The Journal of Systems and Software 85 (2012) 87–101

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Keyword clustering for user interest profiling refinement within paper recommender systems

Xiaoyu Tang, Qingtian Zeng*

College of Information Science and Engineering, Shandong University of Science and Technology, No. 579 Qianwangang Road, Qingdao 266510, PR China

ARTICLE INFO

Article history:
Received 5 October 2010
Received in revised form 8 July 2011
Accepted 13 July 2011
Available online 23 July 2011

Keywords:
Weighted keyword graph
Keyword clustering
User interest profiles
Recommender systems
Ontology extension

ABSTRACT

To refine user interest profiling, this paper focuses on extending scientific subject ontology via keyword clustering and on improving the accuracy and effectiveness of recommendation of the electronic academic publications in online services. A clustering approach is proposed for domain keywords for the purpose of the subject ontology extension. Based on the keyword clusters, the construction of user interest profiles is presented on a rather fine granularity level. In the construction of user interest profiles, we apply two types of interest profiles: explicit profiles and implicit profiles. The explicit profiles are obtained by relating users' interest-topic relevance factors to users' interest measurements of these topics computed by a conventional ontology-based method, and the implicit profiles are acquired on the basis of the correlative relationships among the topic nodes in topic network graphs. Three experiments are conducted which reveal that the uses of the subject ontology extension approach as well as the two types of interest profiles satisfyingly contribute to an improvement in the accuracy of recommendation.

1. Introduction

Recommender systems form or work from a specific type of information filtering system technique that attempts to recommend information items that are likely to be of interest to the user. Typically, a recommender system compares a user profile to certain reference characteristics and seeks to predict the 'rating' that a user would give to an item they had not yet considered (Wikipedia, 2010). The current generation of recommendation methods is usually classified into the following three main categories: content-based, collaborative, and hybrid recommendation approaches (Adomavicius and Tuzhilin, 2005). Among the main techniques of recommender systems, user profiling forms the basis of such systems. At the same time, much work has been done in academia and industry on creating user profiles since the emergence of recommender systems. For example, VSM-based profiling approaches are a basic way of conducting user interest modeling (Balabanović and Shoham, 1997). In GroupLens, profiles of all users are aggregated in a user-item rating matrix, in which each line corresponds to a user, and each column to an item (Resnick et al., 1994). Collaborative recommender systems produce recommendations based on the heuristic that people who agreed in the past will probably agree again. There are also case-based approaches like CASPER (Smyth et al., 2002), Entrée system (Burke et al., 1997), and some other profiling methods based on artificial neutral network, such as Tan et al. (1998), Shepherd et al. (2002) and Kim et al. (2002).

The utilization of ontologies in user profiling techniques has gained much attention since it allows inference to be employed, enabling interests to be discovered that were not directly observed in the user's behavior (Middleton et al., 2004; Zeng et al., 2009; Wu et al., 2009). Additionally, once profiles are represented using an ontology, they can communicate with other ontologies which share similar concepts, which contributes to knowledge reuse and abates the effect of the cold-start problem (Felden and Linden, 2007; Middleton et al., 2004). In the Foxtrot system (Middleton et al., 2004), researchers improved recommendations based on the ontological profiling method by visualizing the profiles themselves and presenting them to users. Blanco-Fernández et al. (2008) used a domain ontology in which the semantic descriptions of the available items (e.g., TV programs, books, CDs, etc.) are formalized. Their reasoning-based recommendation strategy discovers semantic relationships between the users' preferences and the items available in the domain ontology. These relationships provide the system with extra knowledge about the users' interests, thus favoring more accurate personalization processes.

In this paper, we propose a refined ontological profiling method based on our extended subject ontology. Two kinds of interest profiles are introduced: explicit profiles and implicit profiles. The

* Corresponding author.
E-mail addresses: lonewar@163.com (X. Tang), qtzeng@163.com, qtzeng@sdust.edu.cn, qingtianzeng@hotmail.com (Q. Zeng).

0164-1212/$ – see front matter. Crown Copyright © 2011 Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2011.07.029

construction algorithms for profiling explicit interests and for profiling implicit interests are also proposed. Based on our extended subject ontology, we take into account the pertinence of keywords and the classifications of the ontology in the calculation of explicit interest profiles. The method for inferring users' potential interests (implicit interests) proposed employs a quantified evaluation of relationships between classifications in the ontology.

The paper is organized into seven sections. Section 2 discusses the related work. Section 3 presents a brief description of our recommender system, based on which we evaluate our approach. In Section 4, we propose a way of clustering keywords for an existing ontology, in order to create a more detailed and accurate structure of the existing ontology. In Section 5, the approaches to user interest profiles, which contain both explicit and implicit interest profiles, are presented. Section 6 introduces our prototype system called SPRS, and using this system we conducted three experiments. We evaluate the subject ontology extension method and refinement of user profiling approach. Section 7 concludes the paper and discusses future work.

2. Related work

Ontology is a conceptualization of a domain into a human-understandable but machine-readable format consisting of entities, attributes, relationships, and axioms (Guarino and Giaretta, 1995). It is used to alleviate the communication problems between systems due to ambiguous usage of different terms. For building user profiles, ontologies are used to address the so-called "cold-start problem" (Middleton et al., 2002; Susan and Gauch, 2004). Cantador et al. (2008) mapped social tagging information from multiple sources to their ontological structures that described the domains of interest covered by the tags, in order to build user profiles. Moreover, Middleton et al. (2004) used a term ontology to refer to the classification structure and instances within a knowledge base, representing the profiles in terms of a research paper topic ontology, which allows other interests to be inferred that go beyond those only seen in directly observed behavior. They also employ profile visualization to acquire profile feedback from users to improve profiling accuracy.

A number of strategies have been implemented to facilitate the construction of ontology. For example, Mika (2007) extended the traditional bipartite model of ontologies with the social dimension, leading to a tripartite model of actors, concepts and instances. He also demonstrated the application of this representation by showing how community-based semantics emerges from this model through a process of graph transformation. In addition, Zhang et al. (2010) proposed a suite of ontology metrics, at both the ontology-level and the class-level, to measure the design complexity of ontologies.

Finally, in recent years complex network and other forms of network models such as bipartite network have been considered to be important aspects of recommender systems by many researchers (Zanin et al., 2008; Zhou et al., 2007, 2010). When focusing on the problem of recommending items to a user, the underlying transaction data can be seen as a bipartite network, in which users and items are represented as two groups of nodes, connected to each other by certain links (Zanin et al., 2008). In order to utilize the bipartite network, a one-mode projecting method is usually implemented as an alternative to using the bipartite network directly. Zhou et al. (2007) raised a novel one-mode projecting method to compress the bipartite network and better preserve the original information. Zhou et al. (2010) introduced and used evaluation criteria for the diversification of recommendation, aside from accuracy. They believe the next generation of information filtering methods should focus on not only precision but also diversification, and that a balance between them should be sought.

There are some problems in the existing ontologies (Adomavicius and Tuzhilin, 2005; Correa da Silva et al., 2002) and the user interest profiling methods based on them (Middleton et al., 2001, 2003, 2004; Cantador et al., 2008; Susan and Gauch, 2004; Felden and Linden, 2007). (1) Most of the ontologies in use are framed in coarse granularity, making the classification of items obscure and undetermined. Therefore, the effectiveness of user profiling techniques based on such ontologies, no matter how sophisticated the interest profiling algorithms are, would deteriorate because of the coarse classification. (2) Typically, ontologies are usually predefined manually and remain fixed during a certain period of time, which makes them insensitive to new changes. When new research subjects emerge or other changes happen, the modifications must be done by their creators manually. This reaction is slow and tardy and needs human involvement. Moreover, typical ontology-based user interest calculation algorithms compute the interest value on each category for every user; however, these methods are not always effective. For instance, assume that two users, user *A* and user *B* have both viewed a certain number of papers in the subject "artificial intelligence" and we obtain the same interest value for their interests in this subject; therefore we believe their interests in "artificial intelligence" are no different. However, in fact some papers viewed by user *A* are about "support vector machine", and the other papers viewed by this user are about "genetic algorithm", whereas the papers viewed by user *B* are all related to "natural language processing". In this case, the relatively subtle differences between the users' interests in "artificial intelligence" are not discovered, leading to the description of user interests and the subsequent recommendation being inaccurate. (3) It is difficult for conventional approaches to differentiate the items within the same class (or subject). That is, in the case of research paper recommendation, different papers in the same subject contribute essentially equally to the construction of user profiles and the differences in textual information between papers is neglected, which is evidently unreasonable. (4) The inference of topics of interest via ontological relations between topics that have not been browsed explicitly by users is not precise enough. For instance, assume that there are three subclasses *A*, *B* and *C* under their immediate super class *A*, and a user has an interest in subclass *A* with an interest value of 0.4. With the aforementioned conventional profiling method, we get 0.2 for the user's interest value of the super class *A*. In this scenario, the user will receive equivalent recommendations from subclass *B* and *C*, because the conventional profiling method does not take into account the differences between subclass *B* and *C*, leading to inaccuracy of the inference about a user's implicit interest because these differences may be relevant to the user's preferences.

3. Framework for generating user interest profiles

In this section, we first present the framework for generating user interest profiles within online paper recommender systems, which is shown in Fig. 1. The main components in the framework include:

(1) *Subject ontology*. The subject ontology is predefined by domain experts with suitable granularity and scale, which presents the organization structure of domain knowledge and serves as the taxonomy for research papers. Moreover, it is the basis of the user profile. To refine the subject ontology, we will discuss our method of extending the subject ontology through automatically clustering weighted keyword graph in Section 4.

(2) *Paper management module*. Users can upload, browse, download and comment on any research papers through the user interface of the paper management module. All of the research papers are

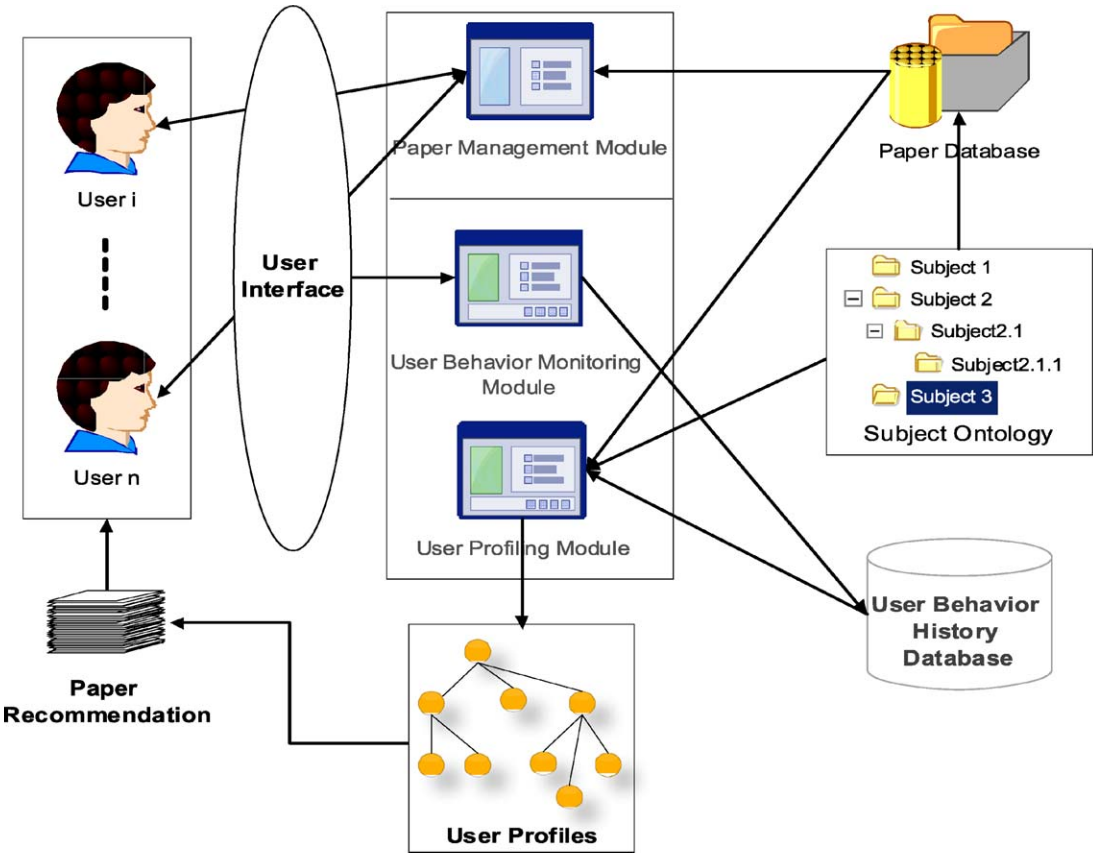


Fig. 1. Framework for generating user interest profiles.

stored in the paper database. Each paper in the paper database is classified according to the subject ontology and can readily be retrieved by users. The paper management module plays the role of fundamental component in the whole framework.

