIP project in Academia Sinica proposes the Chinese parser to facilitate word segmentation and provides not only the functionality of word segmentation but also the morphological information of each word.

For the process of removing stop words, we extract several syntactical functions and morphological features (nouns and besides we select several kinds of verbs) that help us to extract useful terms for representing the documents. Then the remaining words are the index terms. A basic cosine similarity metric of term vectors with standard TFIDF [15] weighting scheme is used to represent each index term of each blog article. Semantic score measures textual similarity of blog posts between the requester and 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 or object-to-object score and is defined as blow:

$$SS(r,o_{ij}) = sim(q,d_{ij}),\ \ i \in [1,n],\ \ j > 0\ \text{ and }\ 0 \le SS(r,o_{ij}) \le 1, \qquad (4.6)$$

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*.

The similarity comparison is limited within constructed blog network. On one hand, the agents in the blog network are introduced according to trust-based filtering mechanism which is more trustworthy and reliable to induce a better recommendation. On the other hand, the process is computationally efficient to deal with the problems of information overload and scale reduction of the blog recommendation source pool.

4.3. Neural Network-based Recommendation Mechanism

A back-propagation neural network (BPNN) model is one of the most frequently used techniques for classification and prediction, and is special in accommodating complex and non-linear data relationships. Thus, in this section, BPNN is adopted to capture the implicit relationships between these factors (TR, SIP and SS) and requester's preferences in blog social network accurately in a comprehensive view to forecast the FRS for each object or agent.

Requester evaluation

Once the initial recommendation list of *k* blog posts (bloggers) is delivered to the requester, it accompanies with a detailed principles of evaluation by a web-based interface to help the users fill the form with the satisfaction scores (see Appendix B). For a 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, calculate the forecasted recommendation score R^F, and then generate recommendation list of k blog posts or bloggers to the requester

The characteristics, preference, and social behaviors vary dramatically among human beings. Neural network-based recommendation mechanism is special for its leaning and forecasting ability 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 recommendation score with respect to three scores will be learned (i.e. weights α, β and γ for TR, SIP and SS scores respectively) through the neural network. Therefore, to train the back-propagation neural network, we combine three scores i.e. TR, SIP and SS, and the results from the requester evaluation process as testing data for BPNN. Once the network is trained, it can be used to calculate the forecasted recommendation score R^F and then generate recommendation list of *k* blog posts or bloggers to the requester.

CHAPTER 5 Experimental Studies

So far we have introduced trust model, social relation and semantic analysis into our model. They present crucial factors to guarantee high-quality recommendations in blog network. In this section, we apply the proposed recommendation framework to Wretch, a famous blog system accommodating millions of uses to interact with others [25] in Taiwan and show the entire recommendation processes. We then conduct an empirical experiment to examine the effectiveness of proposed blog recommendation mechanism and the satisfaction level of service requester.

We begin by explaining how the dataset was collected. Then some statistical data such as the number of bloggers in the recommendation network, average number of friends of bloggers and of blog posts for each blogger will be presented. In the following subsection, we introduce how to build trust network to calculate TR score. Experiment results and evaluations are addressed in the end.

5.1 Data Descriptions

We test our proposed mechanism by using a dataset collected from the Wretch [26] which is a Taiwanese community website. It is the most famous weblog community in Taiwan with millions of users registered now. In this website, users can upload photos to album, write the blog, and interact with others by these services [25].

In early July 2007, we start to 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 for constructing the recommendation network. Note that the objects are crawled according to the agents which have been crawled.

The detailed statistics information of this experimental recommendation network is presented in 5.1 and 5.2. It can be observed that the network size is drastically increasing, and we can predict that the network will achieve a saturated situation when the network spreads 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 5.1 Statistics of recommendation network (up to 3rd layer)

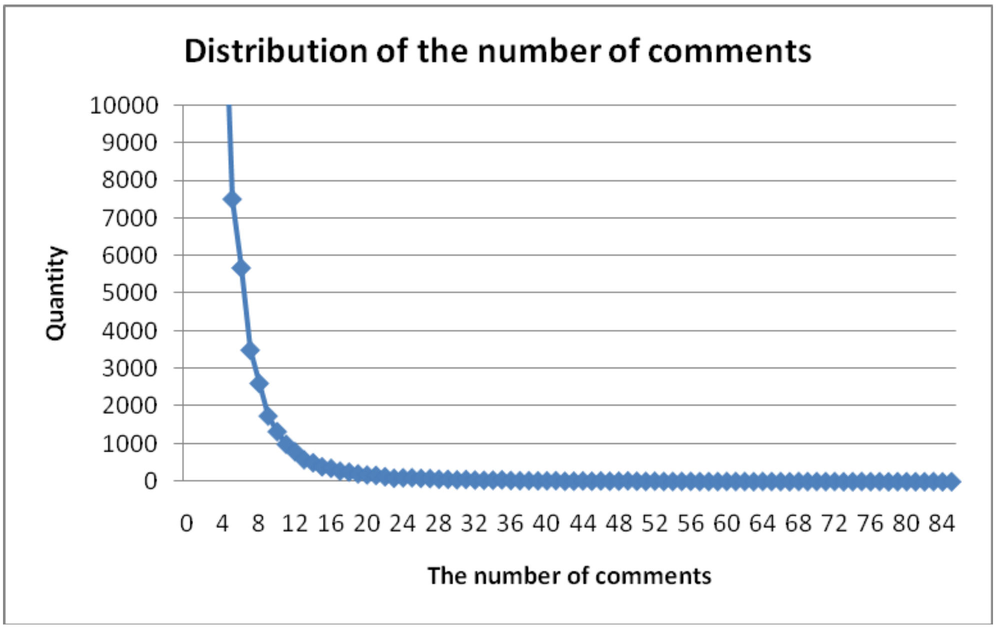
Characteristics of recommendation network	**Statistics**
# of agent (blogger) in the network	22,336
# of object (blog post) in the network	338,614
Average # of friend of an agent	29.722
Average # of objects of an agent	15.160
Average # of comments of an object	2.382
Average # of citations of an object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

Table 5.2 The # of agent and friend relationship in each layer according to the root: "chiang1000"

