Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Future Generation Computer Systems 20 (2004) 47–60

FGCS FUTURE GENERATION COMPUTER SYSTEMS

www.elsevier.com/locate/future

# Semantic profile-based document logistics for cooperative research☆

Hai Zhuge*, Yanyan Li

*Knowledge Grid Research Group, Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, P.O. Box 2704-28, 100080 Beijing, China*

**Abstract**

This paper proposes a document logistics approach for cooperative research based on the Web and Knowledge Grid. The approach realizes effective research document collection, organization and provision as well as knowledge sharing by incorporating the following functions: construction of semantic profiles representing interests, continuous discovery and collection of potentially relevant documents, synthesis of evaluation feedbacks, and support of flexible management operations and document recommendation services. The prototype has been implemented and is available for use online. Experiments show that the proposed approach is feasible and effective.


*Keywords:* Feedback synthesis; Knowledge Grid; Logistics; Profile; Teamwork

## 1. Introduction

Traditionally, logistics refers to the management of inter-related business activities whose objective is to move objects between origins (e.g., production) and destinations (e.g., consumption) in a timely fashion [6,7]. It concerns process management (e.g., supply chain, workflow, etc.), coordination, planning, and execution control as well as application platforms.

The rapid growth of the number of research documents available on the Web has led to researchers constantly fighting information overload in their pursuit of knowledge. Keeping up-to-date with documents and finding relevant documents are becoming increasingly difficult. Though the Scientific Literature Digital Library [11,12] and other citation indices of scientific literature (such as LANL e-Print archive, NCSTRL, UCSTRI, LTRS, etc.) have alleviated the information overload to a certain extent, researchers have to still expend a great deal of time and effort looking for new documents that may interest them. So how to effectively and orderly process and manage information becomes an important issue.

Information logistics is a technology that aims to efficiently collect, organize and provide personalized heterogeneous information on demand. *Document logistics* is a special case of information logistics, which aims at enhancing the cooperation and efficiency of research groups. In general, effective cooperative research concerns: complete collection of relevant research documents, effective sharing of documents and feedbacks to avoid redundant efforts, efficient organization of relevant documents, and recommendation of up-to-date documents related to researchers' interests.

☆ The research work was supported by the National Science Foundation of China (NSFC).
* Corresponding author. Fax: +86-1-062567724.
*E-mail address:* zhuge@ict.ac.cn (H. Zhuge).

0167-739X/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-739X(03)00164-X

Knowledge Grid is a platform that enables sharing and managing the distributed heterogeneous resources (including information, knowledge and services) spread across the Internet in a uniform way. It includes multi-dimensional resource spaces (such as knowledge space and information space) and resource operations that enable users to store and access the resources with different privileges including personal-privacy, group-privacy and public sharing [20]. The representation of resources in the Knowledge Grid is based on the markup languages like XML and RDF in the Semantic Web [1,8,10].

Based on the Web and Knowledge Grid, this paper proposes a document logistics approach serving research groups across the Internet. This approach enables group members to collect, organize, access and share research documents in a more effective cooperative manner.

## 2. Document logistics framework

A framework of document logistics is illustrated in Fig. 1. The *initial definition* module enables research groups to choose the construction mode of profiles and specify the keywords and constraints. The Knowledge Grid herein comprises two resource spaces: *knowledge space* and *information space*. The former stores the evaluations and pre-designed knowledge cooperation rules [23], and the latter stores the research documents while providing information support to the knowledge space. By consulting the cooperation rules in the Knowledge Grid and the initial definition, the *logistics engine* not only monitors and controls the cooperation process but also constructs and updates the profiles. The *document evaluation* module is responsible for evaluating the documents based on user feedbacks, and the evaluation information is put into the Knowledge Grid for later reference. The *document collection* and *classification* module automatically collects documents from the Web and from group members, and then all the collected documents are classified and stored in the Knowledge Grid. With the support of *document provision* module, group members can access or share the documents by means of pull (reacting to user management operations) or push (proactive recommendation of resources that match user personal profile) facility. The *profile base* is composed of *collective profiles* and *personal profiles* representing the interests of groups and members, which provides support to document collection and document classification as well as document provision.

## 3. Construction of semantic profiles

### 3.1. Outline

The use and learning of user profiles to improve the quality of information filtering has been studied