(3) *User behavior monitoring module*. This module is responsible for the background collection of the behavior data of each user. The user behavior data include searching keywords, browsing, downloading and commenting on papers, etc. The monitoring and collecting processes are totally unobtrusive.

(4) *User profiling module*. The user profiling module makes use of the user behavior data recorded by the user behavior monitoring module, the paper database and the subject ontology, to create user profiles. The user profiles obtained can be used to recommend papers to these users. We will elaborate our profiling approach in Section 6.

4. Automatic ontology extension through clustering weighted keyword graphs

In order to solve the problems in the user profiles based on the traditional ontologies, we propose an ontology extension algorithm to refine the user profiles. Before presenting it, we introduce an original subject ontology, which is defined by the Science Paper Online website (Sciencepaper Online, 2010). It is a taxonomy of research subjects and has been in use on the Internet for many years. This simple ontology consists of two levels of classification, primary subjects and secondary subjects, and it holds is–a relationships between the subjects in different levels. In the first level, there are 43 primary subjects. Each primary subject has secondary subjects as its subordinate classifications. Fig. 2 shows the section of the primary subject “computer science” in this subject ontology.

In the paper database storing the research paper data, each paper's information contains its corresponding category mark based on the subject ontology. In this paper, vector space model is used to represent the research papers. Keywords are provided by the paper's authors, representing the key content of each corresponding paper. In the vector space model, a paper is represented by a keyword vector, i.e., $paper = (keyword_1, \ldots, keyword_i, \ldots, keyword_n) \quad (1 \le i \le n)$. The weight of the *i*th keyword $keyword_i$ in $paper_p$ is denoted as $WKP(keyword_i, paper_p)$, which is computed by the TF-IDF

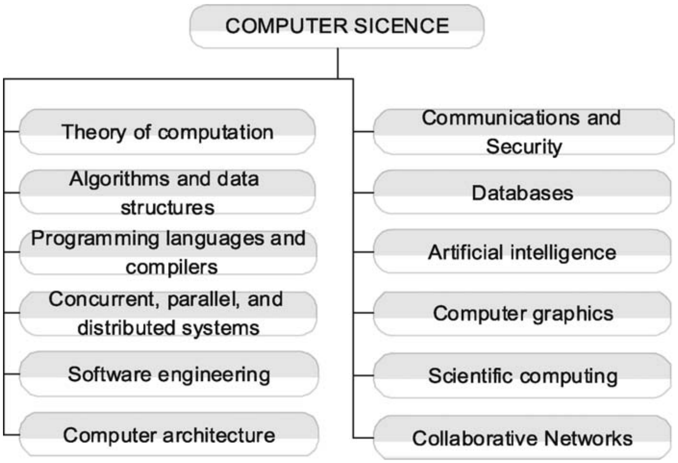


Fig. 2. The classification of “computer science” in the research paper subject ontology.

algorithm (Term Frequency-Inverse Document Frequency) (Salton and Buckley, 1988):

$$WKP(keyword_i, paper_p) = \frac{tf(keyword_i, paper_p) \cdot \log((N/n_i) + 0.01)}{\sqrt{\sum_{k=1}^{n} (tf(keyword_k, paper_p) \cdot \log((N/n_k) + 0.01))^2}},$$

where (1) $WKP(keyword_i, paper_p)$ is the weight of the keyword $keyword_i$ in $paper_p$; (2) $tf(keyword_i, paper_p)$ is the frequency of the keyword $keyword_i$ in $paper_p$; (3) N indicates the total number of the papers; (4) n_k is the frequency of the keyword $keyword_k$ in all papers.

In the following discussions, we use $PKV_i = (WKP(keyword_1, paper_i), \ldots, WKP(keyword_j, paper_i), \ldots, WKP(keyword_n, paper_i))$ to represent $paper_i$, in which the quantitative value of $keyword_j$ is computed by the TF-IDF algorithm (Salton and Buckley, 1988). By computing the TF-IDF values of keywords we obtain the keyword vectors of all papers.

The subject ontology we have discussed so far is still coarsely granular and thus not precise enough to use. In order to produce a more finely granular ontology, we now introduce an automatic clustering algorithm for weighted keyword graph, which includes three main steps, as follows.

First, we define the paper–keyword relation model represented by a bipartite graph. A bipartite graph is a graph whose vertices can be divided into two disjoint sets U and V such that every edge connects a vertex in U to one in V; that is, U and V are independent sets (West, 2000). Usually the authors specify a few keywords for their papers in order to demonstrate the contents or topic of the papers. We collect the keywords and record their relations with papers to create the bipartite graphs between papers and keywords. Here, we introduce the definition of the paper–keyword relation model:

Definition 1 *(Paper–keyword relation model).* A paper–keyword relation model is a bipartite graph $PKRM = (PNS, KNS, RS)$, where

(1) The set of nodes $PNS = \{paper_1, paper_2, ..., paper_n\}$ represents a set of papers.
(2) The set of nodes $KNS = \{keyword_1, keyword_2, \ldots, keyword_m\}$ represents a set of keywords.
(3) The set of edges $RS \subseteq (PNS \times KNS) \cup (KNS \times PNS)$ indicates the binary relations between PNS and KNS. Suppose $paper_i \in PNS$, $keyword_j \in KNS$ and $keyword_j$ exists in PKV_i, then $r_{i,j} \in RS$, where $r_{i,j}$ is an edge connecting the node $keyword_j$ and the node $paper_i$.

An example of a paper–keyword relation model is shown in Fig. 3, in which the circle nodes represent research papers and the rectangle nodes represent keywords. The edges stand for the relations between papers and keywords.

Secondly, using the one-mode projection method (Zanin et al., 2008; Zhou et al., 2007), we project paper–keyword relation models onto the keywords set KNS to generate weighted keyword graphs. The one-mode projection onto keywords means a network containing only keyword nodes, in which two keyword nodes are connected when they have at least one common neighboring paper node (Zhou et al., 2007). In the following, we give the definition of weighted keyword graphs.

Definition 2 *(Weighted keyword graph).* A weighted keyword graph is a graph $WKG = (KNS, CR)$, where

(1) KNS is the node set representing keywords. For each $keyword_i \in KNS$, the weight of $keyword_i$ is calculated by

$$Weight(keyword_i) = \sum_{paper_p \in PS(keyword_i)} WKP(keyword_i, paper_p),$$

where $PS(keyword_i)$ is the set of papers containing $keyword_i$; $WKP(keyword_i, paper_p)$ is the weight of $keyword_i$ in PKV_p.
(2) The set of edges $CR \subseteq (KNS \times KNS)$ represents the relations between different keyword nodes in KNS.
(3) $WGT(R_{i,j}) = count(PS(keyword_i) \cap PS(keyword_j))$ denotes the weight of $R_{i,j}$, where $R_{i,j}$ is the edge linking $keyword_i$ and $keyword_j$; $PS(keyword_i)$ is the paper set containing $keyword_i$; the function $count(\,)$ counts the number of papers of which the keyword vectors contain both $keyword_i$ and $keyword_j$, i.e., $PS(keyword_i) \cap PS(keyword_j)$.

In the paper–keyword relation model ($PKRM$), if any two keyword nodes in KNS are connected to one or more paper nodes in PNS, then there is an edge connecting the two keyword nodes in CR of WKG. $WGT(R_{i,j})$ indicates the number of co-occurrences of the two keywords in papers. During the subsequent keyword clustering process $WGT(R_{i,j})$ are changed.

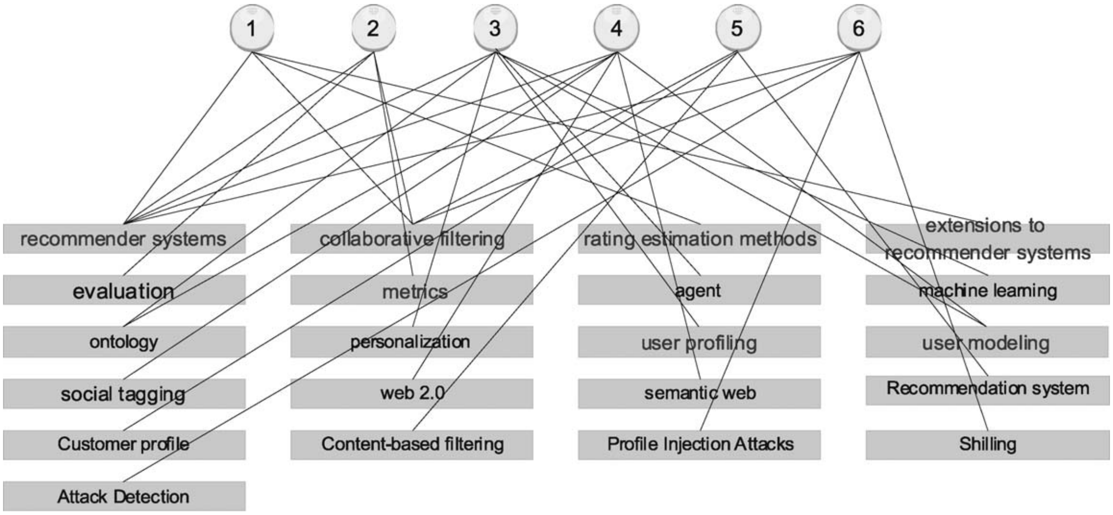
For example, a weighted keyword graph of the subject “artificial intelligence”, which is created with Pajek (Batagelj and Mrvar, 1998), is illustrated in Fig. 4. The keywords engaged are from 200 papers in the subject of “artificial intelligence”. The construction of the weighted keyword graph depends on the paper set. As the paper set is changed, so the result may be different.