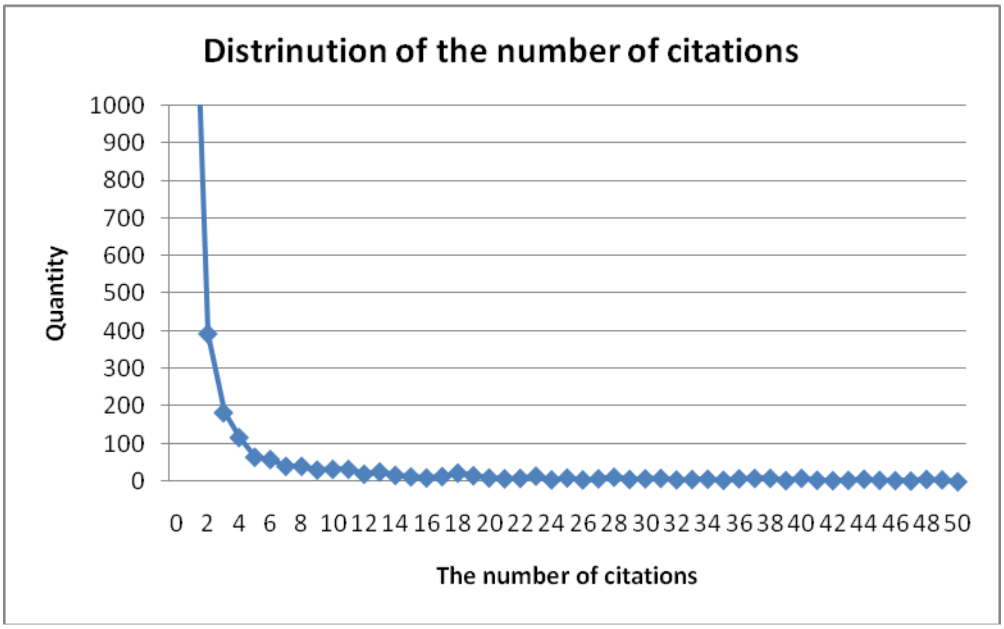
# / layer	**root**	**1st layer**	**2nd layer**	**3rd layer**	**4th layer**
The # of agent	1	23	927	21,384	299,539
The # of friend relationship	23	972	30,299	632,389	NA

In this research, an experimental small recommendation network about 20,000+ agents and 330,000+ objects was constructed and limited the layer to 3rd layer, due to the reasons that the network size grows up exponentially with the layer increased, which will result in a decreasing computability of trust and semantic similarity.

To describe entire network, about 57.22% of objects are isolated and without any comment and citation. From Figure 5.1, we found that 99% of the objects have comments range from 0 to 15, 80% range from 0 to 2, but 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 earlier, our approach seeks to increase the density of the implicit links between bloggers and between blog posts. This will enhance the reliability and comprehensiveness of recommendation mechanism.



(a)



(b)

Figure 5.1 The distributions of the number of (a) comments and (b) citations in our dataset

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 on the trust network starting from requester, into our dataset. We conduct our experiments with pre-selected target requesters who provide recommendation information and evaluate the effectiveness of the proposed recommendation mechanism in this study.

5.2 Building Trust Network

Prior works took user profile similarity or rating similarity over items as degree of trust among users [6][7]. However, in the context of blog recommendation, these methods cannot not be used due to the sparsity problem. As such, we need to develop a trust-generating network suitable in the blog environment. Figure 5.2 shows a visualization of trust network taken from one of Wretch blogger account "chiang1000". A complete line represents a trust value and a dotted line means lack of trust value toward certain blogger. The direction of arrow stands for the direction of trust information.

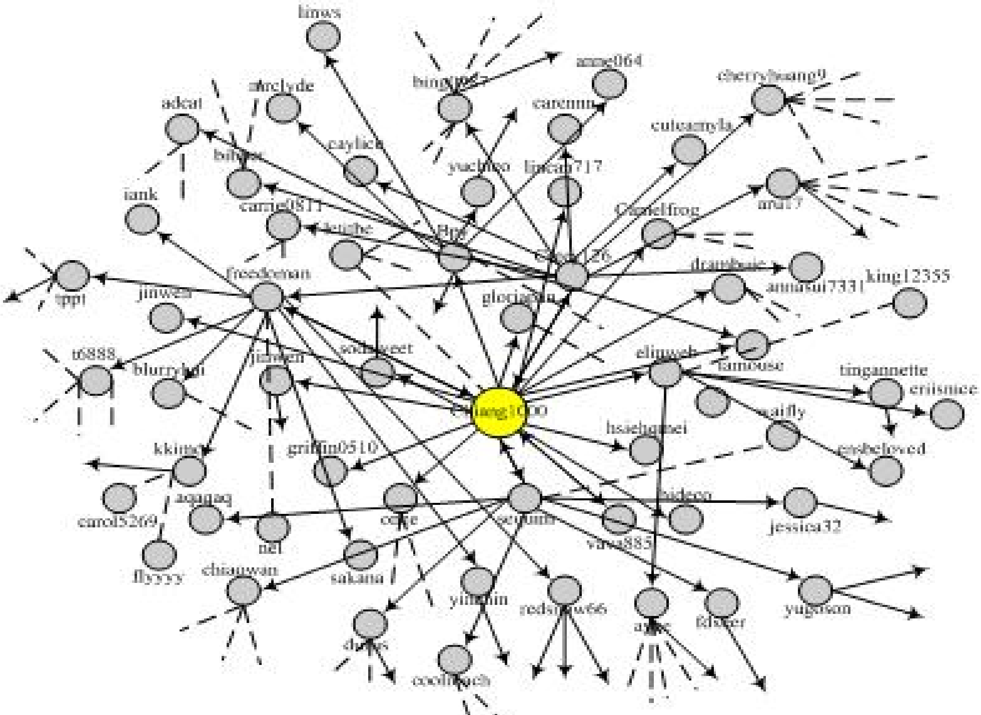


Figure 5.2 A visualization of portion of trust network which is spanned from the core of blogger account “chiang1000”

Trust can be generated by the trust values directly assigned by each blogger to his/her friends (see Appendix A). We design an interface to receive trust from the bloggers. Using the interface, bloggers are asked to assign a trust value to each of his/her friends on the blogroll of blogsite. While bothering for bloggers, this is one of the ways to realistically capture the degree of trust. Then, we calculate the trust value of each blogger for a requester either by trust inference approach or average the trust value toward certain blogger once the inference approach cannot reach it. As to the rest of bloggers who lack of trust information in this recommendation network, we then ignore them.

5.3 Experimental Results and Evaluations

The experiment is conducted with 6 Wretch bloggers who did not have prior knowledge about the recommendation algorithms used in the system and who had different preferences. The targeted users are asked to examine the initial recommendation list to judge the recommendation results on a 10-point scale ranging from strongly satisfaction to strongly dissatisfaction (1, Very unsatisfied; 5, Average; 10, Very satisfied). The averaged satisfaction score can be used to indicate the degree of fitness and user satisfaction between the users' preferences and recommended articles or bloggers.