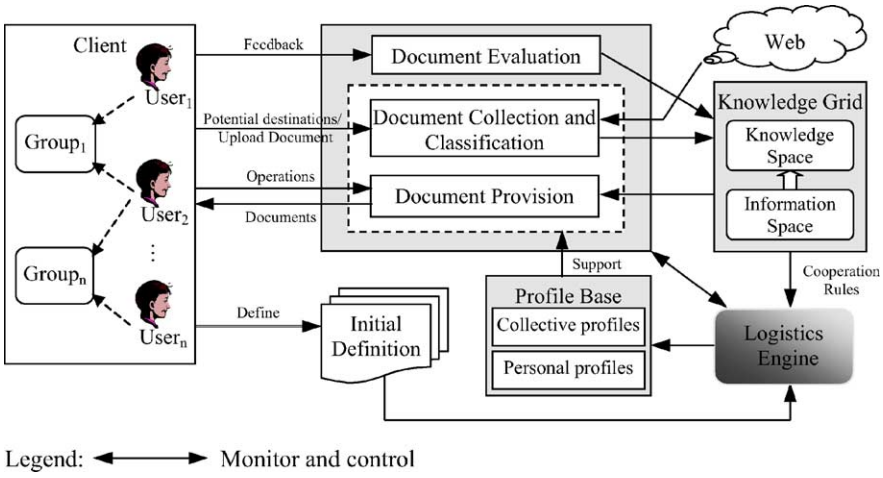



Fig. 1. The framework of document logistics platform.

[2–4,19]. In this paper, the profiles are dynamically constructed according to different purposes (e.g. resources retrieval, helper location, promoting reflection, etc.), which comprise knowledge level, interest, helpfulness, time constraints, etc. We adopt two modes including loosely coupled mode and tightly coupled mode to construct and update two-level profiles: the collective profile reflecting the interests of a research group and the personal profile reflecting the personal interest (each group member can have multiple personal profiles). The first mode is that group members can contribute to the personal profiles explicitly by manually editing and tuning the profiles through the system's interface at anytime, and the relevant documents and queries are recorded by tracking group member's searching and browsing behaviors. These documents reflect each member's personal interests, so they are added into a training set that is used to construct a profile. Queries usually reflect group member's interests directly, so the keywords in the frequently asked queries can be extracted and used as core keywords. By incorporating the time constraints [21], the second mode is suitable for the occasion on which the researcher has clear plan about his research interests in a certain period (e.g. conduct a project within the planned time). The system will automatically collect the documents related to his specified interests in advance, and then recommend different cluster of documents in different period. In build-time, the researcher describes his different interests in different phase, and the logistics engine will then form the candidate profiles denoted as $\langle P(\Delta t_1), P(\Delta t_2), \ldots, P(\Delta t_n)\rangle$, where $P$ denotes time-related profile and $\Delta t_i$ denotes the working duration of the profile. In run-time, the logistics engine checks the time constraint of all candidate profiles and selects the $P(\Delta t_i)$ as working profile if $t \in \Delta t_i$. Furthermore, regarding both of the two modes, peer evaluations can be referred to assess a learner's knowledge level and helpfulness so as to update personal profiles. All the group member's personal profiles are combined to form a collective profile standing for the interests of a research group, so the newly joined member can initially adopt the collective profile as his personal profile and modify it later.

This section mainly describes the interest discovery approach. Researchers can roughly describe their interests by specifying concepts (terms or term-phrases) and corresponding weights. The weight denotes the importance degree of a concept in a profile, and the default value is 1. However, the description information is limited and subjective, so a learning process is conducted on a training set of documents (the number of documents increases over time) to discover more closely related concepts and modify the association weights between concepts in a profile. In this way, a profile is continually refined till reaching a steady state. As a result, a profile comprises concepts and relationships between concepts, which represents one of the interests of researchers in a certain period. The learning process includes *concept extraction*, *co-occurrence analysis* and *authority identification*.

### 3.2. Concept extraction

Concepts usually refer to the terms or term-phrases. Based on Salton's approach [16], we extract the concepts from the documents by following a four-step process: individual term identification, stop-wording, word stemming, and term-phrase formation. A stop-word list is used to eliminate the noises or useless words such as "the", "a", "on", "in", etc. A stemming algorithm is used to unify different forms of a word. The term-phrase formation formulates phrases by combining only adjacent words. After the concepts are extracted, we compute the information gain for each concept because the remaining texts still contain many concepts. Finally, select the concepts with the information gain bigger than the threshold to characterize a document.

It is well-known that researchers always search research documents by inputting keywords or author name, so we deal with the concepts representing the document content feature and author name separately. We standardize all author names according to the format of last name, followed by the first character of the first name. This helps to remove the problem of the same names appearing in different forms.

### 3.3. Co-occurrence analysis

After concepts are identified, we perform the co-occurrence analysis by adjusting the concept space approach [5,9,15]. Usually concepts that occur in different locations have different descriptive abilities, for example, concepts identified in the title of a doc-

ument are more descriptive than concepts identified in the abstract of a document.