Finally, we cluster the weighted keyword graph to extend the subject ontology. The keyword clustering is virtually the process of increasing the edge weight threshold and finding a practical one. In Fig. 4, we can clearly see that keywords related to similar themes tend to cluster together. These keyword bundles naturally emerge and signify the tendency towards a concentration of the contents among similar papers. Through the keyword clustering process, these concentrated keyword bundles will be decomposed into separate ones. Now we present the keyword clustering procedure:

(1) According to Definition 2, the weighted keyword graph WKG of a subject is generated.
(2) Define a threshold Ω for the weights of edges in WKG. If $WGT(R_{i,j}) < \Omega$, the edge linking $keyword_i$ and $keyword_j$ would be removed from WKG, where $keyword_i$, $keyword_j \in KNS$ and $i \neq j$.
(3) A set of connected components are produced as a result of Step (2), with each connected component indicating one keyword cluster. These keyword clusters are named “topics”. We define a topic graph set $TGS = \{TG_1, \ldots, TG_i, \ldots, TG_n\}$ in which TG_i represents a topic, to represent all the topics generated.

Here, we make three rules to facilitate the application of keyword clustering.

(1) Only when there are at least four keyword nodes in a cluster is this keyword cluster considered to be a topic; the nodes in the clusters with keywords less than four, together with the keyword nodes with no connections, are collected into a special topic, called “heterogeneous” topic. The “heterogeneous” topic is not included in user profiles, and therefore in the rest of this paper the term “topic” only refers to usual keyword clusters that are generated by the keyword clustering process.
(2) We name the topics with the keywords which possess the largest value of the sum of the weights of edges connected to them. If there is more than one keyword node with the highest connectivity, we choose one from them randomly.
(3) In the case that names of the clusters (topics) are the same as their immediate upper subjects’ names, we would not consider these kinds of clusters to be topics; that is, the keyword node with the highest connectivity in each cluster and the edges linking to it are virtually neglected.



The papers represented by circles are:

1. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions
2. Evaluating Collaborative Filtering Recommender Systems
3. Ontological User Profiling in Recommender Systems
4. Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations
5. Individual and Group Behavior-based Customer Profile Model for Personalized Product Recommendation
6. Towards Trustworthy Recommender Systems: An Analysis of Attack Models and Algorithm Robustness

Fig. 3. An example of a paper–keyword relation model.

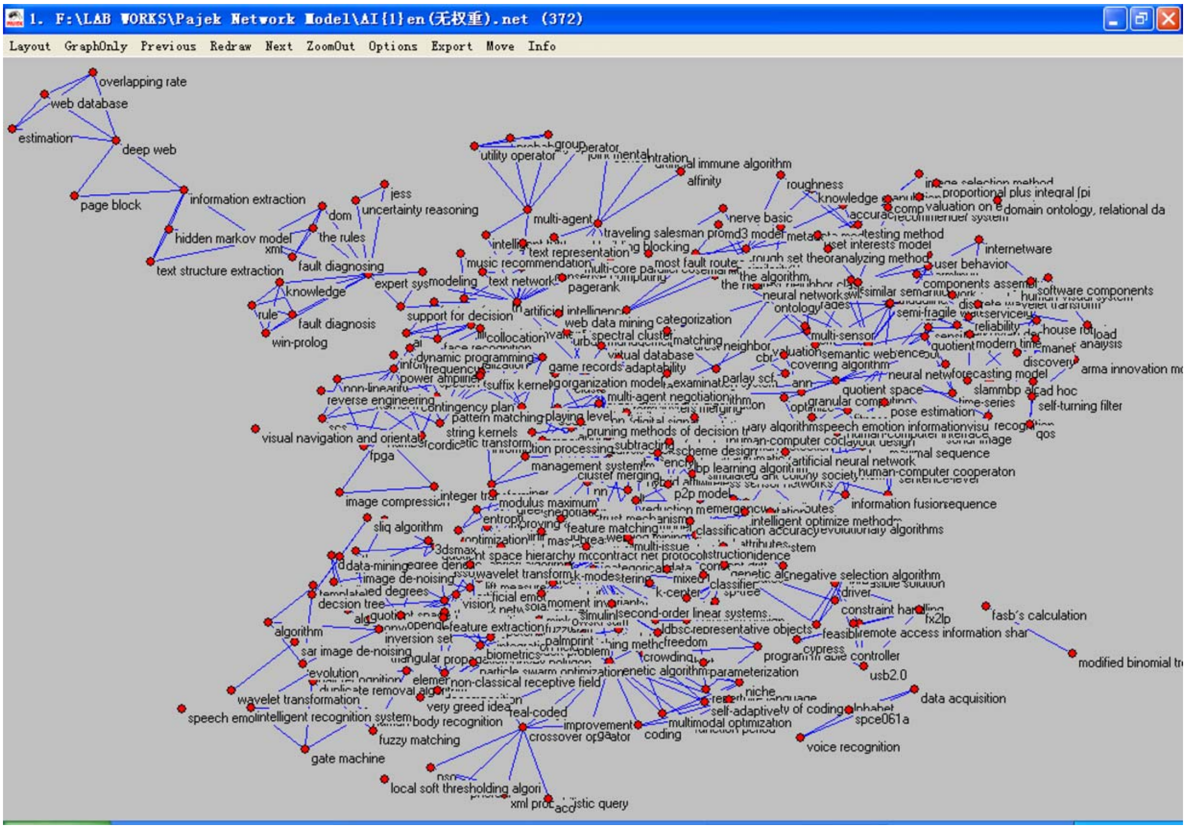


Fig. 4. A weighted keyword graph of the subject “artificial intelligence” demonstrated by Pajek.

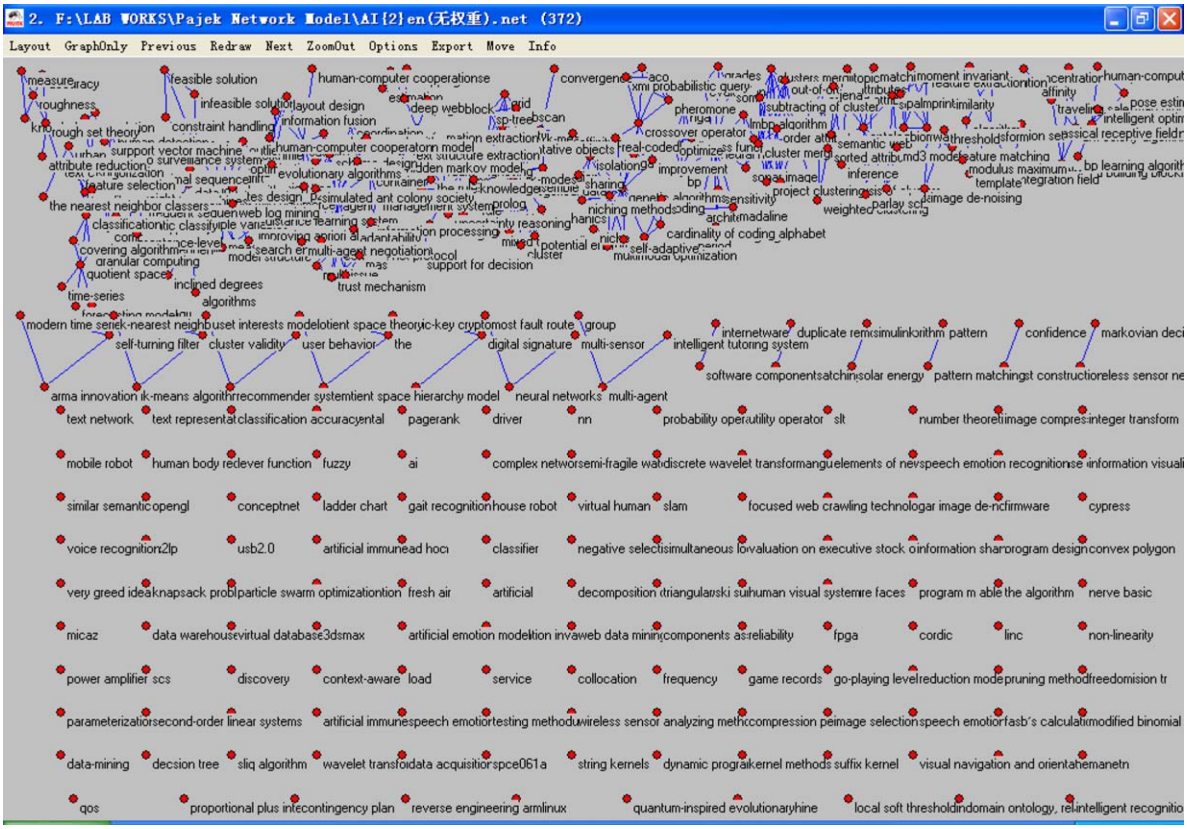


Fig. 5. A weighted keyword graph of the subject “artificial intelligence” (edge weight threshold equals 2).

If we assign a different value to the threshold Ω, different emergences of *TGS* then will be produced. During our experiment, in which 200 papers were used, we discovered that, when 1 was used as the threshold of edge weight, *TGS* is as shown in Fig. 4; when 3 was used as the threshold of edge weight, most of the keyword nodes in *TGS* became extremely discrete and only three keyword clusters, within which very few keyword nodes existed maintained. The selection of the edge weight threshold is strongly pertinent to the content and the quantity of the papers used in this process. Now we use 2 as the threshold when clustering the weighted keyword graph in Fig. 4 and the result is showed in Fig. 5. In this weighted keyword graph, aside from the “heterogeneous” topic, there are ten other newly emerged topics which are listed in Table 1. Therefore, the subordinate classifications of the secondary subject “artificial intelligence” are engendered.

So far, we have finished the extension of the secondary subject “artificial intelligence”. All the secondary subjects in the original subject ontology are extended by the keyword clustering process. An illustrative sketch of the result of the ontology extension through keyword clustering is provided in Fig. 6.

Table 1
Topics of the secondary subject “artificial intelligence”.

Topic name	Connectivity of central keyword node
neural network	23
agent	20
SVM	16
genetic algorithms	14
expert system	11
hybrid attributes	10
ontology	9
wavelet transform	8
inversion set	6
traveling salesman problem	5

The method of ontology extension through keyword clustering holds two advantages. It allows subject ontology to be automatically and sensitively adaptive to the changes of research topics in any subject. However scientific research topics change, whatever new research hotspots appear or whichever old research contents fade out, for example, the changes and the newest status will be immediately reflected in the new formation as the keyword clustering is executed. The ontology extension also makes the user interest profiles more precise and distinct. With the clustering method, we are able to place a new angle of view with more accurate classifications on all subjects; in this way, users’ interests will be captured and recorded even more clearly. In the next section, we will discuss the construction of user interest profiles.