5.3.1 Recommendation Strategies

We design seven different recommendation strategies to evaluate the proposed mechanism. Some of which are commonly used approaches provided by blog service providers (BSP) (i.e., Wretch) [26] as the comparison benchmarks. The followings are the different recommendation strategies we use:

1. ANN+All, which is the approach proposed in this study. We apply back-propagation neuron network to learn the final recommendation scores from the combinations of TR, SIP and SS scores.
2. All, which results in initial recommendation scores. Without BPNN to learn the non-linear relationships between TR, SIP, SS scores and final recommendation scores is addressed. In other words, recommendation scores is formed by the weighted sum of these scores (see Eq. (4.1)). In this study, we set $\alpha = 0.3$, $\beta = 0.3$ and $\gamma = 0.4$.
3. SS, which purely takes the semantic similarity of contents into consideration. In this strategy, we ask six targeted users to select an article published in their blog site. We then focus on processing this selected article to calculate the content similarity with other

articles in recommendation pool. It is similar to “full-text search” in a different form to some extent.

4. Random, which simply recommends *k* articles or bloggers at random.
5. Comment, which recommends top-*k* articles or bloggers with more numbers of comments at certain time period.
6. Citation, which recommends top-*k* articles or bloggers with more numbers of citations at certain time period.
7. Hotness, which recommends top-*k* hottest articles or bloggers. The degree of hotness is dependent on the number of visitors of blogsite at certain time period.

In this experiment, we emphasize the power and robustness of hybridization of these factors accompanied with the preference predicting ability of BPNN in recommending weblogs (i.e., ANN+ALL strategy).

5.3.2 Neural Network Prediction Model

We apply back propagation neural network (BPNN) to recognize the preference patterns and predict the final recommendation score of each target user in the proposed mechanism. We utilized neural network toolbox of Matlab software to implement our model. The following table outlines the relevant network parameters and learning settings used in this experiment.

Table 5.3 Network parameters and learning settings

Parameters	**Value**
Initial learning rate (lr)	0.001
lr_inc	0.1

lr_dec	10
Epochs	500
Number of hidden layer	1
Number of input neurons	3
Number of hidden neurons	5 / 10 / 15 / 20

We apply adaptive learning rate approach to accelerate the convergence of back-propagation learning to adjust the learning rate parameter during learning process. Learning rate (lr) is multiplied by parameter lr_inc (lr_dec) whenever the performance function has an incremental increase (reduced).

A total of 20 subjects of each target users are gathered from the requester evaluation, and they are divided into 80%/20% training/testing data. Thus, there are 16 training subjects used for BPNN and 4 testing subjects for evaluation of the prediction ability of BPNN model. Therefore, we applied BPNN to select better training parameters for generating the final recommendation list. The mean absolute prediction error (MAPE) and the root mean square error (RMSE) are adopted to evaluate the BPNN effectiveness. The formulas are shown in Equation (5.1) and (5.2).

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{x_i}\right| \times 100\%, \qquad (5.1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \qquad i=1,\ldots,n, \qquad (5.2)$$

where y_i is the predicted output, x_i is the actual output, and n is the number of tested data. When the MAPE and RMSE of test data set is more close to 0, it is indicated that BPNN model has more precise prediction ability.

For training parameter settings, 5, 10, 15 and 20 units of neurons in hidden layer are evaluated individually to derive a better BPNN model to minimize the error function for each user. Since the preference patterns are different, prediction models vary with users. In line with the concept, we develop different BPNN models for different users. The evaluation results of recommending articles and bloggers, displayed in average RMSE (MAPE %), of each BPNN models under different number of neurons in hidden layer of each user are listed in table 5.4 and 5.5 For each score listed in both table is generated from taking the average value of five different trials with same parameters and value settings.

Table 5.4 The evaluation results of recommending **articles**, the average RMSE (MAPE %) under different number of neurons in hidden layer

User (account_id) / # of neurons in hidden layer	**5**	**10**	**15**	**20**
1 (Chiang1000)	**0.31** (35.571)	0.366 (37.242)	0.281 (36.131)	0.38 (40.875)
2 (Freedoman)	0.087 (55.394)	**0.079** (47.962)	0.083 (54.88)	0.085 (54.979)
3 (Cutey126)	**0.176** (24.01)	0.27 (37.281)	0.225 (30.901)	0.22 (30.187)
4 (Vivachu)	**0.234** (23.808)	0.246 (25.973)	0.237 (24.83)	0.252 (26.806)
5 (Vava885)	**0.277** (27.206)	0.259 (27.988)	0.369 (38.826)	0.286 (29.49)
6 (Anny0307)	**0.165**	0.519	0.437	0.406

	(27.621)	(89.643)	(75.116)	(66.79)

Table 5.5 The evaluation results of recommending **bloggers**, the average RMSE (MAPE %) under different number of neurons in hidden layer

User (account_id) / # of neurons in hidden layer	**5**	**10**	**15**	**20**
1 (Chiang1000)	0.335 (140.429)	0.158 (39.699)	0.459 (87.63)	**0.13** (46.17)
2 (Freedoman)	0.142 (12.624)	0.112 (10.693)	0.164 (13.765)	**0.104** (8.338)
3 (Cutey126)	0.295 (31.287)	0.253 (25.937)	**0.174** (19.229)	0.18 (17.476)
4 (Vivachu)	0.25 (29.507)	**0.17** (19.845)	0.481 (57.765)	0.364 (42.698)
5 (Vava885)	**0.135** (18.552)	0.249 (32.192)	0.224 (29.106)	0.188 (25.121)

The smallest RMSE value was marked in bold face in each row to denote better prediction ability of the BPNN model in both tables. This allows us to choose the appropriate hidden neuron number. We can observe that the MAPE does not perform well (significantly low). This may be due to the reason that training data is not insufficient enough to capture the complicated human decision patterns. The MAPE varied with different users, ranging from 10% more to 70% more in average. That shows that the preference pattern of each user is rather different and hard to capture if training data is rather small. However, one may gives totally opposite satisfied scores at different time. As such, the prediction model should keep learning and adaptive to user's variability by feeding more and more training data. Although the overall average RMSE (MAPE) taken for predicting final recommendation scores of articles and bloggers is 0.199 (31.03%) and 0.329 (22.435%) respectively, the proposed mechanism still outperform the others.

