A Virtual Community Member's Referability Determination Model

Shih-Ting Yang

Department of Industrial Engineering and Systems Management, Feng Chia University, Taiwan
styang@fcu.edu.tw

Abstract

This paper develops a "virtual community member's referability determination" model applicable to virtual communities by which to measure the referable value of virtual community members. In terms of community administrators, there are standard and fair member domain specialty management indicators to enhance the substantial effectiveness of knowledge contributors' sharing behavior control. In terms of knowledge demanders, the cost of evaluating the members' referability is reduced, and the probability of obtaining the required and correct knowledge is increased. In terms of knowledge contributors, there will be precise and rational member referability evaluation scores, and the defects in the existing virtual community's incentive mechanisms will be remedied. Thus, the circulation of less referable information can be reduced, with more community members being willing to share knowledge, with the accumulation of professional public knowledge becoming enhanced. Finally the utilization of virtual communities can be activated to catalyze continuous comprehensive development of virtual communities. In addition, this paper develops a Web-based system accordingly for case verification to confirm the feasibility of the methodology. The verification results show that when the system uses about 600 training knowledge articles, the performance of the inference indicators of system can be increased to 82%. Generally speaking, the system performance grows continuously with the periods and training load, and eventually reaches stable and good performance level.

Keywords: Virtual community, Clustering, Data mining, Community members' referability

1 Introduction

As many knowledge contributors are devoted to knowledge sharing, a great deal of information will surely be agglomerated, with the knowledge demanders given an increasing number of channels for more multivariate knowledge reference. To achieve the clustering of referential and professional knowledge, there are considerable challenges. As the Internet brings ever greater convenience of information transfer, numerous knowledge contributors are devoted to virtual communities, sharing knowledge spontaneously to obtain personal community status. For the management of numerous and plural knowledge contributors, most virtual communities use incentive systems to encourage the members to share quality knowledge, with statements violating policy removed manually. This practice is difficult to implement effectively, or achieve substantial control on the knowledge contributors who accumulate community status rapidly by taking advantage of the defects in the incentive systems and keep spreading less referential knowledge. Therefore, the knowledge contributors' referability measurement depends on the spontaneous grading and comments of community members. The members measure the knowledge referability according to the understanding of domain expertise of knowledge contributors, and discuss the comments and professional ability of other members. There will not be any recognized standard due to the cognitive differences between community members. Therefore, the knowledge demander must collect, probe into (poster and knowledge content) and learn domain expertise, in order to measure a knowledge contributor's reference value. However, as the existing virtual community has not provided performance statistics on precise knowledge contributors, and the understanding of the contributors' referability and domain expertise is insufficient, the evaluation may be wrong, and a lot of information needs to be collected, and the time cost of evaluating the knowledge contributor's referability is increased. In addition, the less referential public knowledge is spread continuously, and the community status will lack justice and precision, with the knowledge contributor's reduced power to keep sharing knowledge. Regarding the effect on the overall virtual community, the overall public knowledge quality in the community will become degraded, negatively influencing the development of virtual community. The present operating mechanism of knowledge acquisition and public knowledge shaping (i.e. AS-IS model) is shown

*Corresponding author: Shih-Ting Yang; E-mail: styang@fcu.edu.tw
DOI: 10.3966/160792642019102006011

in Figure 1.

Figure 1. AS-IS model

The virtual community contains not only "knowledge content", but also the process of interaction of virtual community members. Differing from general knowledge platforms, knowledge in the virtual community has different values as it contains the interaction factor of members. The members do not have identical backgrounds and expertise, so that the shared knowledge has different reference degrees in various fields, which is to say, the referability of the knowledge shared by the same member differs in different areas. In view of this, in order to avoid the knowledge demander extracting wrong knowledge due to insufficient background knowledge, and to increase the knowledge demander's knowledge extraction effectiveness and efficiency, this paper proposes a "virtual community member referability determination model" to analyze the referability of domain knowledge of virtual community members for the knowledge demander's reference. The proposed model automatically measures and evaluates the knowledge contributor referability, and provides the determined data for the user to solve the problems in the virtual community knowledge acquisition and public knowledge shaping mechanism (as shown in Figure 2).

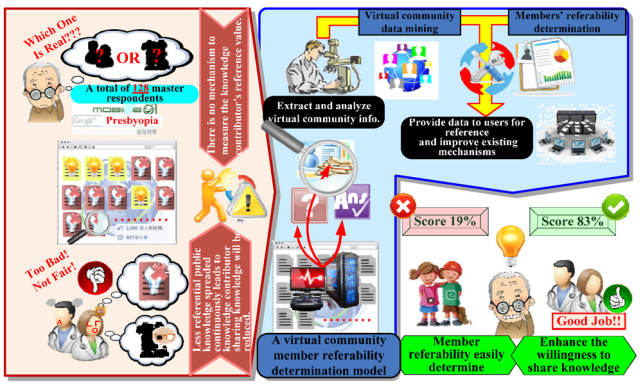

Figure 2. TO-BE model

2 Literature Review

2.1 Virtual Community Members' Knowledge Sharing Intention

The spontaneous knowledge sharing of virtual community members is the prerequisite for the success of a virtual community. In view of this, Fang and Chiu [1] used the psychology-related altruism and conscientiousness, justice theory and trust theory to discuss the principles and fairness of members and managers, and found that the intention of knowledge sharing of members can be changed. Tsai and Pai [2] integrated the far and near end influencing factors and adjustment items, and applied qualitative and quantitative research methods to learn that the overall satisfaction and recognition are rooted in the members' attitude, background, familiarity, and information usability, and that these factors affect the member's spontaneous participation. On the other hand, to discuss the knowledge sharing intention of virtual community members, Hung and Cheng [3] integrated a technological preparedness and technology acceptance model, and concluded that members' optimism, innovation and knowledge content's usefulness and ease to use can affect the members' intention to share knowledge. Chen and Hung [4] applied social cognition theory and social exchange theory, and found that knowledge contribution and collection behavior can be rooted in self-effectiveness, cognitive related advantages, compatibility, and mutual beneficial norms. These behaviors are preconditions for improving the virtual community's operational efficiency. Moreover, Jin et al. [5] integrated expectation confirmation theory, knowledge sharing factor and knowledge self-efficacy, and concluded that evaluation and recognition of achievement will enhance the self-efficacy and confirmation of the contributor to get satisfaction and enhance sharing intention. Based on the Wiki knowledge sharing environment, Moskaliuk et al. [6] discussed the formation of the crowd wisdom, and learned that the crowd wisdom formation efficiency and quality will be at the highest level if there is knowledge of medium and high degrees of similarity. In addition, crowd wisdom is shaped by the dynamic internalization and externalization processes of individuals and social networks.

2.2 Analysis of Virtual Community Members' Profiles

The literature on the subject "member's specialized field bias analysis" and "member's knowledge reliability analysis" are reviewed in this paper. In regard to the "member's specialized field bias analysis", in order to recommend similar professional background users, the private file can be created automatically by user query and retrieval behaviors,

and the users' professional bias similarity, social network and historical article content are analyzed and clustered, with similar users recommended and displayed by Hypergraph [7-8]. In addition, in order to recommend the knowledge professional coincident with the knowledge extractor's required or interested knowledge, the vector space model, TF-IDF algorithm, Markov Chain analysis and Latent Dirichlet Allocation (LDA) can be integrated to analyze the member's private file and historical articles by subject extraction, so as to rank the member recommendation according to the required correspondence [9-10]. On the other hand, in order to ensure that member clustering has correlated domain knowledge, the PageRank algorithm and knowledge database can be used for the semantic analysis of a private file set vector, information importance, and member domain class relationship link degree. The member's social trust network is built via clustering algorithm to provide ranking and recommendation of the member's domain bias [11-12]. In addition, in order to develop the domain knowledge ontology adaptation mechanism for personal knowledge search and recommendation, Chen et al. [13] extracted representative features and knowledge ontology adaptation factors from the user's private file and historical records, and recommended the knowledge matching the User-defined Feature factor after the correlation between the domain bias of features and the knowledge ontology was analyzed.