5. Approaches to generating user profiles based on the extended subject ontology

We now use the extended subject ontology to create user profiles. In this section, we present our refined ontological profiling approach to solve the problems of the low accuracy of recommendation and of the coarse granularity of user interest profiles in traditional ontology-based profiling algorithms. This user interest profiling approach is able to distinguish between the different contributions of the papers on the same topic to the construction of user interest profiles. Also, apart from the user profile obtained directly from the user behavior data, which we call the “explicit interest” profile, we apply implicit profiles to infer possible interests that users may develop in the future, in order to describe user interests more roundly and thereby improve recommendation. A user interest profile therefore consists of two parts: an explicit interest profile and an implicit interest profile. An arbitrary user has an explicit interest in a certain topic if the user directly accesses one or more papers in the topic. By contrast, an arbitrary user’s implicit interest

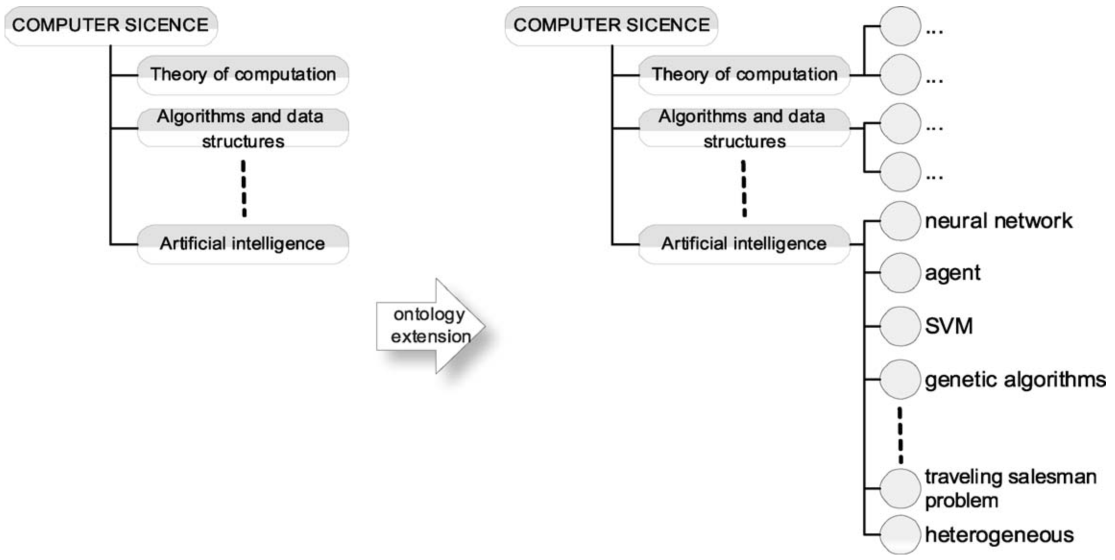


Fig. 6. The result of the ontology extension.

refers to the user's interest in a certain topic about which no papers have yet been directly accessed by the user. Through the ontological relationships between this topic and the other topics in which the user has explicit interests, the user's implicit interest about this "uninteresting" topic can be predicted. Implicit interest profiles express users' potential interests in topics that are obscurely discovered.

Definition 3 *(User interest profile).* A user interest profile consists of two parts: an explicit interest profile part and an implicit interest profile part, denoted by $UEIP(User_u)$ and $UIIP(User_u)$, respectively. Each of the two parts is a set of 2-tuples: $UEIP(User_u) = \{(topic_t, EI(topic_t, user_u))|topic_t$ is a topic, $EI(topic_t, user_u)$ measures $User_u$'s explicit interest in $topic_t\}$; and $UIIP(User_u) = \{(topic_K, II(topic_K, user_u))|topic_K$ is a topic, $II(topic_K, user_u)$ measures $User_u$'s implicit interest in $topic_K\}$.

5.1. The profiling method for users' explicit interests

We now introduce the construction of the explicit interest profile. The generation of explicit interest profiles is achieved through the user profiling module in Fig. 1. Fig. 7 shows a detailed process chart of this module.

The key task of creating the explicit interest profiles of users lies in computing the interest values of the topics in which the users have explicit interest. Considering a paper's possible relevance to different topics, the possible relevance of a keyword to different topics is taken into account. In order to differentiate between the topics for which a user has the same access behaviors, we assign a measurement to each topic called the relevance factor. In the following, we first give a group of algorithms and definitions, and then we present the definition and explanation of the relevance factor and the explicit interest profiler.

Algorithm 1 *(Generating user-featured topic graph).* Based on the definition of weighted keyword graph *WKG*, a user-featured topic graph, which is referred as *UTG*, can be created in the following steps:

(1) Create the weighted keyword graph $WKG(topic_t)$ for $topic_t$.
(2) Eliminate all edges and reset weights of all nodes to 0 to generate a new graph $UTG(topic_t)$.
(3) Based on the user's behavior history data, assign an edge between $keyword_i$ and $keyword_j$ only if $paper_p$ contains both $keyword_i$ and $keyword_j$, where $paper_p \in UPS(user_u)$ and $UPS(user_u)$ represents the set of papers recorded in $user_u$'s behavior history data; $keyword_i, keyword_j \in UTG(topic_t)$. And the weights of the edges are defined as the number of related papers.
(4) The weight of node $keyword_i$ in *UTG* which is referred as $IK(keyword_i)$ measuring the user's interest in the content related to $keyword_i$ is calculated by

$$IK(keyword_i) = \sum_{paper_p \in PS(keyword_i)} (WKP(keyword_i, paper_p) \cdot UIP(paper_p, user_u)),$$

where $keyword_i \in KNS$; $PS(keyword_i)$ is the set of papers containing $keyword_i$ and $PS(keyword_i) \in UPS(user_u)$; $UPS(user_u)$ is the set of papers recorded in $user_u$'s behavior history database; $WKP(keyword_i, paper_p)$ is the weight of $keyword_i$ in the vector space model of $paper_p$. And

$$UIP(paper_p, user_u) = \frac{BF(paper_p, user_u)}{(DP(paper_p, user_u))^{1/\Phi}}, \quad \Phi \in (0, +\infty),$$

where $BF(paper_p, user_u)$ denotes behavior factor which equals the number of times $user_u$ browses $paper_p$, or the number of times $user_u$ downloads $paper_p$, or the score $user_u$ rates $paper_p$ divided by 5, depending on the type of behavior based on which $TG(topic_t)$ is calculated; $DP(paper_p, user_u)$ indicates the number of days that have passed since the behavior occurred; Φ is a parameter for adjustment.

Φ should be set as a value which optimizes the recommendation performance. In the conventional ontology-based user interest profiling algorithm, Φ is set as 1 (Middleton et al., 2004).

Definition 4 *(Inner edge).* An edge is termed an inner edge if it links two keyword nodes in the same weighted keyword graph. The sum of the weights of the inner edges of a keyword node is called inner edge strength. In this paper, we use $IES(keyword_i)$ to refer to the inner edge strength of $keyword_i$.

Definition 5 *(Cross edge).* An edge is termed as a cross edge if it links two nodes from different weighted keyword graphs. The sum of the weights of the cross edges of a node is called cross edge strength. In this paper, we use $CES(keyword_i)$ to refer to the cross edge strength of $keyword_i$.

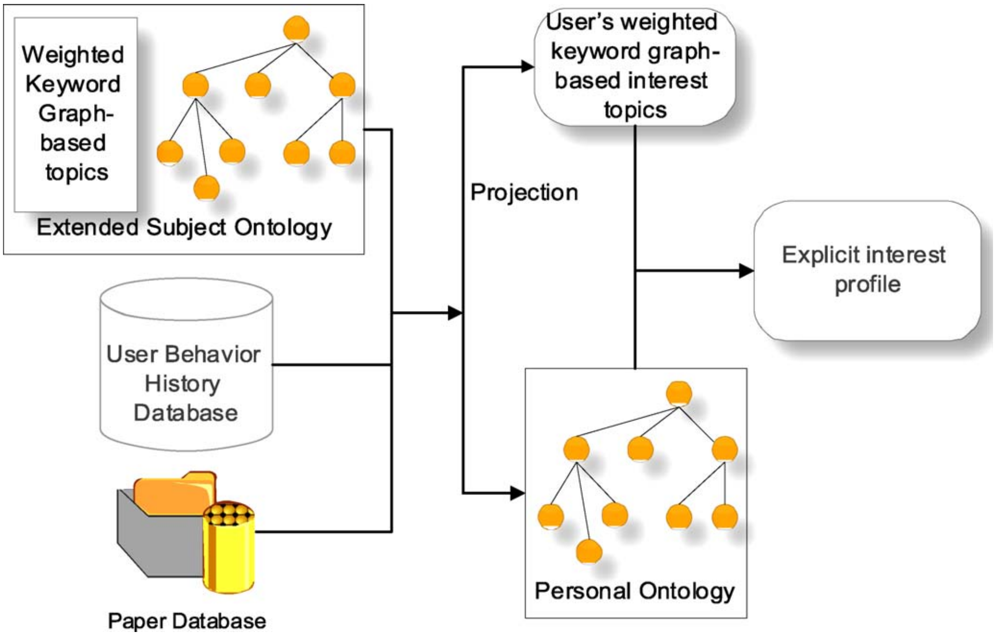

Weighted Keyword Graph-based topics
Extended Subject Ontology
User Behavior History Database
Paper Database
Projection
User's weighted keyword graph-based interest topics
Explicit interest profile
Personal Ontology

Fig. 7. Process of generating explicit user interest profiles.

Definition 6 *(Relevance strength).* Based on the definitions of inner edge and cross edge, the relevance strength $RS(keyword_i, topic_t)$ of $keyword_i$ towards $topic_t$ can be calculated by

$$RS(keyword_i, topic_t) = \frac{IES(keyword_i)}{IES(keyword_i) + CES(keyword_i)},$$

which quantifies the relevance of $keyword_i$ to $topic_t$. From the definition of relevance strength we can determine that $RS(keyword_i, topic_t) \in (0, 1]$ and $IES(keyword_i) > 0$.

Definition 7 *(Relevance factor).* The relevance factor $RF(user_u, topic_t)$ indicating the measurement of the relevance of $user_u$'s interest towards $topic_t$ is given as

$$RF(user_u, topic_t) = ((\overrightarrow{TGV} \cdot \overrightarrow{UTGV})/(\overrightarrow{TGV} \cdot \vec{1}))^{1/\lambda}, \quad \lambda \in [1, +\infty),$$

in which $\overrightarrow{TGV}$ denotes a vector comprising the weights of all nodes in $TG(topic_t)$; $\overrightarrow{UTGV}$ denotes a vector listing the weights of all nodes in $UTG(topic_t)$ in the sequence as same as $\overrightarrow{TGV}$; λ is an adjustable parameter. The setting of λ depends on experimental data, and λ should be assigned a value to maximize the recommendation performance. We will discuss this in Section 6.

The relevance factor quantifies the pertinence of the contents of a certain topic that a user accesses to the most essential content of the topic, with the purpose of measuring the relevance between the user's interest and the topic.

Now, we present the algorithm of the explicit interest profiler.

Algorithm 2 *(Explicit interest profiler).* The explicit interest of $user_u$ in $topic_t$ $EI(\ topic_t, user_u)$ is computed by

$$EI(topic_t, user_u) = \frac{RF(user_u, topic_t) \cdot \sum_{paper_p \in topic_t} UIP(paper_p, user_u)}{\sum_{topic_{ti} \in UTS(user_u)} \left(RF(user_u, topic_{ti}) \sum_{paper_{pi} \in topic_{ti}} UIP(paper_{pi}, user_u) \right)},$$

where $UTS(user_u)$ signifies the set of topics that $user_u$ has explicit interests in.

To some extent, the measurement of the relevance factor is similar to the feature selection techniques in text classification, which forms the basis of the task. Well-developed feature selection techniques include document frequency (DF), information gain (IG), mutual information (MI), chi-square and expected cross entropy (ECE) (Sebastiani and Ricerche, 2002; Yang, 1995). Most of these functions try to capture the intuition according to which the most valuable terms for a certain categorization are those that are distributed most differently in the sets of positive and negative examples of the category (Sebastiani and Ricerche, 2002). Instead of employing these existing techniques, we propose the new measurement since it is expressly designed for our automated classification method and it calculates the tightness between different keywords well. The posterior experiment testifies its good performance in efficiency and accuracy.