5.3.3 User Evaluation Results

The following firgures indicate the stretegy of ANN+ALL and ALL being the best and second best respectively among other strtegies. Figure 5.3 and 5.4 confirm the proposed blog recommendation mechanism is the best in average satisfaction level, compared to other approaches.

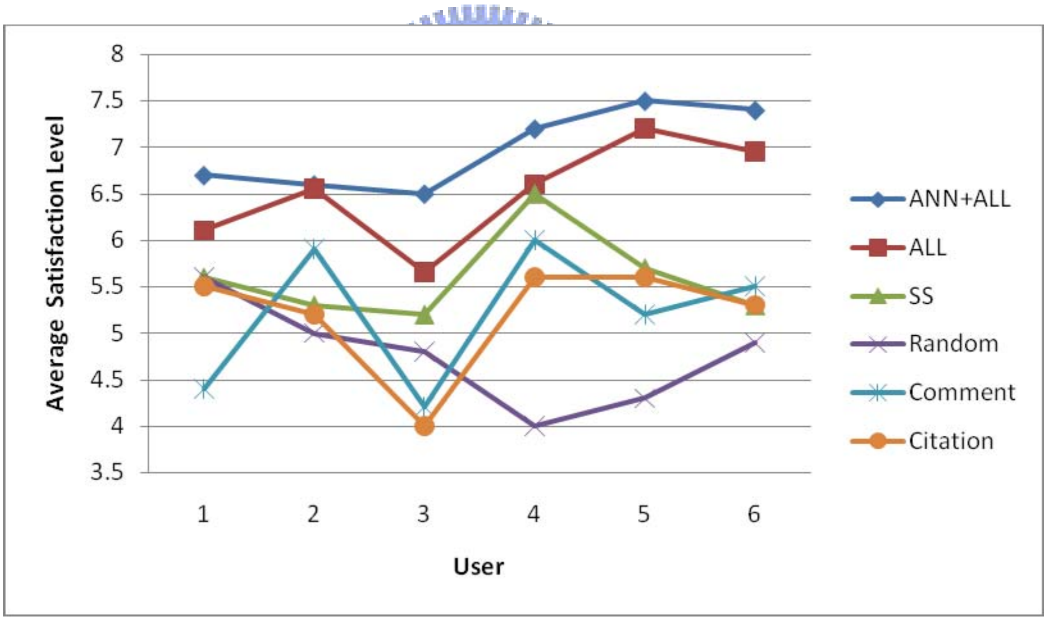


Figure 5.3 The evaluation results of recommending articles

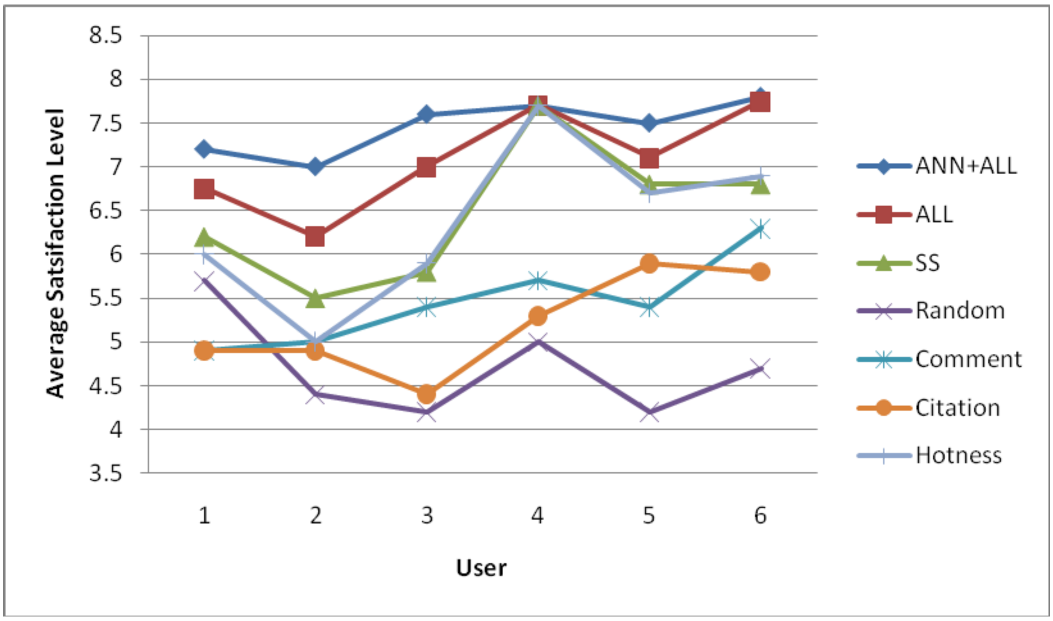


Figure 5.4 The evaluation results of recommending bloggers

The statistical test (e.g. paired sample t-test) is used to further confirm the significance of the differences in the recommendation results. As shown from Table 5.6 to 5.16, at 95 % significant level, both results of recommended articles and bloggers are statistically significant in terms of average satisfaction level. The results reveal that the proposed synthetical neural network-based approach is the best compared to others in the domain of blog recommendation.

Table 5.6 The statistical verification results of recommending articles: ANN+ALL versus ALL

	Number of users	**Mean**	**Std. dev.**	**T-value**	**$t_{0.05,5}$**
ANN+ALL	6	6.983	0.436		
ALL	6	6.508	0.563	4.2*	2.015
Paired Difference	6	0.475	0.277		

*Significant at p < .01

Table 5.7 The statistical verification results of recommending articles: ANN+ALL versus SS

	Number of users	Mean	Std. dev.	T-value	$t_{0.05,5}$
ANN+ALL	6	6.983	0.436		
SS	6	5.6	0.482	6.781*	2.015
Paired Difference	6	1.383	0.5		

*Significant at p < .01

Table 5.8 The statistical verification results of recommending articles: ANN+ALL versus Random

	Number of users	Mean	Std. dev.	T-value	$t_{0.05,5}$
ANN+ALL	6	6.983	0.436		
Random	6	4.767	0.561	6.141*	2.015
Paired Difference	6	2.216	0.884		

*Significant at p < .01

Table 5.9 The statistical verification results of recommending articles: ANN+ALL versus Comment

	Number of users	Mean	Std. dev.	T-value	$t_{0.05,5}$
ANN+ALL	6	6.983	0.436		
Comment	6	5.2	0.756	6.401*	2.015
Paired Difference	6	1.783	0.682		

*Significant at p < .01

Table 5.10 The statistical verification results of recommending articles: ANN+ALL versus Citation