In order to accurately analyze the specialty and credibility of community members, the PageRank and Hyperlink-Induced Topic Search (HITS) can be used to analyze the link degree between the member's specialty and that of other members of social network according to the member's activity and the context of historical articles. The specialty and reliability levels of knowledge contributors are evaluated by using evidence theory and fuzzy uncertainty concepts [14-15]. In addition, in order to explore the user's latent trustiness in social network, the user's knowledge experience can be analyzed by using Random Forests and Bayesian sorting algorithms to obtain the similarity of user class evaluation, preference and comment, while the class likeness is calculated to obtain the trust correlation and implicit confidence of users in social network [16-17]. On the other hand, in order to improve the low reliability of member recommendation results due to few user evaluation records, and based on the interactions of title, comments and user, the vector space model, TF-IDF and PageRank algorithms are used to analyze the user's professional representativeness in community, with the similar user clusters obtained by automatic private file creation and clustering to display the knowledge specialty data via integrating linear combination, cascade ranking and scaling strategy. Liu et al. [18] combined the Academia Sinica CKIP word segmentation system with TF-IDF for data preprocessing of class similarity integration, and calculated the member specialty score by analyzing the user knowledge file, evaluation information and domain class correlation, to create the domain expert list.

To sum up, firstly, for virtual community member profile analysis, the member's specialized field bias analysis is mostly based on the member's private file and historical use behavior (e.g. historical publication and browsing behaviors) in the virtual community for determining individual member's domain bias. A few studies have considered the interactions of members, so the concepts of social network and clustering are integrated to determine the member's domain bias (see the first subtopic in Section 2.2). Secondly, the member's knowledge reliability analysis mostly aims at the historical publications of the member to obtain the credibility in the domain. The similarity is measured mostly based on term frequency to determine the similarity between the member's historical publications and the articles of the domain, so as to obtain the member's knowledge reliability (see the second subtopic in Section 2.2). As a whole, this paper develops a "virtual community member referability analysis model" with term frequency and semantic analyses, and considering the views of members and wisdom of crowds.

3 A Virtual Community Member's Referability Determination Model

The "virtual community member's referability determination model" proposed in this paper uses the clustering algorithm [19] as the basis of data clustering, the "virtual community domanial discussion threads distribution cluster" and "target member domanial historical publication distribution cluster" can be obtained by "word segmentation and screen valid term", "screen frequent item set", "construct similar matrices", "determine cluster type" and "divide cluster of domain data" in Step (1) to Step (5). Secondly, in order to enable the knowledge extractor to browse more visually, and obtain the specific indices of a virtual community member's knowledge specialty bias and referability, this paper nominates the corresponding clusters through Step (6) according to the cluster nomenclature proposed by Zhang et al. [19], to obtain the cluster names of representative virtual community, the target member domanial discussion threads and historical publication distribution clusters. Finally, this paper integrates the vector space model and the NGD algorithm proposed by Cilibrasi and Vitanyi [20], and uses the concept of cluster similarity [19], to analyze the virtual community domanial discussion threads through Step (A7), and the similarity relationship between clusters in the target member domanial historical publications to obtain the

domain knowledge referability of target member. The "target member domanial historical publication" data are the public information derived from the target virtual community member's publications (including sharing, questioning and response). The "virtual community domanial discussion threads" refers to the discussion threads in various discussion forums in the virtual community. In addition, the complete virtual community domanial discussion threads consist of the titles, questions and answers in the discussion threads. The overall operation of this proposed model is shown in Figure 3.

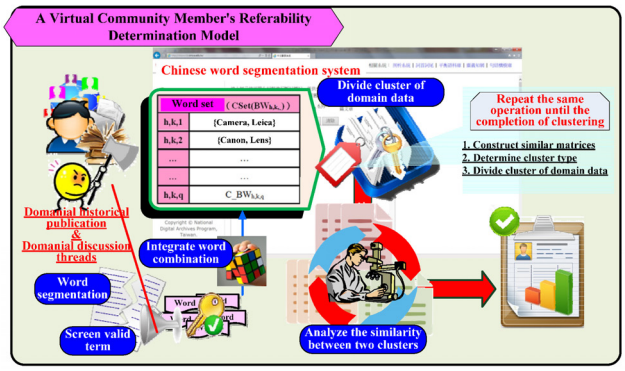

Figure 3. The architecture of this proposed model

Step (1) Word segmentation and valid screen terms. The data preprocessing operation in this model is divided into two major aspects: "word segmentation of data", and "screening of valid terms from the data". In the "word segmentation of data", the CKIP Chinese word segmentation system developed by Academia Sinica is used to segment words for all the data in the integrated dataset $Set_{h,k}$ to obtain the distribution of words in the data $BW(AH_{h,k,\bullet})$ (Equation (1)). This derived result will be used in Step (2) and Equation (17) of Step (7).

$$BW(AH_{h,k,\bullet}) = \{ BW_{h,k,1}, BW_{h,k,2}, ..., BW_{h,k,p} \} \quad (1)$$

In the screening of valid terms from the data, following the result of word segmentation according to the definition mode of parts of speech in the "Sinica Corpus POST set", when the parts of speech of word frequencies are summarized, the words belonging to valid terms (AWT) are maintained. The words of the parts of speech other than those defined in valid terms (AWT) are not kept, in order to filter out the meaningless words and obtain the significant domain data valid terms (AWT) set $F\_BW(AH_{h,k,\bullet})$ (Equation (2)). This result will be used in Equation (18) of Step (7).

$$F\_BW(AH_{h,k,\bullet}) = \{ VW_{h,k,1}, VW_{h,k,2}, ..., VW_{h,k,q} \mid VW_{h,k,q} \in AWT \} \quad (2)$$

Step (2) Screen frequent item set. This step screens "frequent item set" for all data. The frequent item set refers to the word combination with occurrence frequency greater than or equal to minimum support (MinSup). Before the frequent item set is screened, the set $CSet(BW_{h,k,\bullet})$ disregarding sequence word combination must be created according to all domain data (Equation (3)).

$$CSet(BW_{h,k,\bullet}) = \{ C\_BW_{h,k,1}, C\_BW_{h,k,2}, ..., C\_BW_{h,k,e} \} \quad (3)$$

Secondly, for the word combination $C\_BW_{h,k,e}$, the indicator function $I(C\_BW_{h,k,e}, AH_{h,k,i})$ in the domain data $AH_{h,k,i}$ is defined. If the word combination occurs in it, the indicator function is defined as 1; otherwise the indicator function is defined as 0 (Equation (4)).

$$I(C\_BW_{h,k,e}, AH_{h,k,i}) = \begin{cases} 1, If\ C\_BW_{h,k,e}\ exist\ in\ AH_{h,k,i} \\ 0, Otherwise \end{cases} \quad (4)$$

The frequent item set screening threshold is created by the product of "total number of data in domain ($N(AH_{h,k,\bullet})$)" and minimum support (MinSup). If the sum of the indicator functions of word combination is greater than or equal to the screening threshold, the word combination is of frequent item set $FI(AH_{h,k,\bullet})$ (Equation (5)).