5.2. Profiling method for users' implicit interests

Many researchers have emphasized the importance of the novelty of the recommended items in recommender systems. Herlocker et al. (2004) asserted that new dimensions were needed for analyzing recommender systems that considered the "nonobviousness" of the recommendation: novelty and serendipity. Zhou et al. (2010) sought to gain in both diversity and accuracy of recommendations, and argued that "real value is found in the ability to suggest objects users would not readily discover for themselves, that is, in the novelty and diversity of recommendation". They employed two metrics for measuring recommendation diversity.

Compared with other profiling strategies, ontology-based profiling methods have the advantage that they can infer user interests on the basis of hierarchical structures and are able to utilize the relations between their ontologies and external concepts for recommendations. Due to ontological inference, user profiles can be rounded off and can be matched better to the wide range of user interests (Felden and Linden, 2007). In Middleton et al. (2001, 2003, 2004), researchers used is–a relationships between subclasses and their immediate super classes by adding 50% of the interest values of the classes to those of the super-classes. Trials substantiated that ontological inference boosted the recommendation accuracy of individual recommendations.

Instead of taking advantage of the is–a relationship between subordinate topics and their super-topics, we believe that the semantic relations between topics under the same super-topics are more preferable for inferring user interests. In our extended

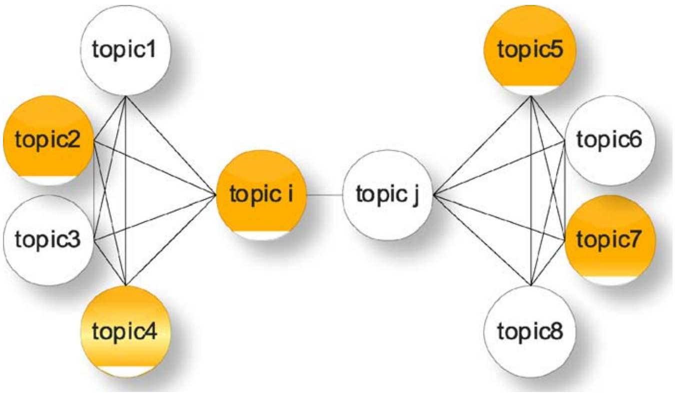


Fig. 8. An example of topic network graph.

subject ontology, the semantic relations between topics can be obtained by an inverse process of keyword clustering. Here, we give the definition of topic network graphs to reveal these semantic relations:

Definition 8 *(Topic network graph).* A topic network graph is a graph $TNG = (ETNS, ITNS, ES)$, where

(1) $ETNS$ is the node set in which each node represents a topic in which one or more papers are ever accessed by $user_u$, the weight of the node $topic_i$ equals $EI(\ topic_t, user_u)$.
(2) $ITNS$ is the node set in which each node represents a topic in which no papers are ever accessed by $user_u$, the weight of the node $topic_x$ equals the measurement of $user_u$'s implicit interest in $topic_x$, which is denoted as $II(\ topic_t, user_u)$. The calculation of $II(\ topic_t, user_u)$ is given in Algorithm 3.
(3) The set of edges $ES \subseteq ((ETNS \cup ITNS) \times (ETNS \cup ITNS))$, representing the relevance between different topics in TNS.
(4) $WET(E_{i,j}) = sum(MES(topic_i,\ topic_j))$, where $E_{i,j}$ represents the edge connecting $topic_i$ and $topic_j$; $WET(E_{i,j})$ denotes the weight of $E_{i,j}$; $MES(topic_i, topic_j)$ is the edge set in which each edge corresponds to an edge in WKG connecting one keyword belongs to $topic_i$ and one keyword belongs to $topic_j$; the function $sum(\ )$ calculates the summation of the weights of the edges in $MES(topic_i, topic_j)$.

Thereby, we can now obtain a topic network graph for each secondary subject. Fig. 8 shows a sample of a topic network graph. The demonstration is provided for illustrative purposes only. In Fig. 8, the yellow circles represent topics belong to *ETNS* and the white circles represent topics belong to *ITNS*.

The relevance between topics is a meaningful way to infer implicit interests. As long as a user has explicit interests in some topics, the implicit interests of the user in the topics that the user has not yet explored can be estimated reasonably by the following algorithm.

Table 2
Topics of the secondary subject "databases".

Topic name	Connectivity of central keyword node
data mining	32
SQL	14
clustering	11
data warehouse	10
distributed database	9
relational database	9
spatial data mining	7

Algorithm 3 *(Implicit interest profiler).* Let $VC(topic_i)$ denotes the connectivity of the node $topic_i$, which equals the sum of the weights of all edges linked to the node $topic_i$; let $II(topic_x, user_u)$ be $user_u$'s implicit interest value for $topic_x$. Then, $II(topic_x, user_u)$ can be computed as follows:

$$II(topic_x, user_u) = \sum_{topic_i \in Neightbors(topic_x)} \left(\frac{WET(E_{i,x})^2}{VC(topic_x) \cdot VC(topic_i)} \cdot EI(topic_i, user_u) \right),$$

in which $topic_x \in ITNS$ and $topic_i \in ETNS$; $Neighbors(topic_x)$ stands for the set of topic nodes neighboring $topic_x$ in TNG.

6. Prototype system and empirical evaluation

6.1. SPRS—a prototype for scientific paper recommender system

In order to assess the precision and effectiveness of our approach, we developed a scientific paper recommender prototype system (SPRS), which is illustrated in Fig. 9. There are three columns in the homepage of SPRS. The left column consists of three parts: user information, the papers which are currently read by the current user and the papers that the current user wants to read. The middle column presents the paper recommendations separately based on the records of three behaviors: download, browse and comment. In the right column, three lists of papers in the behavioral records are presented.

Based on the user profiles obtained, SPRS is able to provide paper recommendations to users. Recommendations are produced by correlating the users' topics of interest and the papers classified to these topics. The explicit interests and implicit interests of a certain user are both, separately, used for recommendation. The top-*N* items rule is implemented to control the number of recommendations. Papers are ranked in the order of the recommendation confidence of them when being presented to a specific user; a paper's recommendation confidence for a user equals the product of the classification confidence of this paper in a topic that maximizes this paper's classification confidence and this user's degree of interest in this topic. Depending on which interest profile is used for recommendation, the interest degree can be of either explicit interest or implicit interest.

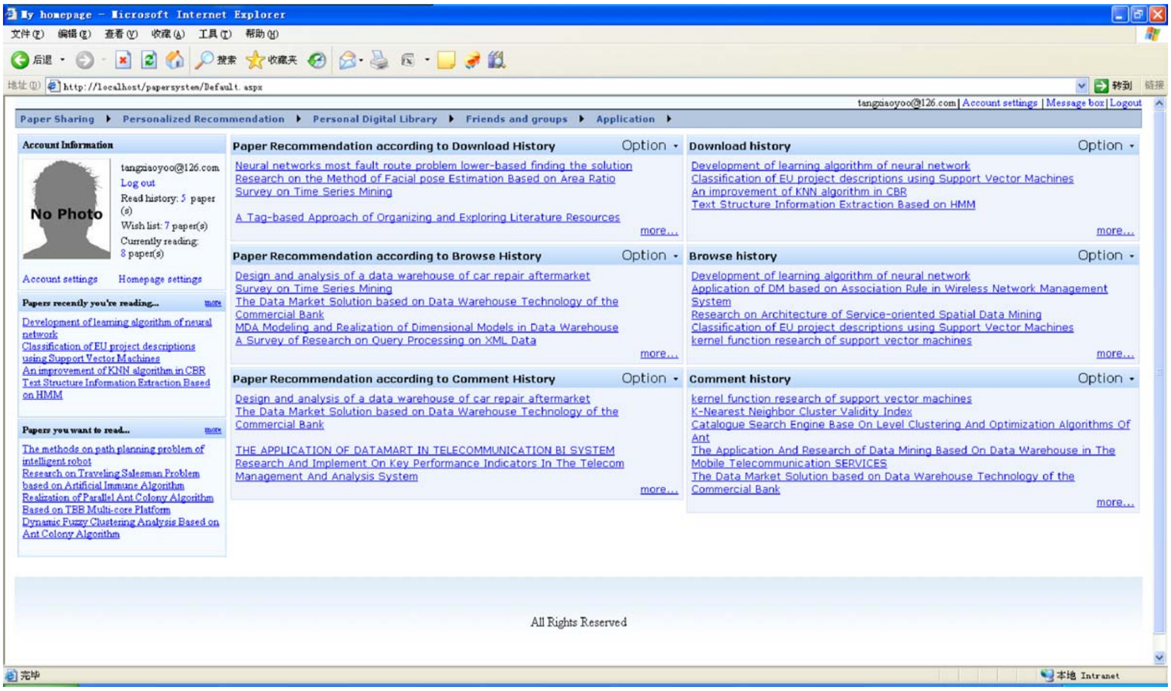
The behavior history of users, including browse, download and comment histories, are recorded and used to calculate their interest profiles. Also, papers that have been accessed by users before will be ignored when recommendations are provided. The recommendations are presented separately. If users are interested in these recommendations, they certainly tend to click on them, unless these items have already been visited.

6.2. Empirical evaluation

In this section, we conducted three experiments by SPRS to examine the ontology extension method and the extended ontology-based profiling approach proposed in this paper.

6.2.1. Experiment 1: the acquisition of the extended subject ontology

First, we evaluated the extension of the original subject ontology by generating keyword clusters for the weighted keyword graphs of "artificial intelligence" and "databases" subject. We used 200 papers in each subject in the following experiment. The papers were from the Science Paper Online website (Sciencepaper Online, 2010). The clustering result of the "artificial intelligence" subject has already been used as the examples shown in Section 4. In this section, we only discuss the clustering result of "databases". The weighted keyword graph of "databases" in which the edge weight threshold value is 1 is showed in Fig. 10. Then we assign 2 as the


My homepage – Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退
搜索
收藏夹
地址(D)
转到
链接
No Photo
完毕
本地 Intranet

http://localhost/papersystem/Default.aspx

tangxiaoyoo@126.com | Account settings | Message box | Logout

Paper Sharing ▸ Personalized Recommendation ▸ Personal Digital Library ▸ Friends and groups ▸ Application ▸

Account Information

tangxiaoyoo@126.com
Log out
Read history: 5 paper(s)
Wish list: 7 paper(s)
Currently reading: 8 paper(s)

Account settings Homepage settings

Papers recently you're reading... more

Development of learning algorithm of neural network
Classification of EU project descriptions using Support Vector Machines
An improvement of KNN algorithm in CBR
Text Structure Information Extraction Based on HMM

Papers you want to read... more

The methods on path planning problem of intelligent robot
Research on Traveling Salesman Problem based on Artificial Immune Algorithm
Realization of Parallel Ant Colony Algorithm Based on TBB Multi-core Platform
Dynamic Fuzzy Clustering Analysis Based on Ant Colony Algorithm

Paper Recommendation according to Download History Option ▾

Neural networks most fault route problem lower-based finding the solution
Research on the Method of Facial pose Estimation Based on Area Ratio
Survey on Time Series Mining

A Tag-based Approach of Organizing and Exploring Literature Resources
more...