	Number of users	**Mean**	**Std. dev.**	**T-value**	**$t_{0.05,5}$**
ANN+ALL	6	6.983	0.436		
Citation	6	5.2	0.61	9.115*	2.015
Paired Difference	6	1.783	0.479		

*Significant at p < .01

Table 5.11 The statistical verification results of recommending bloggers: ANN+ALL versus ALL

	Number of users	**Mean**	**Std. dev.**	**T-value**	**$t_{0.05,5}$**
ANN+ALL	6	7.467	0.308		
ALL	6	7.083	0.587	3.02*	2.015
Paired Difference	6	0.384	0.311		

*Significant at p < .01

Table 5.12 The statistical verification results of recommending bloggers: ANN+ALL versus SS

	Number of users	**Mean**	**Std. dev.**	**T-value**	**$t_{0.05,5}$**
ANN+ALL	6	7.467	0.308		
SS	6	6.467	0.799	3.893*	2.015
Paired Difference	6	1.0	0.629		

*Significant at p < .01

Table 5.13 The statistical verification results of recommending bloggers: ANN+ALL versus Random

	Number of users	Mean	Std. dev.	T-value	$t_{0.05,5}$
ANN+ALL	6	7.467	0.308		
Random	6	4.7	0.58	9.714*	2.015
Paired Difference	6	2.767	0.698		

*Significant at p < .01

Table 5.14 The statistical verification results of recommending bloggers: ANN+ALL versus Comment

	Number of users	Mean	Std. dev.	T-value	$t_{0.05,5}$
ANN+ALL	6	7.467	0.308		
Comment	6	5.45	0.509	17.725*	2.015
Paired Difference	6	2.017	0.279		

*Significant at p < .01

Table 5.15 The statistical verification results of recommending bloggers: ANN+ALL versus Citation

	Number of users	Mean	Std. dev.	T-value	$t_{0.05,5}$
ANN+ALL	6	7.467	0.308		
Citation	6	5.2	0.58	10.37*	2.015
Paired Difference	6	2.267	0.535		

*Significant at p < .01

Table 5.16 The statistical verification results of recommending bloggers: ANN+ALL versus Hotness

	Number of users	**Mean**	**Std. dev.**	**T-value**	**$t_{0.05,5}$**
ANN+ALL	6	7.467	0.308		
Hotness	6	6.367	0.937	3.795*	2.015
Paired Difference	6	1.1	0.71		

*Significant at $p < .01$

CHAPTER 6 Concluding Remarks

6.1 Summary of Research Results

This paper proposes a synthetical blog recommendation mechanism to personally recommend suitable blog articles or bloggers to the users in blogosphere of practice. We have combined trust model, social relation, and semantic analysis to develop our model and illustrated how it can be applied to a prestigious online blogging system – Wretch in Taiwan. Trust model measures the trustworthiness and reliability of the targets. Social relation addresses the social intimacy and the similarity of social behaviors in blog social network where both explicit and implicit links are considered. Semantic analysis simply compares the textual similarity of blog articles.

Major findings from the evaluation of the proposed blog recommendation mechanism are summarized as follows. From constructing recommendation network, we found that the recommendation network will almost contain the majority of bloggers of Wretch when the network spreads up to 5 to 6 layer. The network dramatically grows with about 25.1 times in average with the increase of spreading layer. As the network becomes more and more saturated, the expanding scope converges until all of the interconnected bloggers are in the network. This “small-world” phenomenon of blog recommendation network thus verifies the well known theory of “six-degrees”. That is, most bloggers in recommendation network can be linked on average six degrees of traversal, except for isolated bloggers.

Our mechanism, combined TR, SIP and SS, is showed to be an effective in recommending blog articles or bloggers to users. An experimental study is shown how these components combined together will induce the final recommendation score. The trust models defined in this work can not only be used to enhance recommendation trustworthiness and reliability but also be utilized to increase the robustness of CF-based recommendation systems [19].

However, the information related to trust degree is not available if we utilize real data from online blogging system, it means existing online blogging system does not contain the concept of trust degree between any pair of friend relationship.

6.2 Discussions and Limitations

There are some limitations in this work. First, to capture trust information in the real world, we quantify it by asking users to assign the trust values. The invasive requirements toward users thus may cause some disfavor and the trustworthy issues (i.e. some misleading or skewed situations of recommender system). Obviously, the phenomenon that over than half of objects is isolated will debase the value of SIP score. This may causes the recommendation score lays particular stress on the other two scores (i.e. TR and SS), which distort the recommendation scores and denotation of recommendation mechanism. Second, our trust models are constructed on an agent-to-agent level which cannot reflect trustworthiness in an object-to-object level. That is, with regard to objects of certain agent, we treat each object as the same trust level and each of them has the same trust value relative to the requester. In the future work, we will design a more comprehensive trust model to tackle with this issue to induce a complete and robust recommendation mechanism.

As to SS score, we design an interface for the requester to select some of his/her posts to compare content similarity with others. Unlike search engine, some brief keywords would induce numbers of results which make users hard to digest and unable to find what they really want. More index terms would be helpful for users to accurately locate the needed information [27]. In this work, article selecting process (i.e. select the posts to list into comparison target) would indeed increase the efficiency and accuracy of calculation of semantic similarity. As for processing procedures of Chinese words, each step could be refined and advanced for more accuracy calculation of SS scores.

In recommendation strategies, four existing recommendation (i.e. Random, Comment, Citation, Hotness) approaches applied by Wretch is used here as benchmark approaches. From the experimental evaluation results, the neural model (ANN+All) and the linear model (All) outperform the others. Still, we may wonder that if we take the traditional collaborative-based filtering, content-based filtering, or even other recommendation techniques into the comparison approaches, is the proposed model still has a better performance than the others? We should extend the recommendation strategies for further comparison for a more convincible evaluation in future work.

Moreover, we can observe that the MAPE performance of the model seems insignificant under the limitation of the insufficient training data, even though the recommendation prediction model still outperformed the other approaches, including the synthetical approach without BPNN training.

6.3 Future Works

Recommendation is an interesting topic in blog applications. Depending on their preferences and interests, the synthetical blog recommendation system will help bloggers to find out not only interesting blogs and blog posts but also trustworthy and socially homoeo-bloggers. In future works, there are still several issues in blog application and social networking:

First, finding the influential bloggers for marketing is attention demanding. Finding marketing influential bloggers for marketing will not only allow us to better understand the interesting activities happening in a social network, but also present unique opportunities for industry, sales, and advertisements. With the advent of online social network, viral marketing/word-of-mouth is increasingly being recognized as a crucial issue in social influence and marketing domains. Especially on blogs, it provides a finest platform both for advertisers to market a new product/service and for customers to locate the purchasing suggestions and comments. In conclusion, finding influential blog sites in the blogosphere is an important research problem, which investigates how these blog sites influence the external world and within the blogosphere [8] In future works, we will address a novel problem of finding influential bloggers for marketing on the blogosphere by proposing a preliminary MIV (Marketing Influential Value) model. We will induce two dimensions, network-based factors, and content-based factors, to identify the potential marketing-aided nodes to help marketers/advertisers in promoting their products/services with less efforts and costs.