$$FI(AH_{h,k,\bullet}) = \{ C\_BW_{h,k,1}, C\_BW_{h,k,2}, ..., C\_BW_{h,k,e} \mid \sum_{all\ i} I(C\_BW_{h,k,e}, AH_{h,k,i}) \geq N(AH_{h,k,\bullet}) \cdot MinSup \} \quad (5)$$

Step (3) Construct similar matrices. This step uses the similarities among all data in the domain to construct "similar matrices" for the similarity analysis before clustering. The similarities between data must be calculated before the similar matrices are constructed. First, the similarity is determined by analyzing the similarity "between two data", so all data ($AH_{h,k,\bullet}$) in the parts of speech must be defined as $AH^{A}_{h,k,a}$ and $AH^{B}_{h,k,b}$ independent data representation modes for subsequent similarity analysis. Secondly, when the data are defined, the similarity determination rule integrating "frequent item set" into similar matrices proposed by Zhang et al. [19] is used to calculate the similarity between data $Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})$ (Equation (6)).

$$Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b}) = N(Set(C\_BW_{h,k,\bullet}, AH^{A}_{h,k,a}) \cap Set(C\_BW_{h,k,\bullet}, AH^{B}_{h,k,b})) \quad (6)$$
$$where\ a < b\ for\ all\ a,b\ and\ C\_BW_{h,k,\bullet} \in FI(AH_{h,k,\bullet})$$

Finally, the similarity data of various data calculated by Equation (6) are integrated to construct similar matrices $SM[AH_{h,k,\bullet}]$ (Equation (7)). As the calculation of similarity is free from the ordinal relation between two data, the calculation $Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})$ and $Sim(AH^{B}_{h,k,b}, AH^{A}_{h,k,a})$ of similarity between two data will produce the same result. In addition, the similarity between two identical data is the same; the similarity (maximum similarity)

can be represented, but the subsequent clustering method must use the data of "maximum similarity" in similar matrices for clustering, yet clustering the identical data is insubstantial. Equation (6) does not calculate two identical data; the similarity between the uncalculated data is expressed as "0" in similar matrices (Equation (7)), to make subsequent clustering smooth and reasonable. On the other hand, in order to increase the clustering efficiency, the repeated calculation of similarity (e.g. $Sim(AH^{A}_{h,k,1}, AH^{B}_{h,k,2})$ and $Sim(AH^{A}_{h,k,2}, AH^{B}_{h,k,1})$) is processed as above.

$$SM[AH_{h,k,\bullet}] = \begin{bmatrix} 0 & Sim(AH^{A}_{h,k,1}, AH^{B}_{h,k,2}) & Sim(AH^{A}_{h,k,1}, AH^{B}_{h,k,3}) & \cdots & Sim(AH^{A}_{h,k,1}, AH^{B}_{h,k,b}) \\ 0 & 0 & Sim(AH^{A}_{h,k,2}, AH^{B}_{h,k,3}) & \cdots & Sim(AH^{A}_{h,k,2}, AH^{B}_{h,k,b}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \quad (7)$$

Step (4) Integrate data and determine cluster type. The related data must be completed and integrated before clustering to determine the cluster type. Therefore, this step integrates the required information according to the similarity implied in similar matrices. First, the data pairs with maximum similarity in similar matrices are integrated to obtain the set $MaxSet(SM[AH_{h,k,\bullet}])$ of data pairs with maximum similarity in similar matrices (Equation (8)). The data pair must contain at least one unclustered datum. The "data pair" refers to two data in similarity determination (e.g. $Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})$, the data pair is $(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})$).

$$MaxSet(SM[AH_{h,k,\bullet}]) = \begin{Bmatrix} (AH^{A}_{h,k,1}, AH^{B}_{h,k,2}), (AH^{A}_{h,k,1}, AH^{B}_{h,k,3}), (AH^{A}_{h,k,2}, AH^{B}_{h,k,3}), \ldots, (AH^{A}_{h,k,a}, AH^{B}_{h,k,b}) \\ | Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b}) \in Max(Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})) \\ and\ AH^{A}_{h,k,a}\ or\ AH^{B}_{h,k,b}\ has\ not\ been\ assigned\ to\ any\ cluster \end{Bmatrix} \quad (8)$$

Secondly, the data with minimum similarity that are not 0 in the similar matrices are collected. The minimum similarity value $CMin(SM[AH_{h,k,\bullet}])$ of the matching condition in the similar matrices can be obtained by Equation (9).

$$CMin(SM[AH_{h,k,\bullet}]) = Min(Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})) \quad Where\ Min(Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})) \neq 0 \quad (9)$$

Finally, the maximum similarity in similar matrices is compared with the minimum value of match condition to obtain the cluster type (CType) for subsequent clustering (Equation (10)). CType is divided into 1, 2 and 3, if the maximum similarity is not equal to or greater than the minimum similarity of match condition, CType=1; if the maximum similarity equals the minimum similarity of match condition, CType=2; if the maximum similarity is 0, CType=3.

$$CType = \begin{cases} 1,\ If\ Max(Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})) \neq CMin(SM[AH_{h,k,\bullet}]) \\ \quad and\ Max(Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})) > CMin(SM[AH_{h,k,\bullet}]) \\ 2,\ If\ Max(Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})) = CMin(SM[AH_{h,k,\bullet}]) \\ 3,\ If\ Max(Sim(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})) = 0 \end{cases} \quad (10)$$

Step (5) Divide the cluster of domain data. This step constructs related clustering rules (four items) according to the clustering algorithm proposed by Zhang et al. [19]. The precondition of clustering rules is determined by the cluster type (CType) obtained in Step (4). In addition, when all the data in the batch ($MaxSet(SM[AH_{h,k,\bullet}])$) are clustered (including determined data but not included in cluster), the similarity of all data pairs of this batch must be set as 0, followed by returning to Step (4) to obtain the new cluster data again, till all the data are clustered. The clustering rules and details are described below:

(1) When CType = 1, if the two data of data pair ($(AH^{A}_{h,k,a}, AH^{B}_{h,k,b})$) do not belong to the existing cluster ($CR'_{h,k,\bullet}$), the new cluster $CR_{h,k,c+n}$ is added to the data pairs of the matching condition, respectively (Equation (11)), where n value 1 represents the first new cluster, and n value 2 represents the second new cluster; the rest can be deduced accordingly.

$$
\begin{aligned}
&CR_{h,k,c+n} \\
&= \left\{ \begin{matrix} (AH^A_{h,k,1}, AH^B_{h,k,2}), (AH^A_{h,k,1}, AH^B_{h,k,3}), (AH^A_{h,k,2}, AH^B_{h,k,3}), \ldots, (AH^A_{h,k,a}, AH^B_{h,k,b}) \\ | CType = 1 \text{ and } AH^A_{h,k,a} \notin CR'_{h,k,\bullet} \text{ and } AH^B_{h,k,b} \notin CR'_{h,k,\bullet} \end{matrix} \right\} \\
&\text{and } Sim(AH^A_{h,k,a}, AH^B_{h,k,b}) = 0 \ \forall a,b \\
&\text{Where } (AH^A_{h,k,a}, AH^B_{h,k,b}) \in MaxSet(SM[AH_{h,k,\bullet}])
\end{aligned} \tag{11}
$$

(2) When CType=1, if the data ($AH^A_{h,k,a}$ or $AH^B_{h,k,b}$) of a data pair ($(AH^A_{h,k,a}, AH^B_{h,k,b})$) belongs to the existing cluster ($CR'_{h,k,\bullet}$), this data pair is united with the existing cluster to obtain the combined cluster $CR_{h,k,c}$ (Equation (12)). If any data belong to the new cluster ($CR_{h,k,c+n}$), this condition is false (i.e. not to merge this data pair).