Download history Option ▾

Development of learning algorithm of neural network
Classification of EU project descriptions using Support Vector Machines
An improvement of KNN algorithm in CBR
Text Structure Information Extraction Based on HMM
more...

Paper Recommendation according to Browse History Option ▾

Design and analysis of a data warehouse of car repair aftermarket
Survey on Time Series Mining
The Data Market Solution based on Data Warehouse Technology of the Commercial Bank
MDA Modeling and Realization of Dimensional Models in Data Warehouse
A Survey of Research on Query Processing on XML Data
more...

Browse history Option ▾

Development of learning algorithm of neural network
Application of DM based on Association Rule in Wireless Network Management System
Research on Architecture of Service-oriented Spatial Data Mining
Classification of EU project descriptions using Support Vector Machines
kernel function research of support vector machines
more...

Paper Recommendation according to Comment History Option ▾

Design and analysis of a data warehouse of car repair aftermarket
The Data Market Solution based on Data Warehouse Technology of the Commercial Bank

THE APPLICATION OF DATAMART IN TELECOMMUNICATION BI SYSTEM
Research And Implement On Key Performance Indicators In The Telecom Management And Analysis System
more...

Comment history Option ▾

kernel function research of support vector machines
K-Nearest Neighbor Cluster Validity Index
Catalogue Search Engine Base On Level Clustering And Optimization Algorithms Of Ant
The Application And Research of Data Mining Based On Data Warehouse in The Mobile Telecommunication SERVICES
The Data Market Solution based on Data Warehouse Technology of the Commercial Bank
more...

All Rights Reserved

Fig. 9. Scientific paper recommender system (SPRS).

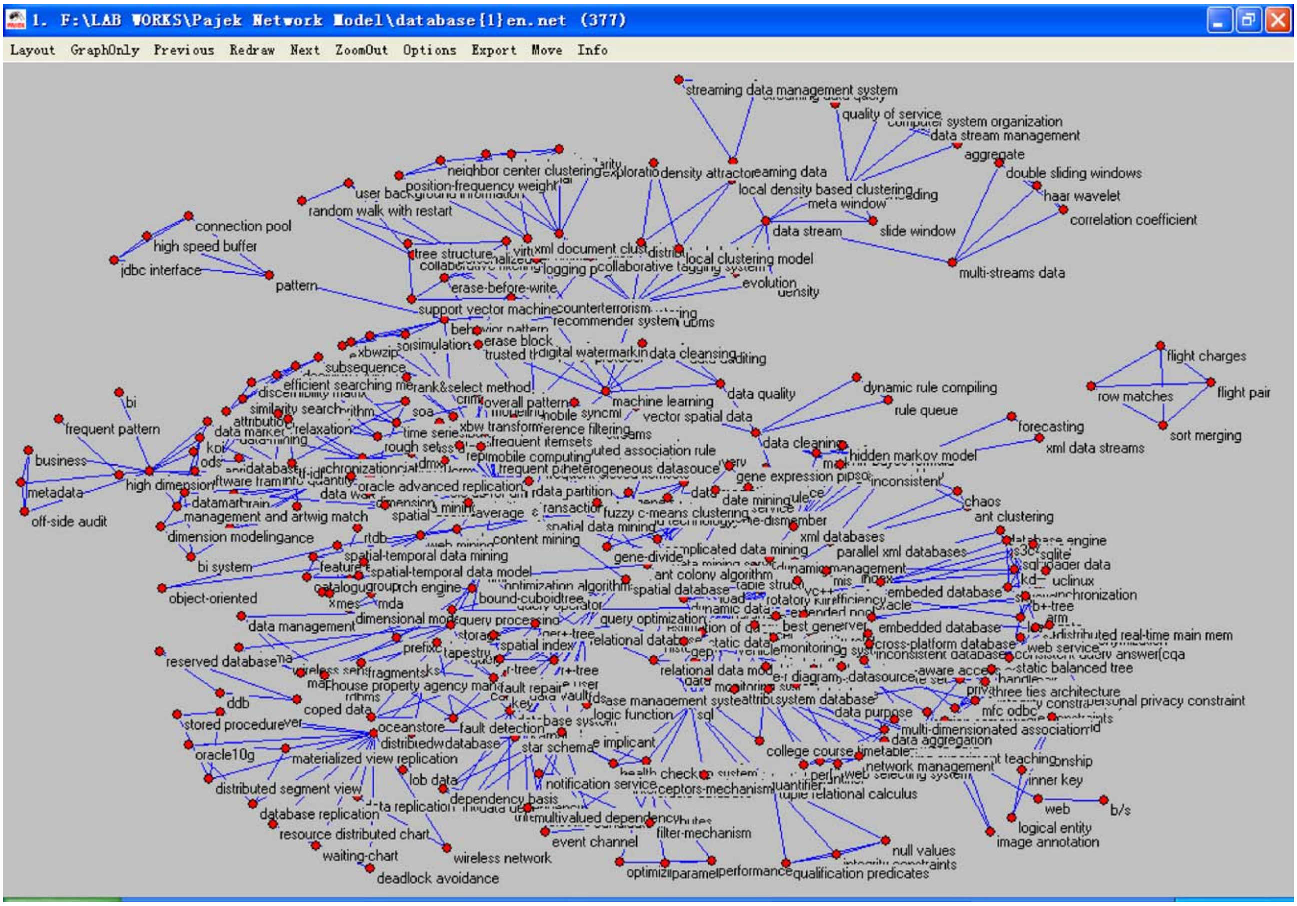


Fig. 10. The weighted keyword graph of “databases” whose edge weight threshold value is 1.

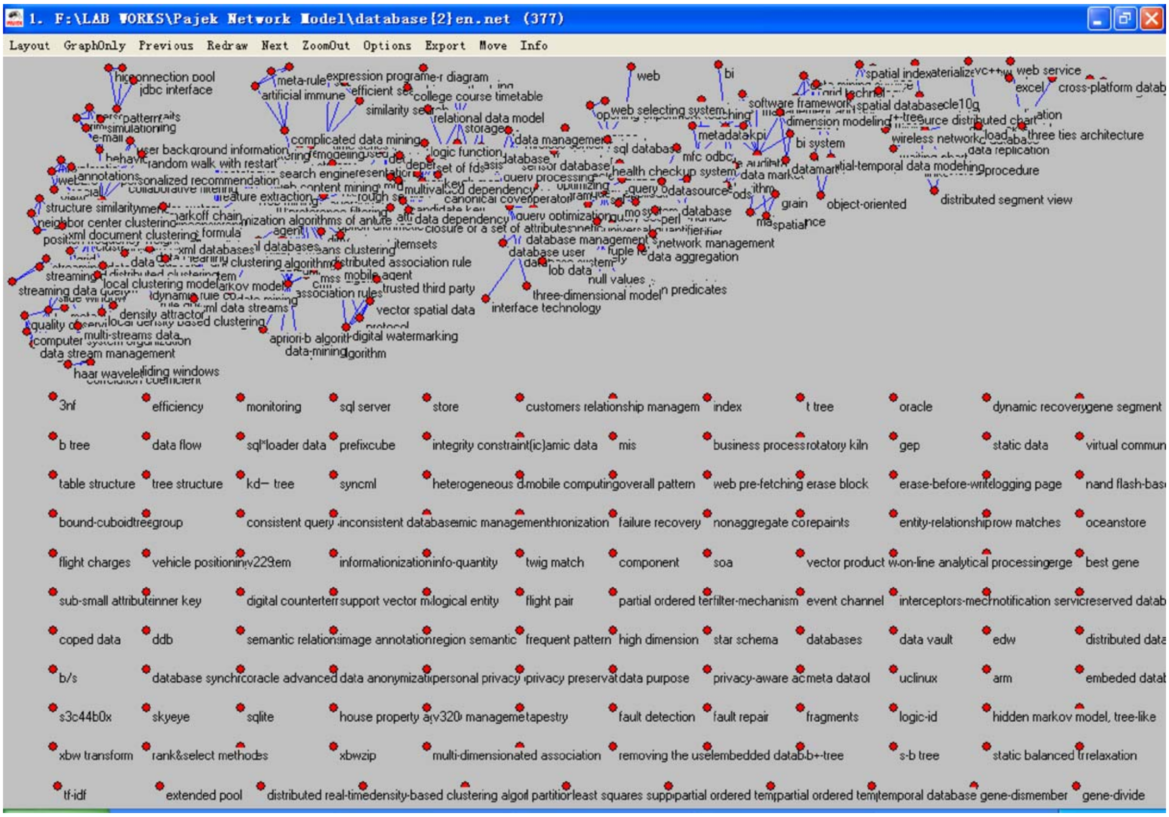


Fig. 11. The weighted keyword graph of "databases" whose edge weight threshold value is 2.

threshold value; the corresponding weighted keyword graph is presented in Fig. 11. As shown in Fig. 10, there are 2 clusters (topics), while the number of clusters is 7 in Fig. 11. Small newly emerged keyword clusters and their corresponding connectivities are presented in Table 2. Since the value 2 is proved to be a rather decent value of the edge weight threshold, we then used it as the edge weight threshold. Thus, the seven topics plus the "heterogeneous" topic, constitute the new classifications of the secondary subject "databases". In this way, the "databases" subject of the original ontology is extended.

6.2.2. Experiment 2: user interest profile construction

Based on the extended ontology formed from the clustering results, user interest profiles were calculated according to our user interest profiling algorithm. This experiment lasted for 30 days, and 5 master students were engaged in it. These students were asked to freely and naturally browse, download, comment on or collect the research papers of the subjects of "databases" and "artificial intelligence" within SPRS. Because we mainly focused on the validity of our profiling approach at the detailed classification level, we did not include other subjects. We also conducted a training process to find the best value of the parameter λ in the new user interest profiling algorithm. During the training process, we found that using 2 as λ shows the best performance of recommendation, and we therefore used 2 as the value of λ.

In order to make a comparison, we also compute the user interests using a traditional ontology-based profiling method. In both methods, the parameter Φ is set as 1, which equals the value of this

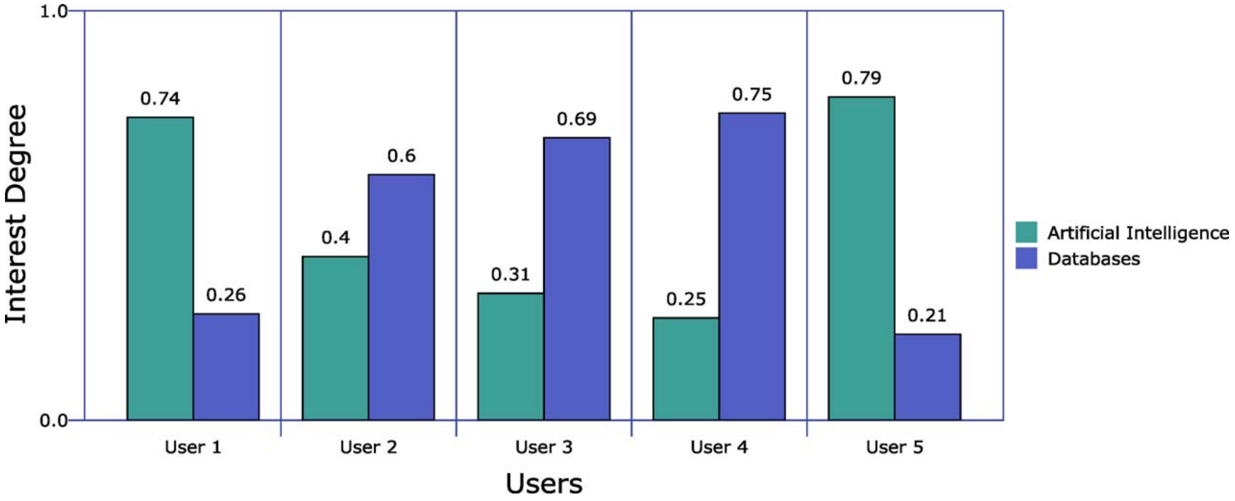


Fig. 12. Computational results of the users' interests on the secondary subject level (explicit interests).