Second, proposing a dynamic blog recommender system. There exists a tradeoff between a precise recommender system and computational efficiency. A precise recommendation must gather as more information as possible from bloggers; however, it will result in decreasing of the computation ability as well. We should develop an efficient approach and process, to make the recommendation more realistic and to update the relationships dynamically. As to the computational ability, recommendation may perform better by searching for a more scalable tools and technologies. Such as, cloud computing technique can handle the data processing and computation well under large amount of data. In conclusion, a scalable recommender system is needed in blog service of the real-world application.

Third, integrating the social relations in different social networking services is interesting. In the era of web2.0, users may use many social webs to satisfy their own needs. Upload the images to an image sharing website, join a online community on a social networking site, write a review in a product information sharing site, or publish a diary on their own blogs. It is interesting if we collect all information in these sites and aggregate the social relations among them. We could find out the domain experts, influential nodes, or authorities by proposing a ranking mechanism. We believe that the idea is quite promising and proactive in applications of social-networked service.

References

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

[2] ALi-Hasan, N., & Adamic, L. (2007). Expressing Social Relationships on the Blog through Links and comments, ICWSM.

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

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

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

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

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

[8] Kathy E. Gill. How can we measure the influence of the blogosphere? In Proceedings of the WWW’04: workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.

[9] Keinberg, J. M. (1999). Authoritative sources in hyperlinked environment, Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, 46(5).

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

[11] Kritikopoulos, A., Sideri, M., & 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, 198.

[12] Lee, H.-Y., Ahn, H., & Han, I. (2007). VCR: Virtual community recommender using the technology acceptance model and the user's needs type, Expert Systems with Applications, 33(4), 984-995(12).

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

[14] Ma, W.-Y., & 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, 168-171. Online demo system: http://ckipsvr.iis.sinica.edu.tw/

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

[16] Massa, P., & Bhattacharjee, B. (2004). Using trust in recommender systems: an experimental analysis, In Proc. of 2nd Int. Conference on Trust Management.

[17] Matsuo, Y., & Yamamoto, H. (2007). Diffusion of Recommendation through a Trust Network, ICWSM.

[18] Negnevitsky, M., (2005). Artificial Intelligence- A Guide to Intelligent System, Second edition published, Person education, Ltd, pp.175~180.

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

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

[21] Song, W., Phoha, V. V. (2005). Opinion filtered recommendation trust model in

peer-to-peer networks, Lecture Notes in Computer Science, 237-244.

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

[23] Weihua, S., Phoha, V. V, & Xu X. (2004). An Adaptive Recommendation Trust Model in Multiagent System, Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT' 04), 9, 462-465.

[24] Wikipedia, Neural network: http://en.wikipedia.org/wiki/Neural_network

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

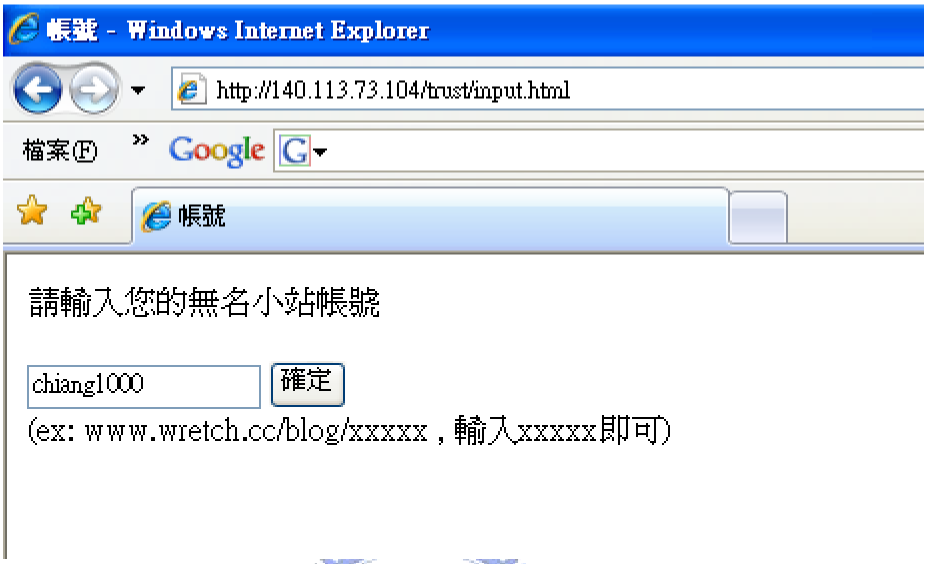
[26] Wretch blog: http://www.wretch.cc/blog/

[27] Yang, K., Yu, N., Valerio, A., Zhang, H., & Ke, W. (2007). Fusion approach to finding opinions in blogosphere, ICWSM.

[28] Zou, H., Xia, G., Yang, F., & Yang, H. (2007). A neural network model based on the multi-stage optimization approach for short-term food price forecasting in China.