$$
\begin{aligned}
&CR_{h,k,c} \\
&= \left\{ \begin{matrix} (AH^A_{h,k,1}, AH^B_{h,k,2}), (AH^A_{h,k,1}, AH^B_{h,k,3}), (AH^A_{h,k,2}, AH^B_{h,k,3}), \ldots, (AH^A_{h,k,a}, AH^B_{h,k,b}) \\ | CType = 1 \text{ and } AH^A_{h,k,a} \in CR'_{h,k,\bullet} \text{ or } AH^B_{h,k,b} \in CR'_{h,k,\bullet} \\ \text{and } AH^A_{h,k,a} \notin CR_{h,k,c+n} \text{ and } AH^B_{h,k,b} \notin CR_{h,k,c+n} \end{matrix} \right\} \cup CR'_{h,k,c} \\
&\text{and } Sim(AH^A_{h,k,a}, AH^B_{h,k,b}) = 0 \ \forall a,b \\
&\text{Where } (AH^A_{h,k,a}, AH^B_{h,k,b}) \in MaxSet(SM[AH_{h,k,\bullet}])
\end{aligned} \tag{12}
$$

(3) When Ctype=2, all the data not belonging to any cluster in this batch ($MaxSet(SM[AH_{h,k,\bullet}])$) are merged to form a new cluster $CR_{h,k,c+n+1}$ (Equation (13)).

$$
\begin{aligned}
&CR_{h,k,c+n+1} = \left\{ \begin{matrix} AH^A_{h,k,1}, AH^B_{h,k,2}, \ldots, AH^A_{h,k,a}, AH^B_{h,k,b} \\ | CType = 2 \text{ and } AH^A_{h,k,a} \notin CR'_{h,k,\bullet} \text{ or } AH^B_{h,k,b} \notin CR'_{h,k,\bullet} \end{matrix} \right\} \\
&\text{and } Sim(AH^A_{h,k,a}, AH^B_{h,k,b}) = 0 \ \forall a,b \\
&\text{Where } (AH^A_{h,k,a}, AH^B_{h,k,b}) \in MaxSet(SM[AH_{h,k,\bullet}])
\end{aligned} \tag{13}
$$

(4) When CType=3, all the data not belonging to any cluster are merged to form a new cluster $CR_{h,k,c+n+2}$ (Equation (14)).

$$
\begin{aligned}
&CR_{h,k,c+n+2} \\
&= \left\{ AH_{h,k,1}, AH_{h,k,2}, \ldots, AH_{h,k,i} \mid CType = 3 \text{ and } AH_{h,k,i} \notin CR'_{h,k,\bullet} \right\}
\end{aligned} \tag{14}
$$

Step (6) Define the cluster name. This step nominates the clusters obtained in Step (5), in order to display the "target member domain knowledge referability" obtained by the follow-up analysis more specifically. For the nomination of clusters, the clusters are expressed more specifically in this step, and the names are not duplicated. Therefore, the primary condition of nomination is to nominate the cluster according to the frequent item with the maximum length in the cluster and the name of the frequent item with the maximum occurrence frequency in other data of the cluster, to obtain the cluster name $Topic_{h,k,c}$ (Equation (15)). If the cluster has no frequent item matching condition, the frequent item of the cluster is classified into the candidate cluster name set ($CTSet\_Topic_{h,k,c}$), and processed in Equation (16).

$$
\begin{aligned}
&\text{If } N(CR_{h,k,c}) \neq 1 \text{ and } FI(AH_{h,k,\bullet}) \in ML(FI(AH_{h,k,\bullet}), CR_{h,k,c}) \\
&\text{and } FI(AH_{h,k,\bullet}) \in MN(FI(AH_{h,k,\bullet}), CR_{h,k,c}) \\
&\text{Then } FI(AH_{h,k,\bullet}) = Topic_{h,k,c} \\
&\text{Else } FI(AH_{h,k,\bullet}) \in CTSet\_Topic_{h,k,c} \\
&\text{Where } FI(AH_{h,k,\bullet}) \in CR_{h,k,c} \text{ and } CTSet\_Topic_{h,k,c} \notin Topic'_{h,k,\bullet}
\end{aligned} \tag{15}
$$

If the cluster has no frequent item meeting the primary condition (i.e. the cluster cannot be nominated directly), the cluster is nominated according to the frequent item with maximum occurrence frequency in the other data of the cluster. The determined name shall not be the existing cluster name ($Topic'_{h,k,\bullet}$) to meet the principle of non-repetitive name (Equation (16)).

$$If\ FI(AH_{h,k,\bullet}) \in MN(FI(AH_{h,k,\bullet}), CR_{h,k,c})\ and\ FI(AH_{h,k,\bullet}) \notin Topic'_{h,k,\bullet}$$
$$Then\ \ FI(AH_{h,k,\bullet}) = Topic_{h,k,c} \tag{16}$$
$$Where\ FI(AH_{h,k,\bullet}) \in CTSet\_Topic_{h,k,c}$$

Step (7) Obtain target member's domain knowledge referability data. This paper uses the vector space model, and the NGD algorithm proposed by Cilibrasi and Vitanyi [20] based on "text" and "semantic" similarity analysis, and integrates the cluster similarity calculation method proposed by Zhang et al. [19] to develop the integrated text and semantic similarity analyses of clusters for analyzing the virtual community member domain knowledge referability. Before the related analysis, the clusters obtained in Step (5) must be divided and defined as "$CR^{X}_{1,k,x}$" and "$CR^{Y}_{2,k,y}$" independent clusters according to the "virtual community" and "target virtual community member", and the contained clusters ($CR_{h,k,c}$) for subsequent calculation of cluster similarity. Secondly, in the text similarity analysis of clusters, this step uses the vector space model to calculate the similarity among data of clusters, and the corresponding similarities are added up according to the clusters of data; the values are then averaged according to the total number of data in the two clusters, to avoid the huge difference in the total number of data in the clusters resulting in asymmetric similarity determination result (Equation (17)). The two clusters refer to the clusters of domain data divided from "virtual community (h=1)" and "target virtual community member (h=2)". In addition, the set vector of data is created according to the word segmentation result of Equation (1) in Step (1).

$$AH^{A\,\omega}_{1,k,a} = \left[BW^{A}_{h,k,a,1}, BW^{A}_{h,k,a,2}, \cdots, BW^{A}_{h,k,a,v}\right]^{T}$$
$$AH^{B\,\omega}_{2,k,b} = \left[BW^{B}_{h,k,b,1}, BW^{B}_{h,k,b,2}, \cdots, BW^{B}_{h,k,b,w}\right]^{T}$$
$$SimVS(CR^{X}_{1,k,x}, CR^{Y}_{2,k,y}) = \frac{\sum_{AH^{A}_{1,k,a} \in CR^{X}_{1,k,x}} \sum_{AH^{B}_{2,k,b} \in CR^{Y}_{2,k,y}} \frac{AH^{A\,\omega}_{1,k,a} \bullet AH^{B\,\omega}_{2,k,b}}{\left\|AH^{A\,\omega}_{1,k,a}\right\| \cdot \left\|AH^{B\,\omega}_{2,k,b}\right\|}}{N(AH^{A}_{1,k,\bullet}) + N(AH^{B}_{2,k,\bullet})} \tag{17}$$
$$Where\ AH^{A}_{1,k,\bullet} \in CR^{X}_{1,k,x}\ \ and\ AH^{B}_{2,k,\bullet} \in CR^{Y}_{2,k,y}$$