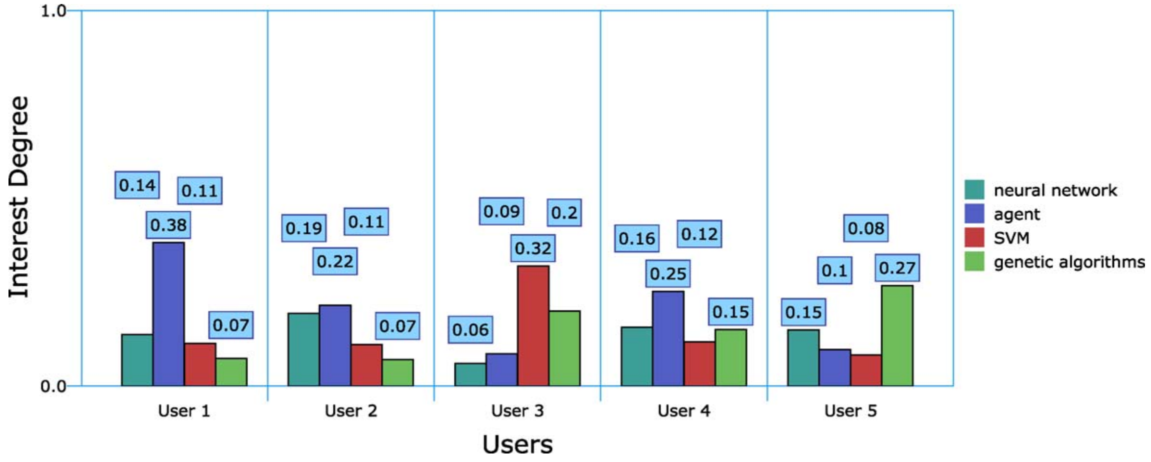


Fig. 13. Computational results of the users' interests in "artificial intelligence" (explicit interests).

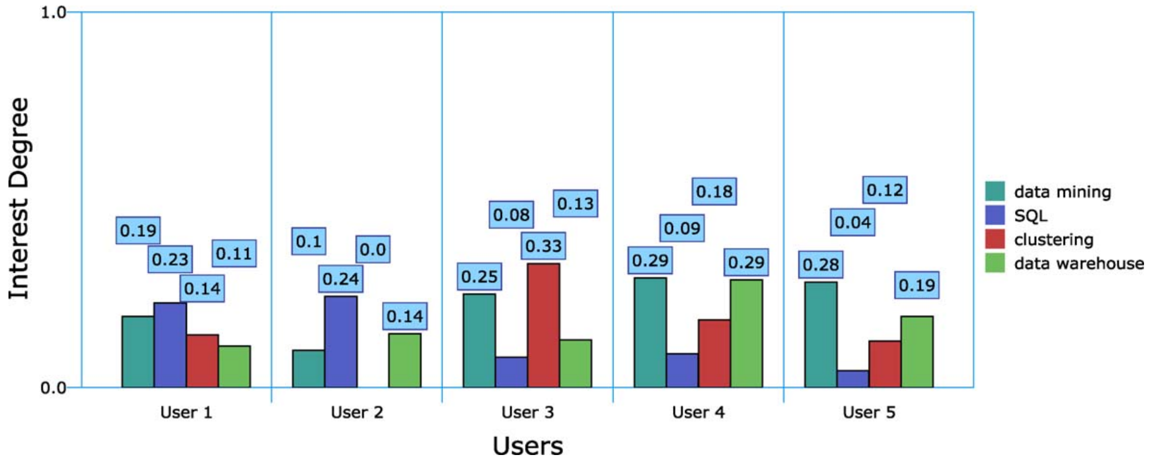


Fig. 14. Computational results of the users' interests in "databases" (explicit interests).

parameter in the user interest profiler in Middleton et al. (2004). The calculation of the interests of the five users is then conducted and the result is shown in Fig. 12.

Based on our refined interest profiling algorithm, using the same data, we compute the five users' explicit interests. The four largest topics in each of the two secondary subjects are adopted in the observation; for "artificial intelligence", the topics are "neural network", "agent", "SVM" and "genetic algorithms", and for "databases", they are "data mining", "SQL", "clustering" and "data warehouse". They represent eight important topics in "artificial intelligence" and "databases" reflected in our experimental paper pool. The results are shown in Figs. 13 and 14. We can see that, although on the secondary subject "databases" the interests of User 1 and User 5 are very similar, that is, 0.26 and 0.21, respectively, as showed in Fig. 12, on the third level (topic level) shown in Fig. 14, the interest distribution on the four topics shows the subtle differences between them: User 1 is very interested in "SQL", while the user 5's interest in it is very low. Evidently, our weighted keyword graph-based ontological user profiling method improves the accuracy of user profiling in that it produces more rounded and detailed user profiles compared to the traditional approach.

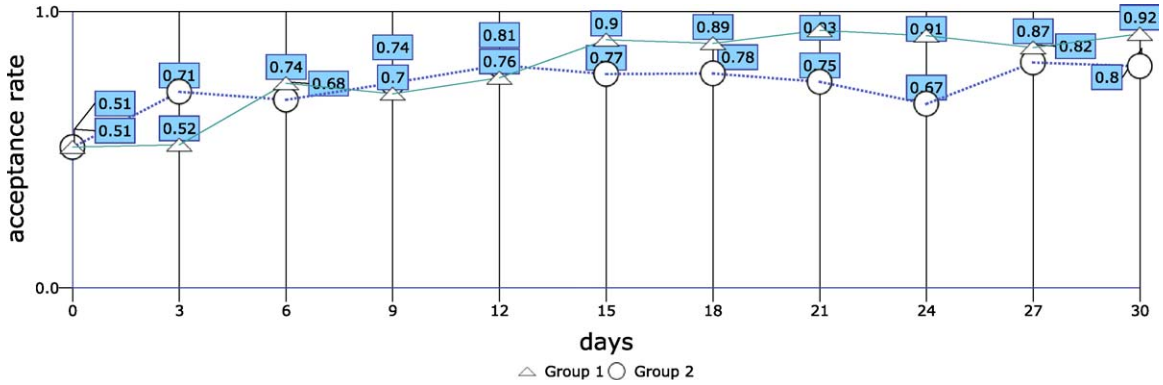


Fig. 15. Users' average acceptance rates for the recommendation in two groups within 30 days.

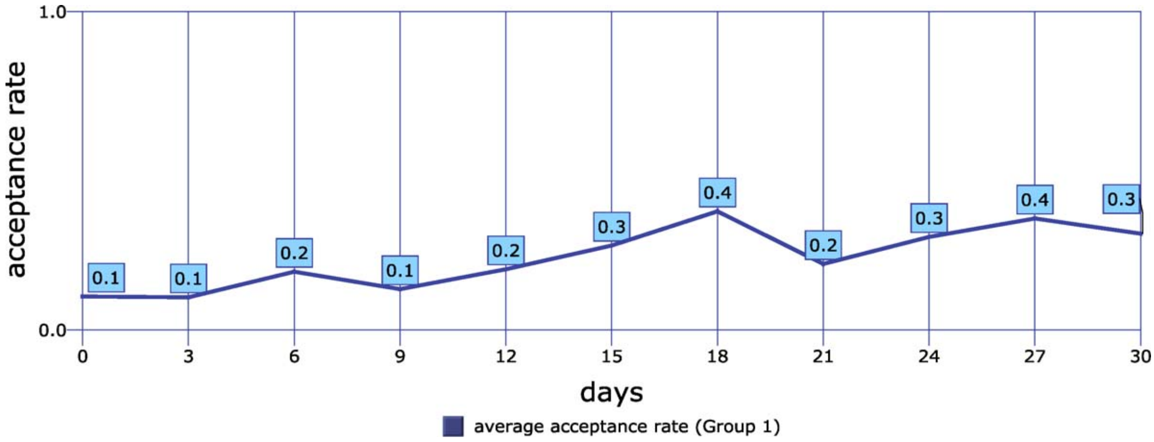


Fig. 16. Users' average acceptance rate for the recommendation based on the implicit interest profiling method within 30 days.

6.2.3. Experiment 3: user profile evaluation

In order to make a comparison between the effects of the refined interest profiling approach and the conventional ontology-based profiling approach, we selected another 5 students to make up a second group, which was a control group (Group 2). During the 30 days, they used recommendation simultaneously with the experimental group, for whom recommendation was produced by the novel approach. The experimental group (Group 1) is made up of the five students in experiments 1 and 2. The difference between the two groups is the method of profiling on which the calculation of their recommendations is based. Conventionally, recommender algorithms can be evaluated with recall and precision metrics (Zhou et al., 2010). In this experiment, we use precision and name it "acceptance rate". Acceptance rate is defined as the average value of the percentages of the items being accepted (clicked) by users in all recommended items, which is recorded every day for both groups. The acceptance rate indicates the degree to which users agree with the recommendation, and it reflects whether the refinement is successful or not. During the experiment, the acceptance rates of the two groups were calculated every day to compare the efficiency of the algorithms. We assume that all the clicks are derived from users' interests but not other intentions, which is usually the case. In this way, as long as our new profiling algorithm is precise and effective enough, the acceptance rate will be increasingly stable and high to a certain degree. Every day, the two average levels of recommendation acceptance rates based on two different approaches were calculated, as illustrated in Fig. 15. We can see that, as time goes by and SPRS accumulates more user behavioral records, the paper recommendation produced by the weighted keyword graph-based profiling approach becomes more highly accepted and steady.

Finally, SPRS also provides paper recommendation according to the implicit interest profiles. These recommendations are presented on single pages, and users are notified of them on the homepage. Similar to the method of evaluation on explicit interest profiles, we also use the acceptance rate to assess the effect of implicit interest profiling. The result of acceptance rate evaluation is shown in Fig. 16. As we can see, users accepted the recommendation according to implicit interest profiles to a rather satisfying degree. This means that, rather successfully, we have predicted what topics these users will potentially like.

7. Conclusion and future work

Recommendation service on academic publications has become a very important research topic due to the development of personalization services and the increased tendency towards electronic academic reading. In this paper, we first introduced a weighted keyword clustering method to extend the subject ontology. The original ontology is a two-level structure which is currently in use, and through our approach, we can extend it to more levels. We then proposed the refined ontological profiling approach to provide paper recommendations to academic paper readers. Two types of interest profiles, explicit and implicit profiles, were introduced. Finally, we conducted three experiments to examine the effectiveness of our methods. One examined the effectiveness of ontology extension; the others assessed the precision of our profiling approaches.