Appendix A

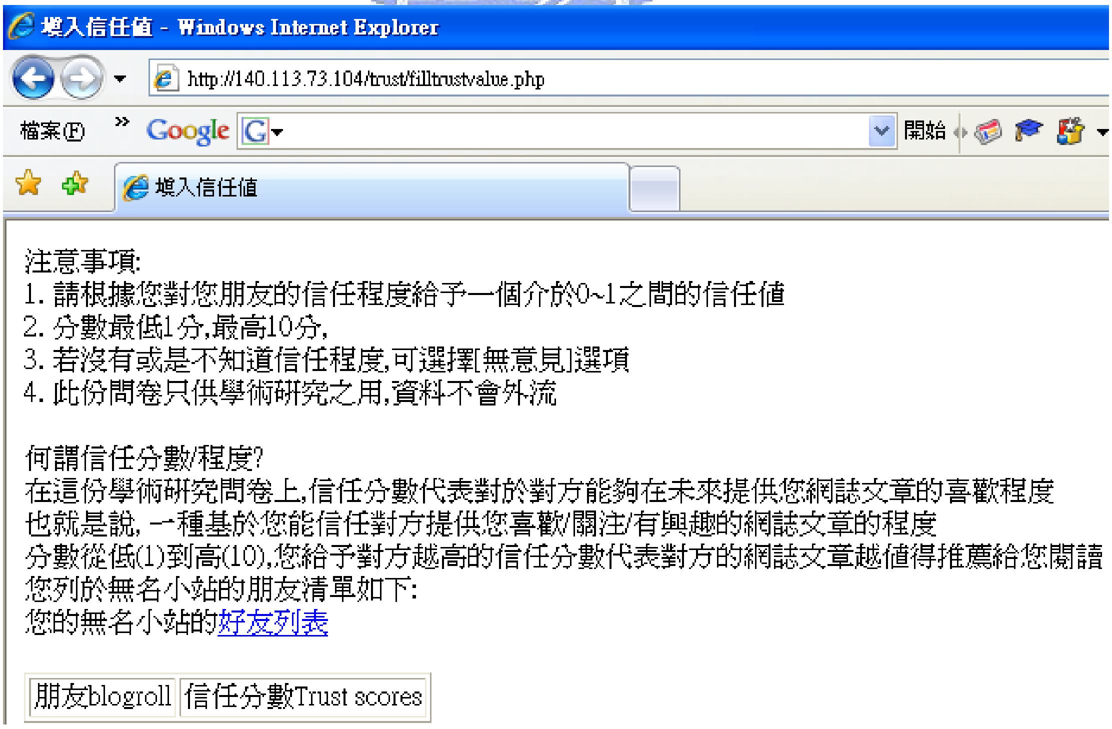
The User Interface of Trust Score Form


帳號 - Windows Internet Explorer
http://140.113.73.104/trust/input.html
檔案(F)
Google
帳號
確定

請輸入您的無名小站帳號

chiang1000

(ex: www.wretch.cc/blog/xxxxx，輸入xxxxx即可)


填入信任值 - Windows Internet Explorer
http://140.113.73.104/trust/filltrustvalue.php
檔案(F)
Google
開始
填入信任值

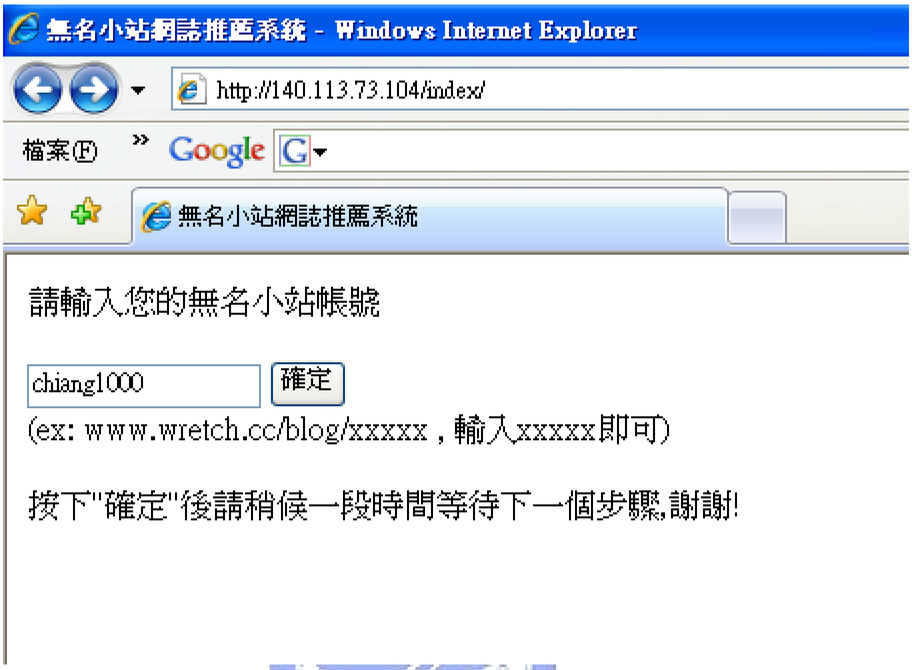
注意事項:
1. 請根據您對您朋友的信任程度給予一個介於0~1之間的信任值
2. 分數最低1分,最高10分,
3. 若沒有或是不知道信任程度,可選擇[無意見]選項
4. 此份問卷只供學術研究之用,資料不會外流

何謂信任分數/程度?
在這份學術研究問卷上,信任分數代表對於對方能夠在未來提供您網誌文章的喜歡程度
也就是說, 一種基於您能信任對方提供您喜歡/關注/有興趣的網誌文章的程度
分數從低(1)到高(10),您給予對方越高的信任分數代表對方的網誌文章越值得推薦給您閱讀
您列於無名小站的朋友清單如下:
您的無名小站的好友列表