In addition, this step utilizes the calculation method for semantic relation distance between two words via the "NGD algorithm"; it is improved and combined with the concept of cluster similarity calculation, developed into a method for analyzing the semantic relation distance between two clusters for analyzing the "semantic relation similarity" other than "text similarity" between two clusters (Equation (18)). The smaller the calculated value, the longer the semantic relation distance between two clusters. The "two clusters" in this part is defined in Equation (18). In addition, as this part analyzes the "semantic" relation between two clusters, the non-representative words (e.g. function words) in the data are excluded to avoid too much noise influencing the analysis result, and to obtain the representative semantic relation distance between two clusters. The representative words of the data are obtained from the Equation (2) screening result of Step (1).

$$SimNGD(CR^{X}_{1,k,x}, CR^{Y}_{2,k,y}) = \frac{\sum_{AH^{A}_{1,k,a} \in CR^{X}_{1,k,x}} \sum_{AH^{B}_{2,k,b} \in CR^{Y}_{2,k,y}} \frac{Max\{\log GSH(VW^{A}_{1,k,m}, AH^{A}_{1,k,a}), \log GSH(VW^{B}_{2,k,n}, AH^{B}_{2,k,b})\} - \log GSH((VW^{A}_{1,k,m}, AH^{A}_{1,k,a}), (VW^{B}_{2,k,n}, AH^{B}_{2,k,b}))}{\log GN\text{-}Min\{\log GSH(VW^{A}_{1,k,m}, AH^{A}_{1,k,a}), \log GSH(VW^{B}_{2,k,n}, AH^{B}_{2,k,b})\}}}{N(VW^{A}_{1,k,\bullet}) \cdot N(VW^{B}_{2,k,\bullet})} = \frac{}{N(AH^{A}_{1,k,\bullet}) + N(AH^{B}_{2,k,\bullet})} \tag{18}$$
$$Where\ VW^{A}_{1,k,m} \in AH^{A}_{1,k,a}\ \ and\ VW^{B}_{2,k,n} \in AH^{B}_{2,k,b}\ \ and\ AH^{A}_{1,k,\bullet} \in CR^{X}_{1,k,x}\ \ and\ AH^{B}_{2,k,\bullet} \in CR^{Y}_{2,k,y}$$

Finally, the results of Equations (17) and (18) are integrated; the target member's domain knowledge referability $MDR(D_{k,x}, (M_T, DK_{k,y}))$ is calculated after normalization (Equation (19)).

$$MDR(D_{k,x}, (M_T, DK_{k,y})) = \frac{\frac{SimVS(CR^{X}_{1,k,x}, CR^{Y}_{2,k,y})}{SimNGD(CR^{X}_{1,k,x}, CR^{Y}_{2,k,y})}}{\sum_{all\ h,y} \frac{SimVS(CR^{X}_{1,k,x}, CR^{Y}_{h,k,y})}{SimNGD(CR^{X}_{1,k,x}, CR^{Y}_{h,k,y})}}$$
$$Where\ h > 1 \tag{19}$$

4 A Virtual Community Member's Referability Determination System

According to the proposed methodology, this paper develops a Web-based virtual community member's referability determination system to validate the feasibility of the model. Under this system, users are divided into the common user and the system administrator to execute different functions. Firstly, the common users can upload members' historical publication data (see procedure (A1) of Figure 4) for system determination. Secondly, the system administrator can execute the kernel functions including member data clustering and member referability analysis functions (see procedure (B3) and (B4)). After that, this system analyzes the similarity between clusters of a target member's domain publications based on the target virtual community member's domain of historical publication, and compares the similarity data with those of other members. The target virtual community member's domain knowledge referability can be obtained by normalization at last. Finally, system administrator can receive the members' referability determination results and give some feedback for system training and common users can query the system inference result (see procedure (A2)).

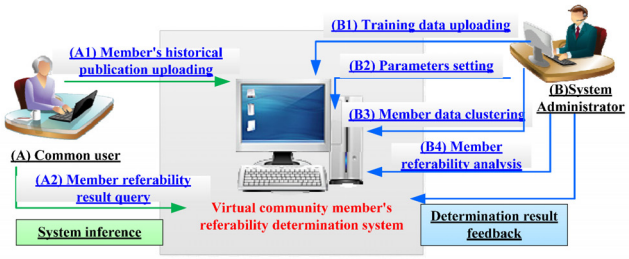

Figure 4. The architecture of developed system

4.1 System Analysis Data Collection

Before this system is executed, the system administrator must collect the training files of related domains from the "soso Q&A" virtual community (http://wenwen.sogou.com) (as shown in Figure 5); the questions and answers are extracted from domain training files and imported into the system. The set of domanial keywords as the basis of domanial cluster relation distance analysis can be created and the member's domain knowledge referability can be determined.

In addition, the common user collects the question and answer data of undetermined virtual community discussion threads before using this system to determine the match of questions and answers in the discussion threads and the member's referability. For case validation, the discussion threads data are collected from the "soso Q&A" virtual community (as shown in Figure 6) as the samples of validation data to test the system feasibility and performance.

Figure 5. The knowledge document collected from "soso Q&A"

Figure 6. Discussion threads data collected from "soso Q&A"

4.2 System Application

4.2.1 Member Data Clustering Function

The authorized user executes the member data clustering function, and selects the "total number of publications in ascending order" and clicks "arrange", the system displays the names of all community members and the total number of publications (e.g. "galant7072" and "11"). The user can review the selected member's detailed historical publication data via the hyperlink of "publication content query". If there is no publication title, the publication is of "answer content" (as shown in Figure 7). Afterwards, when the user executes "display member's historical publication domain distribution pattern" function, the system counts the domain distribution pattern of historical publications of the selected member "galant7072"; for example, the member released "8" articles in the "camera and photography" domain. When the user selects "camera and photography" and clicks "confirm and preprocess data" (as shown in Figure 8), the system carries out data preprocessing of "word segmentation" and "valid terms (AWT) screening" for all of the target member's historical publications in the "camera and photography" domain. Secondly, when the user clicks "integrate valid terms (AWT) set of [camera and photography domain]" (as

shown in Figure 9), the system eliminates the repeated words, integrates the "camera and photography domain" valid terms (AWT) set according to the target member, and displays all "8" valid terms (AWT) integration information of domain historical publications for the user. When the user enters "25" MinSup, and clicks "screen frequent item set" (as shown in Figure 10), the system calculates the occurrence frequency of various item sets in the historical publications, and screens out the frequent item set as "Sensoji Temple" according to the MinSup "25%". Afterwards, when the user clicks "construct similar matrices and determine cluster type" (as shown in Figure 11), the system analyzes the similarity between publications based on the intersection result of frequent item sets according to the target member's total publications in the domain, for example, the similarity between publications No. "1" and "2" is 0, and the maximum similarity is "1", and the nonzero minimum similarity is "1", so that the present cluster type (CType) is "2". When the user clicks "divide cluster of domain data" (as shown in Figure 12), the system classifies the publications No. "4" and "10" as the same cluster according to the cluster type "CType=2" determined in the previous step, and replaces the original similarity data "Sim (4,10)=1" by "Sim (4,10)=0"; meanwhile, the system obtains the maximum similarity "1" from the new similar matrix after the similarity data are replaced, and when the nonzero minimum similarity is "null (no value meeting condition)", the cluster type (CType) is determined as "3". In addition, as a part of member data has not been clustered, when the user clicks "resume clustering" (as shown in Figure 13), the system executes clustering again according to the cluster type in the previous step, to classify the publications No. "1", "3", "5", "6", "7" and "11" as the same cluster. All data are clustered; when the user clicks "integrate final clustering result" (as shown in Figure 14), the system integrates the clustering result of the historical publications of the target community member "galant7072" in the "camera and photography domain", and displays the publication content of publication No. "1" belonging to "Cluster 2", and the clustering information of all publications of the target community member for the user's reference. In addition, when the user clicks "display cluster distribution diagram", the system displays the domain historical publication cluster distribution result. For example, the clustering result contains "Cluster1" and "Cluster2", and the "Cluster1" contains historical publications No. "4" and "10" (as shown in Figure 14).