The result of the first experiment shows that the keyword clustering process is able to produce separate small topics for secondary subjects of the ontology and the clustering results vary as a different parameter is adopted. The second experiment result indicates that, compared with the conventional ontology-based profiling method, our ontological profiling approach focuses on the comprehensive description of user interests and has improved the precision of the recommendation. Also, the application of explicit interest profilers is intriguing. Without question, the application of implicit interest profiles makes user interests more comprehensive and rounded, which in a way provides introductory topics to users that they currently may not appear to be interested in.

In the future, we may make modifications and improvements to the weighted keyword graph-based interest profiling approach and the subject ontology extension method. We will improve the keyword clustering algorithm through identifying synonyms among the keywords. Again, the discovery of semantic relationships between keywords is also a possible research objective. In addition, we should try to apply our weighted keyword graph-based interest profiling approach to a collaborative filtering system.

Acknowledgements

This paper is supported partly by the NSFC under Grant No. 60603090 and 50875158; Sci. & Tech. Development Fund of Shandong Province of China (2010GSF10811); Specialized Research Fund for the Doctoral Program of Higher Education of China (20103718110007), Sci. & Tech. Development Fund of Qingdao (10-3-3-32-nsh), Excellent Young Scientist Foundation of Shandong Province of China under Grant No. BS2009DX004; the Special Fund for Fast Sharing of Science Paper in Net Era by CSTD (20093718110008); the Research Foundation of Shandong Educational Committee under Grant No. J08LJ77; and Natural Science Foundation for Distinguished Young Scholars of Shandong and SDUST (JQ200816 and 2010KYJQ101).

Appendix A. List of symbols

Symbol	Meaning	Location (page number)
$WKP(keyword_i,paper_p)$	The weight of keyword $keyword_i$ in $paper_p$	6
PKV_i	The vector space model of $paper_i$	7
$tf(keyword_i,paper_p)$	The frequency of the keyword $keyword_i$ in paper $paper_p$	7
$PKRM$	Paper–keyword relation model	7
PNS	The set of paper nodes in $PKRM$	7
KNS	A set of keyword nodes	7
RS	The set of edges in $PKRM$	8
$r_{i,j}$	An edge linking $keyword_j$ and $paper_i$ in $PKRM$	8
WKG	Weighted keyword graph	8
$Weight(keyword_i)$	The weight of $keyword_i$ in WKG	9
$PS(keyword_i)$	The set of papers containing $keyword_i$	9
CR	The set of edges in WKG	9
$R_{i,j}$	An edge linking $keyword_i$ and $keyword_j$ in WKG	9
$WGT(R_{i,j})$	The weight of $R_{i,j}$	9
Ω	The threshold for the weights of edges in WKG during the keyword clustering process	10
TG_i	A topic	10
TGS	A set of topics	10
$UEIP(User_u)$	The explicit interest profile part of the interest profile of $User_u$	14
$UIIP(User_u)$	The implicit interest profile part of the interest profile of $User_u$	14
$EI(topic_t,user_u)$	The measurement of $User_u$'s explicit interest in $topic_t$	14
$II(topic_\kappa,user_u)$	The measurement of $User_u$'s implicit interest in $topic_\kappa$	14
UTG	User-featured topic graph	15
$IK(keyword_i)$	The weight of keyword node $keyword_i$ in UTG	15
$PS(keyword_i)$	The set of papers containing $keyword_i$	15
$UPS(user_u)$	The set of papers recorded in $user_u$'s behavior history database	15
$UIP(paper_p,user_u)$	Measurement of $user_u$'s interest in $paper_p$ by conventional method	15
$BF(paper_p,user_u)$	A behavior factor in the calculation of $UIP(paper_p,user_u)$	15
$DP(paper_p,user_u)$	The number of days passed since $user_u$ accessed $paper_p$	15
Φ	A parameter for adjustment in the equation of $UIP(paper_p,user_u)$	15
$IES(keyword_i)$	The inner edge strength of $keyword_i$	16
$CES(keyword_j)$	The cross edge strength of $keyword_i$	16
$RS(keyword_i,topic_t)$	The relevance strength of $keyword_i$ to $topic_t$	16
$RF(user_u,topic_t)$	The relevance factor indicating the measurement of the relevance of $user_u$'s interest to $topic_t$	16
$\overrightarrow{TGV}$	A vector comprising the weights of all nodes in $TG(topic_t)$	16
$\overrightarrow{UTGV}$	A vector consisting of the weights of all nodes in $UTG(topic_t)$	16
λ	An adjustable parameter in the calculation of $RF(user_u,topic_t)$	16
$UTS(user_u)$	The set of topics that $user_u$ has explicit interests	16
TNG	Topic network graph	18
$ETNS$	The node set in TNG in which each topic user shows explicit interest	18
$ITNS$	The node set in TNG in which each topic user shows implicit interest	18
ES	The set of edges in TNG	18
$E_{i,j}$	The edge connecting $topic_i$ and $topic_j$ in TNG	18
$WET(E_{i,j})$	The weight of the edge $E_{i,j}$ in TNG	18
$MES(topic_i,topic_j)$	The edge set in which each edge connects one keyword in $topic_i$ and one keyword in $topic_j$ in TNG	18
$VC(topic_i)$	The connectivity of the node $topic_i$ in TNG	19
$Neighbors(topic_x)$	The set of topic nodes neighboring $topic_x$ in TNG	19

References

Adomavicius, G., Tuzhilin, A., 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. 17, 734–749.

Balabanović, M., Shoham, Y., 1997. Fab: content-based, collaborative recommendation. Commun. ACM 40, 66–72.

Batagelj, V., Mrvar, A., 1998. Pajek – program for large network analysis. Connections 21, 47–57.

Blanco-Fernández, Y., Pazos-Arias, J.J., Gil-Solla, A., Ramos-Cabrer, M., López-Nores, M., García-Duque, J., Fernández-Vilas, A., Díaz-Redondo, R.P., 2008. Exploiting synergies between semantic reasoning and personalization strategies in intelligent recommender systems: a case study. J. Syst. Software 81, 2371–2385.

Burke, R.D., Hammond, K.J., Young, B.C., 1997. The FindMe approach to assisted browsing. IEEE Intell. Syst. 12, 32–40.

Cantador, I., Szomszor, M., Alani, H., Fernández, M., Castells, P., 2008. Enriching ontological user profiles with tagging history for multi-domain recommendations. In: 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008), Tenerife, Spain.

Correa da Silva, F.S., Vasconcelos, W.W., Robertson, D.S., Brilhante, V., de Melo, A.C.V., Finger, M., Agustí, J., 2002. On the insufficiency of ontologies: problems in knowledge sharing and alternative solutions. Knowl.-Based Syst. 15, 147–167.

Felden, C., Linden, M., 2007. Ontology-based user profiling. Bus. Inform. Syst. 4439, 314–327.

Guarino, N., Giaretta, P., 1995. Ontologies and knowledge bases – towards a terminological clarification. In: Mars, N.J.I. (Ed.), Towards Very Large Knowledge Bases. IOS Press, Amsterdam, The Netherlands, pp. 25–32.

Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T., 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inform. Syst. 22, 5–53.

Kim, W., Kerschberg, L., Scime, A., 2002. Learning for automatic personalization in a semantic taxonomy-based meta-search agent. Electron. Commer. R. A. 1, 150–173.

Middleton, S.E., Alani, H., Shadbolt, N.R., De Roure, D.C., 2002. Exploiting synergy between ontologies and recommender systems. In: Proceedings of the WWW2002 International Workshop on the Semantic Web, Maui, USA.

Middleton, S.E., Roure, D.C.D., Shadbolt, N.R.,2001. Capturing knowledge of user preferences: ontologies in recommender systems. In: Proceedings of the 1st International Conference on Knowledge Capture. ACM, Victoria, British Columbia, Canada, pp. 100–107.

Middleton, S.E., Shadbolt, N.R., Roure, D.C.D.,2003. Capturing interest through inference and visualization: ontological user profiling in recommender systems. In: Proceedings of the 2nd International Conference on Knowledge Capture. ACM, Sanibel Island, FL, USA, pp. 62–69.

Middleton, S.E., Shadbolt, N.R., Roure, D.C.D., 2004. Ontological user profiling in recommender systems. ACM Trans. Inform. Syst. 22, 54–88.

Mika, P., 2007. Ontologies are us: a unified model of social networks and semantics. Web Semant. 5, 5–15.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.,1994. GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. ACM, Chapel Hill, NC, United States, pp. 175–186.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manag. 24, 513–523.

Sciencepaper Online, 2010. Sciencepaper Online, Last Visited 1 October, 2010. http://www.paper.edu.cn/en.

Sebastiani, F., Ricerche, C.N.D., 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1–47.

Shepherd, M., Watters, C., Marath, A.,2002. Adaptive user modeling for filtering electronic news. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02). IEEE Computer Society, Hawaii, USA, pp. 1180–1188.

Smyth, B., Bradley, K., Rafter, R., 2002. Personalization techniques for online recruitment services. Commun. ACM 45, 39–40.

Susan, J.T., Gauch, S.,2004. Improving ontology-based user profiles. In: Proceedings of the Recherche d'Information Assiste par Ordinateur, RIAO 2004. University of Avignon (Vaucluse), France, pp. 380–389.

Tan, A., Teo, C., Keng, H.M.,1998. Learning user profiles for personalized information dissemination. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN'98). IEEE Computer Society, Anchorage, AK, USA, pp. 183–188.

Wikipedia, 2010. Recommender System – Wikipedia, The Free Encyclopedia, Last Visited 1 October, 2010. http://en.wikipedia.org/wiki/Recommender_systems.

West, D.B., 2000. Introduction to Graph Theory, second ed. Prentice Hall, Englewood Cliffs, NJ.

Wu, Z., Zeng, Q., Hu, X., 2009. Mining personalized user profile based on interesting points and interesting vectors. Inform. Technol. J. 8 (6), 830–838.

Yang, Y.,1995. Noise reduction in a statistical approach to text categorization. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Seattle, WA, United States, pp. 256–263.

Zanin, M., Cano, P., Buldú, J.M., Celma, O.,2008. Complex networks in recommendation systems. In: Proceedings of the 2nd WSEAS International Conference on Computer Engineering and Applications. World Scientific and Engineering Academy and Society (WSEAS), Acapulco, Mexico, pp. 120–124.

Zeng, Q., Zhao, Z., Liang, Y., 2009. Course ontology-based user's knowledge requirement acquisition from behaviors within E-learning systems. Comput. Educ. 53 (3), 809–818.

Zhang, H., Li, Y., Tan, H.B.K., 2010. Measuring design complexity of semantic web ontologies. J. Syst. Software 83, 803–814.
Zhou, T., Kuscsik, Z., Liu, J., Medo, M., Wakeling, J.R., Zhang, Y., 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. Proc. Natl. Acad. Sci. U.S.A. 107, 4511–4515.
Zhou, T., Ren, J., Medo, M., Zhang, Y., 2007. Bipartite network projection and personal recommendation. Phys. Rev. E 76, 046115.

Qingtian Zeng is a full professor at the College of Information Science and Technology, Shandong University of Science and Technology. He obtained his PhD in computer software and theory from Institute of Computing Technology at Chinese Academy of Sciences in 2005. His research interests are in the areas of Petri nets, Process Mining, Domain Ontology, and Personalized Information Service.

Xiaoyu Tang is a master student at the College of Information Science and Technology, Shandong University of Science and Technology. His research interest is Personalized Information Service.