朋友blogroll	信任分數Trust scores

Appendix B

The User Interface of User Evaluation Form

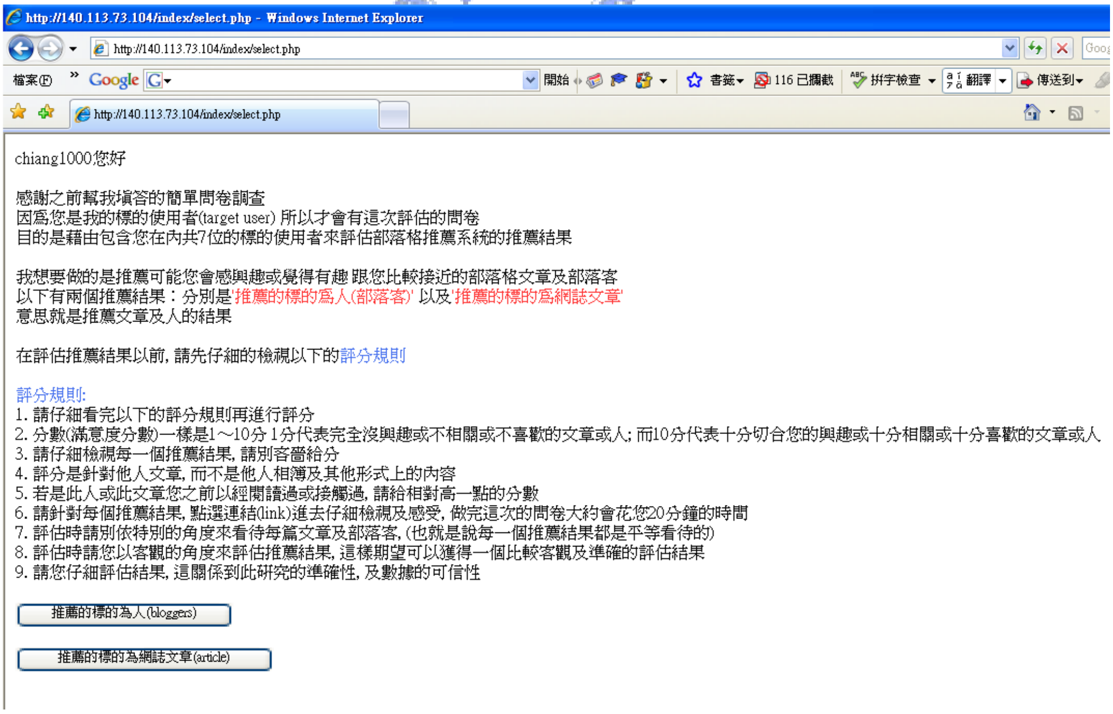

無名小站網誌推薦系統 - Windows Internet Explorer
http://140.113.73.104/index/
檔案(F)
Google
無名小站網誌推薦系統

請輸入您的無名小站帳號

chiang1000 確定

(ex: www.wretch.cc/blog/xxxxx，輸入xxxxx即可)

按下"確定"後請稍候一段時間等待下一個步驟,謝謝!


http://140.113.73.104/index/select.php - Windows Internet Explorer
http://140.113.73.104/index/select.php
檔案(F)
Google
開始
書籤
116 已攔截
拼字檢查
翻譯
傳送到
http://140.113.73.104/index/select.php

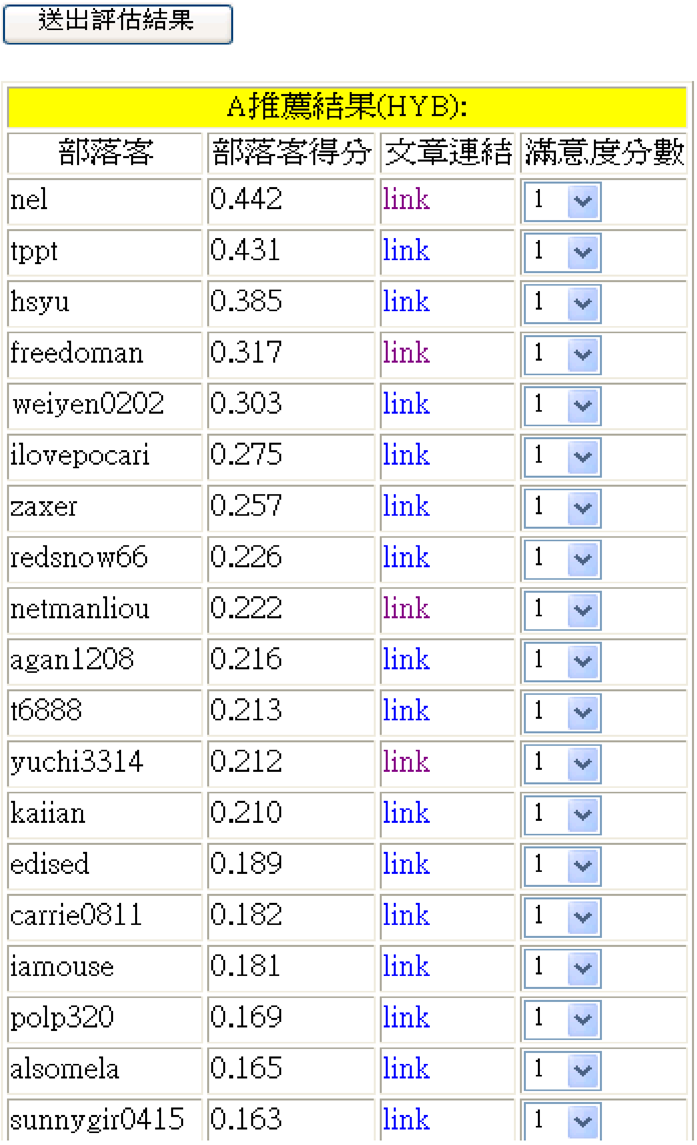
chiang1000您好

感謝之前幫我填答的簡單問卷調查
因爲您是我的標的使用者(target user) 所以才會有這次評估的問卷
目的是藉由包含您在內共7位的標的使用者來評估部落格推薦系統的推薦結果

我想要做的是推薦可能您會感興趣或覺得有趣 跟您比較接近的部落格文章及部落客
以下有兩個推薦結果：分別是'推薦的標的爲人(部落客)' 以及'推薦的標的爲網誌文章'
意思就是推薦文章及人的結果

在評估推薦結果以前, 請先仔細的檢視以下的評分規則

評分規則:
1. 請仔細看完以下的評分規則再進行評分
2. 分數(滿意度分數)一樣是1～10分 1分代表完全沒興趣或不相關或不喜歡的文章或人; 而10分代表十分切合您的興趣或十分相關或十分喜歡的文章或人
3. 請仔細檢視每一個推薦結果, 請別吝嗇給分
4. 評分是針對他人文章, 而不是他人相簿及其他形式上的內容
5. 若是此人或此文章您之前以經閱讀過或接觸過, 請給相對高一點的分數
6. 請針對每個推薦結果, 點選連結(link)進去仔細檢視及感受, 做完這次的問卷大約會花您20分鐘的時間
7. 評估時請別依特別的角度來看待每篇文章及部落客, (也就是說每一個推薦結果都是平等看待的)
8. 評估時請您以客觀的角度來評估推薦結果, 這樣期望可以獲得一個比較客觀及準確的評估結果
9. 請您仔細評估結果, 這關係到此研究的準確性, 及數據的可信性



A推薦結果(HYB):

部落客	部落客得分	文章連結	滿意度分數
nel	0.442	link	1 ▾
tppt	0.431	link	1 ▾
hsyu	0.385	link	1 ▾
freedoman	0.317	link	1 ▾
weiyen0202	0.303	link	1 ▾
ilovepocari	0.275	link	1 ▾
zaxer	0.257	link	1 ▾
redsnow66	0.226	link	1 ▾
netmanliou	0.222	link	1 ▾
agan1208	0.216	link	1 ▾
t6888	0.213	link	1 ▾
yuchi3314	0.212	link	1 ▾
kaiian	0.210	link	1 ▾
edised	0.189	link	1 ▾
carrie0811	0.182	link	1 ▾
iamouse	0.181	link	1 ▾
polp320	0.169	link	1 ▾
alsomela	0.165	link	1 ▾
sunnygir0415	0.163	link	1 ▾