Figure 7. Query member's historical publication data

Figure 8. Count the domain distribution pattern of historical publications

Figure 9. Data preprocessing of member's historical publication data

Figure 10. Screen out the frequent item set

Figure 11. Construct similar matrices and determine cluster type

Figure 12. Divide cluster of domain data (1)

Figure 13. Divide cluster of domain data (2)

Figure 14. Integrate the final clustering result

4.2.2 Member Referability Analysis Function

The authorized user executes the member referability analysis function, and selects "total number of publications in ascending order" and clicks "arrange"; the system displays the names of all the community members, the total number of publications as "galant7072" and "11" information. The user can click the hyperlink of "publication content query" to review the selected member's detailed historical publication information. If there is no publication title, the publication is of "answer content" (as shown in Figure 15). Afterwards, when the user uses "display member's historical publication domain distribution pattern", the system calculates the domain distribution pattern of the historical publications according to the selected member "galant7072", for example, the member released "8" articles in the "camera and photography" domain. When the user selects "camera and photography" and clicks "integrate clustering information" (as shown in Figure 16), the system integrates the clustering result of the "camera and photography" domain after the clustering of target community member by the "member data clustering function", and calculates the occurrence frequency of the frequent item set in the cluster; for example, "Sensoji Temple" and "restaurant" occurred "twice". When the user clicks "define cluster name" (as shown in Figure 17), the system defines and obtains the cluster name of cluster No. "1" as "Sensoji Temple, restaurant" according to the relationship between the frequent item set and the cluster. When the user clicks "integrate domain knowledge referability discussion threads information" (as shown in Figure 18), the system integrates all the discussion threads in the target domain according to the user selected domain class, and executes data preprocessing of word segmentation and screening valid terms (AWT). Afterwards, when the user clicks "calculate text similarity between discussion threads and domain historical publication cluster" (as shown in Figure 19), the system uses the vector space model to calculate the text similarity "0.81626" between cluster No. "1" and discussion thread No. "1", and integrates the text similarity between all clusters and discussion threads; the overall text similarity between the target member's domain historical publications and domain discussion threads is "0.35933". On the other hand, when the user clicks "divide valid terms combination" (as shown in Figure 20), the system provides the interface for valid terms combination input in cluster units. When the user clicks "input search results" and clicks "calculate semantic relation distance between discussion threads and domain historical publication cluster", the system calculates the semantic relation distances "0.28336" and "0.3641", and the overall semantic relation distance "0.32373" (as shown in Figure 21). Finally, when the user clicks "calculate target community member's domain knowledge referability", the system calculates the domain knowledge referability score "1.10997" of target community member "galant7072" in "motion" domain type, and the domain knowledge referability "0.21268" (as shown in Figure 22).

Figure 15. Query member's historical publication data

Figure 16. Integrate the clustering result of historical publications

Figure 17. Define cluster name

Figure 18. Data preprocessing of domain discussion threads

Figure 19. Calculate text similarity between discussion threads and domain historical publication cluster

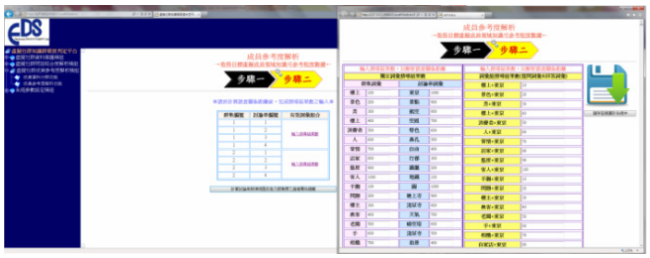

Figure 20. Divide valid terms combination

Figure 21. Calculate semantic relation distance between discussion threads and domain historical publication cluster

Figure 22. Calculate member's domain knowledge referability

5 Case Study

5.1 Collection and Construction of Validation Data

This paper selected the "soso Q&A" virtual community (http://wenwen.sogou.com) as the validation data to measure the "virtual community member's referability" determination result performance and to validate the feasibility of this model. The "set of domanial keywords" must be trained and constructed before the system performance is validated. The knowledge channels in the "soso Q&A" virtual community contain the "essential knowledge" of various domains shared by numerous common community members and periodically released by official administrators. The domain knowledge matches colloquial knowledge sharing of virtual community members, and the knowledge content and domain category are constructed by the manual approval of most members, so the "essential knowledge" has considerable knowledge quality and classification accuracy. Based on the advantages, this paper collects the "essential knowledge" in the "soso Q&A" virtual community (as shown in Figures 23 and 24), and refers to the classification of knowledge as domain training data to construct the "set of domain keywords" meeting the colloquial expression characteristic of virtual community members, and to obtain more accurate determination and validation results.

Figure 23. Essential knowledge in "soso Q&A" virtual community (1)

Figure 24. Essential knowledge in "soso Q&A" virtual community (2)

5.2 Description of Verification Procedure

System validation implementation description First, five community members who had answered over 200 questions are selected randomly from various domains in "soso Q&A" virtual community [21]. There are 10 domains and 50 community members regarded as the subjects of system performance validation. The community member referability in various domain categories is ranked (i.e. No. 1 to No. 5) according to the community members' "comprehensive reputation" (provided by the "soso Q&A" virtual community) in domain categories (as shown in Table 1). Secondly, 10 data are collected randomly from the domains of 50 community members' domain historical publication content and imported into the system; there are 500 data as the basic test data on the subjects. In addition, this paper collected 680 pieces of question and answer content with considerable quality and domain concept from the "soso Q&A" virtual community (as shown in Table 2).

Table 1. The summarized data of subject (Computer/Digitals domain) (Partial data)

Domain Category		**Computer/digitals**
Member (Respondent)		Wang **B**eat
Comprehensive Reputation		**378**
Domain Ranks	No. of Reply	Content of Reply (Random selection)
1	1	LAN router is set up for network sharing equipment, general connections and …
	2	There are two ways to achieve: 1. Change the second stage connection router …
	3	Error 769 Tip: disable the network card, computer card usually without success ...
	4	Your computer may be configured with Ethernet NIC speed on the associated equipment is not ...
	5	Limited or no network connection has the following possibilities: 1 which shows your computer network ...

Table 2. Member's referability determination training data (Partial data)

No.	Domain Category	Question content	Best answer	Approval/ disapproval	Review Result A	Review Result B
1	Olympics/ sports	Where is 8B of WWE?	Batista let WWE last May, and now…	13/0	High quality	Non-high Quality
2	Society/ humanities	Why clinks in drinking?	Why clinks in drinking? There are two versions; one is that ancient Greek…	1491/4	High quality	High quality
3	Computer/ digitals	How to keep the mobile phone powered when you are out?	Smart phone is really handy, but battery is a real problem. Usually…	82/51	Non-high quality	High quality

There are two judgment standards: manual review (reviewed by two volunteers), and reference to the ratio of approval number to disapproval number (clicked by other community members), for Cohen's Kappa statistics. The two reviewers and other community members are regarded as two observers to evaluate the consistency of different observers' judgment on the same event (measure answer quality). The Kappa statistic is kept at 0.97 [22]. Afterwards, 200 discussion threads are selected randomly from the aforesaid 680 discussion threads as training data, regarded as Stage I validation, imported into the system one by one to obtain the "set of domanial keywords" as the analysis reference of the member referability corpus. The "Recall Rate" and "Accuracy Rate" indexes are used to evaluate the performance of this model in the "virtual community member's referability determination." Afterwards, when the aforesaid Stage I validation is completed, the rest of the 480 webpage knowledge files are imported into the system in 6 periods in Stage II. There are 80 discussion threads data imported in each period. Finally, the "virtual community member's referability determination" is inferred again by the aforesaid selected 50 subjects in each period, in order to analyze the long-term learning trend of the system with different training data volumes.

5.2.1 Definition of Evaluation Indices

(1) Kappa Statistic of Domain Training Data

In order to further match the perspective of "wisdom of crowds" emphasized in this paper, the Kappa value application method of Toba et al. [22] is improved. Two manual reviewers and the ratio of approval number to disapproval number (trend of crowd determination) are regarded as two observers to evaluate the consistency of different observers' judgments on the same event (measure answer quality). When the two manual reviewers simultaneously mark the target discussion thread as high quality knowledge or the ratio of approvals to disapprovals is higher than 80%, the observer judges (marks) the target discussion thread as "high quality knowledge". This is the ratio of "observed consistency" to "expected consistency". The Kappa statistic of 680 validation data collected in two stages is 0.97.

(2) Recall Rate of Member Domain Referability Ranking

The recall rate ($W = \frac{p}{u}$) of member domain referability ranking is a relative ratio; it is the ratio of "the number of actual member domain referability rankings matching the inferred member domain referability rankings" (P) to "the total ranking number of actual member domain referability" (u).

(3) Accuracy Rate of Member Domain Referability Ranking

The accuracy rate ($T = \frac{p}{s}$) of member domain referability ranking is a relative ratio; it is the ratio of "the number of actual member domain referability rankings matching the inferred member domain referability rankings" (P) to "the total ranking number of inferred member domain referability" (s).

5.3 System Validation Result Analysis

5.3.1 Stage I Validation Result Analysis (50 Subjects)

The training data base are 200 webpage knowledge files, the system judges 20 test data, the average recall rate of member referability determination is 26%, and the average accuracy rate of member referability determination is 26%. The number of actual rankings matching the system inferred rankings is 6. The distribution trend of the average recall rate and accuracy rate of the virtual community Q&A match determination is shown in Table 3.

Table 3. The performance evaluation result at Stage 1 (200 samples of training data)

Domain/category	(A) Number of actual member domain referability rankings	(B) Number of inferred member domain referability rankings	(A) and (B) matching Numbers	Recall Rate	Accuracy Rage
Computer/Digital	5	5	2	40%	40%
Life/Housing	5	5	1	20%	20%
Olympics/Sports	5	5	1	20%	20%
Recreation/Hobby	5	5	1	20%	20%
Arts/Literature	5	5	1	20%	20%
Society/humanities	5	5	1	20%	20%
Education/Science	5	5	1	20%	20%
Health/Medicine	5	5	2	40%	40%
Commerce/Finance	5	5	2	40%	40%
Entertainment/Star	5	5	1	20%	20%
Average				26%	26%

According to Table 3, the validation results of Q&A match determination in Stage I show that the "recall rate" and "accuracy rate" of member referability are 20% to 40%, and the overall mean is about 26%. Therefore, according to the recall rate and accuracy rate in the validation results of Stage I, the precision and performance of member referability determination are poor; the domain referability determination cannot be judged accurately.

5.3.2 Stage II Validation Result Analysis (50 Subjects)

Stage II validation is divided into six periods, and 80 domain training files are imported in each period to observe the variation of validation indicator in various periods with the increase in training files. The validation results of various periods are summarized in Table 4.

Table 4. The performance evaluation result at Stage I and Stage II

Member's referability determination		# of training data							Ave.
		Stage 1	Stage 2						
		1st period 200	2nd period 280	3rd period 360	4th period 440	5th period 520	6th period 600	7th period 680	
Recall Rate	Ave.	26%	44%	58%	68%	76%	82%	82%	59%
	GR.	-	18%	14%	10%	8%	6%	0%	9%
Accuracy Rate	Ave.	26%	44%	58%	68%	76%	82%	82%	59%
	GR.	-	18%	14%	10%	8%	6%	0%	9%

Note: GR = Growth Rate

According to Table 4, in units of 80 domain training files increasing each period, the overall growth rate of the recall rate and accuracy rate validation indices of member referability determination each period is 8% and 9%, respectively. For the validation results of the final period 7 (680 training data imported), the recall rate and accuracy rate increased from 26% in the 1st period to 82%. To sum up the validation results, the virtual community member's referability determination system developed in this paper has learning ability and considerable correctness. That is, the virtual community member's referability determination model and system can provide the users with accurate member referability.

Finally, the validation results of the two above-mentioned stages are integrated and the results of various validation indices are compiled in Table 5. According to Table 5, the "average growth rate per period before convergence" and "overall average growth rate per period" of various validation indices are positive, and various validation indices converge in the 7th period. Therefore, taking the random validation data in this paper as an example, when the system uses about 600 training knowledge articles, the performance level of the inference indicators of system can be increased to 82%. Generally speaking, the system performance grows continuously with the periods and training load, and eventually reaches stable and good performance level.

Table 5. The compiled results of various validation indexes

Validation indexes	Average	Convergence period	Average growth rate per period before "convergence"	Overall average growth rate per period
Recall Rate	59%	7th	11.2%	8.7%
Accuracy Rate	59%	7th	11.2%	8.7%

5.4 Qualitative Analysis of Validation Results

There is considerable difference between this paper and "soso Q&A" virtual community in the member referability determination "method" and "result". In order to analyze and compare them integrally to evaluate the actual contribution, this paper analyzes of data analysis result to figure out the dissimilarity. The overall performance is concluded by discussing the overall difference qualitatively and the data analysis results are summarized in Table 6.

Table 6. The determination performance of "soso Q&A" virtual community and this model

Domain Category	Subjects	Historical Publications		Total diff.	Referability		Difference
		This Paper	soso Q&A		This Paper	soso Q&A	
Computer	Wang	10	59665	59655	76%	31%	45%
Life	Gu Shunjun	10	36781	36771	70%	24%	46%
Sport	Stinger	10	15951	15941	67%	22%	45%
Hobby	Qinghua	10	19943	19933	65%	23%	42%
Art	Wei	10	30242	30232	86%	37%	49%
Social	Lu	10	37106	37096	78%	33%	45%
Education	Binghua	10	14671	14661	75%	30%	45%
Medical	Meteor	10	33232	33222	71%	26%	45%
Business	Du Jiang	10	27246	27236	61%	18%	43%
Entertainment	Mclass	10	16944	26934	61%	28%	33%
Average		10	29178	30168	71%	27%	43%
Standard Deviation					7.94%	5.71%	

According to Table 6, there are three points can be discussed. Firstly, the "difference in total number of adopted publications" between this model and "soso Q&A" virtual community is considerably large. This model can execute determination by adopting 10 historical publications on average, whereas "soso Q&A" virtual community adopts 29178 historical publications on average. Secondly, for "domain referability", the determination result of this system is 71% on average, whereas the calculated result of "soso Q&A" virtual community is 27% on average. Thirdly, for "domain referability standard deviation", the gap between the standard deviations of this system and "soso Q&A" virtual community is slight.

6 Conclusion

The knowledge and information shared by the knowledge contributors are spread and circulated via the Internet. The existing operating mechanism can continuously accumulate public knowledge for a virtual community, but there will still be some problems in the long run. The problems are listed below:

- The knowledge demanders find it difficult to determine the community member referability. Because the community's incentive system is difficult to perfect, the control of knowledge contributors' sharing behavior is difficult to implement due to the information explosion.
- It is difficult to maintain the knowledge contributors' willingness to continuously share knowledge [23]. Due to the challenges to the members' sharing behavior control and the community's incentive mechanism, fewer professional knowledge contributors can obtain the community status identical with professional members while releasing less referable information.

Considering the above problems, this paper proposes the "virtual community member's referability determination" model to measure the community member's referability. The model built in this paper uses the clustering algorithm proposed by Zhang et al. [19] as the basis of data clustering, and executes "word segmentation and screen valid terms (AWT)", "screen frequent item set", "build similar matrices", "determine cluster type" and "divide cluster of domain data" to obtain "virtual community's domanial discussion threads distribution cluster" and "member's domanial historical publication distribution cluster". Secondly, the corresponding clusters are nominated according to the cluster nomenclature proposed by Zhang et al. [19] to obtain representative cluster names. Finally, the vector space model and NGD algorithm proposed by Cilibrasi and Vitanyi [20] are integrated and improved, combined with the cluster similarity calculation concept proposed by Zhang et al. [19] to analyze the similarity relationship between clusters of virtual community's domanial discussion threads and member's domanial historical publications, in order to obtain "member's domain knowledge referability" as the ultimate objective of this model. Generally speaking, this paper remedies the defects in the existing community's incentive mechanism, reduces the continuous circulation of less referable information, promotes more members to share knowledge, and increases the accumulation of professional public knowledge to activate the virtual community utilization [], so as to catalyze the sustainable overall development of virtual communities.

Acknowledgements

This research is supported in part by the Ministry of Science and Technology, Taiwan under Grant: MOST 107-2221-E-035 -086 -MY2.

References

[1] Y.-H. Fang, C.-M. Chiu, In Justice We Trust: Exploring Knowledge-Sharing Continuance Intentions in Virtual Communities of Practice, *Computers in Human Behavior*, Vol. 26, No. 2, pp. 235-246, March, 2010.

[2] H.-T. Tsai, P. Pai, Explaining Members' Proactive Participation in Virtual Communities, *International Journal of Human-Computer Studies*, Vol. 71, No. 4, pp. 475-491, April, 2013.

[3] S.-W. Hung, M.-J. Cheng, Are You Ready for Knowledge Sharing? An Empirical Study of Virtual Communities, *Computers & Education*, Vol. 62, pp. 8-17, March, 2013.

[4] C.-J. Chen, S.-W. Hung, To Give or to Receive? Factors Influencing Members' Knowledge Sharing and Community Promotion in Professional Virtual Communities, *Information & Management*, Vol. 47, No. 4, pp. 226-236, May, 2010.

[5] X.-L. Jin, Z. Zhou, M. K. O. Lee, C. M. K. Cheung, Why Users Keep Answering Questions in Online Question Answering Communities: A Theoretical and Empirical Investigation, *International Journal of Information Management*, Vol. 33, No. 1, pp. 93-104, February, 2013.

[6] J. Moskaliuk, J. Kimmerle, U. Cress, Collaborative Knowledge Building with Wikis: The Impact of Redundancy and Polarity, *Computers & Education*, Vol. 58, No. 4, pp. 1049-1057, May, 2012.

[7] P. D. Meo, A. Nocera, G. Terracina, D. Ursino, Recommendation of Similar Users, Resources and Social Networks in a Social Internetworking Scenario, *Information Sciences*, Vol. 181, No. 7, pp. 1285-1305, April, 2011.

[8] M. Maleszka, B. Mianowska, N. T. Nguyen, A Method for Collaborative Recommendation Using Knowledge Integration Tools and Hierarchical Structure of User Profiles, *Knowledge-Based Systems*, Vol. 47, pp. 1-13, July, 2013.

[9] Y.-M. Li, T.-F. Liao, C.-Y. Lai, A Social Recommender Mechanism for Improving Knowledge Sharing in Online Forums, *Information Processing & Management*, Vol. 48, No. 5, pp. 978-994, September, 2012.

[10] X. Ni, Y. Lu, X. Quan, L. Wenyin, B. Hua, User Interest Modeling and Its Application for Question Recommendation in User-Interactive Question Answering Systems, *Information Processing & Management*, Vol. 48, No. 2, pp. 218-233, March, 2012.

[11] J. Caverlee, L. Liu, S. Webb, The SocialTrust Framework for Trusted Social Information Management: Architecture and Algorithms, *Information Sciences*, Vol. 180, No. 1, pp. 95-112, January, 2010.

[12] I. Cantador, P. Castells, Extracting Multilayered Communities of Interest from Semantic User Profiles: Application to Group Modeling and Hybrid

Recommendations, *Computers in Human Behavior*, Vol. 27, No. 4, pp. 1321-1336, July, 2011,

[13] Y.-J. Chen, H.-C. Chu, Y.-M. Chen, C.-Y. Chao, Adapting Domain Ontology for Personalized Knowledge Search and Recommendation, *Information & Management*, Vol. 50, No. 6, pp. 285-303, September, 2013.

[14] D. Schall, Expertise Ranking Using Activity and Contextual Link Measures, *Data & Knowledge Engineering*, Vol. 71, No. 1, pp. 92-113, January, 2012.

[15] Y. A. Kim, M. A. Ahmad, Trust, Distrust and Lack of Confidence of Users in Online Social Media-Sharing Communities, *Knowledge-Based Systems*, Vol. 37, pp. 438-450, January, 2013.

[16] Y. A. Kim, R. Phalak, A Trust Prediction Framework in Rating-Based Experience Sharing Social Networks without a Web of Trust, *Information Sciences*, Vol. 191, pp. 128-145. May, 2012.

[17] N. Korovaiko, A. Thomo, Trust Prediction from User-item Ratings, *Social Network Analysis and Mining*, Vol. 3, No. 3, pp. 749-759, June, 2013.

[18] D.-R. Liu, Y.-H. Chen, W.-C. Kao, H.-W. Wang, Integrating Expert Profile, Reputation and Link Analysis for Expert Finding in Question-Answering Websites, *Information Processing & Management*, Vol. 49, No. 1, pp. 312-329, January, 2013.

[19] W. Zhang, T. Yoshida, X. Tang, Q. Wang, Text Clustering Using Frequent Itemsets, *Knowledge-Based Systems*, Vol. 23, No. 5, pp. 379-388, July, 2010.

[20] R. L. Cilibrasi, P. M. B. Vitanyi, The Google Similarity Distance, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 3, pp. 370-383, January, 2007.

[21] X. Liu, T. Jiang, F. Ma, Collective Dynamics in Knowledge Networks: Emerging Trends Analysis, *Journal of Informetrics*, Vol. 7, No. 2, pp. 425-438, April, 2013.

[22] H. Toba, Z.-Y. Ming, M. Adriani, T.-S. Chua, Discovering High Quality Answers in Community Question Answering Archives Using a Hierarchy of Classifiers, *Information Sciences*, Vol. 261, pp. 101-115, March, 2014.

[23] H.-F. Zhang, F. Xiong, Y. Liu, H.-C. Chao, Modeling and Analysis of Information Dissemination in Online Social Networks, *Journal of Internet Technology*, Vol. 16, No. 1, pp. 1-10, January, 2015.

Biography

Shih-Ting Yang is an associate professor in the Department of Industrial Engineering and Systems Management at Feng Chia University. Dr. Yang received his Ph.D. in Industrial Engineering and Engineering Management at National Tsing-Hua University and his research interests are knowledge management, data mining and text mining.