Let $T$ = {Title, Keywords, Abstract, Body, Conclusion, Reference} be a set of identified document fragments, $W_X$ be the weight of the $X$ ($X \in T$) in a document. Users can determine the order of the weight of $W$, for example: $1 > W_{\text{title}} > W_{\text{keywords}} > W_{\text{abstract}} > W_{\text{body}} > W_{\text{conclusion}} > W_{\text{reference}} > 0$. We use the following formula to compute the weight of a concept $j$ (exclude author name) in document $i$ denoted as $d_{ij}$ based on TFIDF (the product of "term frequency" and "inverse document frequency")

$$d_{ij}=\frac{\sum_{X\in T}(W_X \times tf_j^X) \times \log(|D|/\mathrm{d}f_j \times n_j)}{\sqrt{\sum_{k=1}^{t}\left(\sum_{X\in T}(W_X \times tf_k^X)\times\log(|D|/\mathrm{d}f_k\times n_k)\right)^2}}, \tag{1}$$

where $tf_j^X$ denotes the number of occurrences of concept $j$ in the $X$ location of document $i$, $|D|$ represents the total number of documents in a training set, $n_j$ represents the number of words of concept $j$, $\mathrm{d}f_j$ represents the number of documents that include the concept $j$, $t$ represents the total number of concepts in the $i$th document.

Based on formula (1), the association weight between two concepts $j$ and $k$ ($\mathrm{AW}_{jk}$) can be computed as follows:

$$\mathrm{AW}_{jk} = \begin{cases} \dfrac{\sum_{i=1}^{|D|} d_{ijk}}{\sum_{i=1}^{|D|} d_{ij}} \times \dfrac{\log(|D|/\mathrm{d}f_k)}{\log|D|}, & \text{association weight from the concept } j \text{ to the concept } k \text{ (concept space)}, \\ \dfrac{\sum_{i=1}^{|D|}(tf_{ijk} \times d_{ij})}{\sum_{i=1}^{|D|} tf_{ijk}}, & \text{association weight between the concept } j \text{ and the author name } k, \end{cases} \tag{2}$$

where $d_{ijk} = tf_{ijk} \times \log(|D|/\mathrm{d}f_{jk} \times n_j)$ represents the combined weight of both concept $j$ and concept $k$ in the $i$th document, $tf_{ijk}$ represents the number of occurrences of both concept $j$ and concept $k$ in document $i$ (the smaller number of occurrences between the concepts is chosen), and $df_{jk}$ represents the number of documents (in a collection of $|D|$ documents) in which concept $j$ and $k$ occur together. The association weight between two concepts is asymmetric. For example, the association weight from "meta-search" to "search engine" is obviously different from the association weight from "search engine" to "meta-search".

Using the co-occurrence analysis approach, the system computes the association weights of concept associations between a profile and documents, and the weight of each extracted concept according to the following two alternative strategies: (1) maximum-value strategy: $w_j = \mathrm{Max}(w_i^k \times \mathrm{AW}_{ij})$; (2) average-value strategy: $w_j = (\sum_{i=1}^{m} w_i^k \times \mathrm{AW}_{ij})/m$, where $w_j$ denotes the weight of the $j$th expanded concept, $w_i^k$ denotes the weight of the concept $i$ in the $k$th profile, and $m$ denotes the number of concepts in the $k$th profile. Following that, the system selects the concepts whose weights are bigger than the threshold, and finally adds the selected concepts into the profile. The system sets the initial threshold and adjusts it later during execution.

As the number of documents in a training set $D$ is increased, the profiles constructed at different time are different. We use the formula $\delta = \sum_{i=1}^{m}(w_i^k(t+1) - w_i^k(t))^2/m$ to compute the difference between the profile constructed at time $t+1$ and the profile constructed at time $t$, where $w_i^k(t)$ denotes the weight of concept $i$ belonging to the $k$th profile constructed at time $t$, $m$ denotes the larger number of concepts in the profiles constructed at time $t$ and $t+1$, respectively. The weights of concepts that do not belong to the profile are assigned 0. If $\delta$ is less than a predefined threshold, then the profile construction process is terminated.

### 3.4. Authority identification

Authorities indicate the well-known journals, conferences, experts, documents, communities and websites. Group members can input the authorities through the interface. The heuristic rules and statistics method based on the association weight are used to automatically identify the authorities. The DBLP server (http://dblp.uni-trier.de/) provides bibliographic information on major computer science journals and proceedings, based on which, the system records well-known journals in the sub-domains of computer science. The authority conferences can be identified

by referring to the conference organizer, participated leading researchers, and the proceeding publisher. Based on the association weight between authors and concepts, we select the author with the maximum association weight as authority expert. As for the authority documents, we will consider several factors including the number of being cited, publishing date, author and publication. The authority websites refer to the homepages of the authority experts or the famous domain-specific websites.

## 4. Document collection and classification

Based on the search engine and Web crawling technologies, the system automatically collects research documents to increasingly enrich the Knowledge Grid with the following three methods:

1. *Upload manually*. Each member can manually upload documents with corresponding semantic annotations (e.g. content description, access privilege) through the interface.
2. *Customized collection*. Research groups can specify the potential destinations in which they have constant interests. We have developed the collection tool GruDexer to continuously search for new relevant documents from the specified websites. The found documents are downloaded, parsed, and placed into the Knowledge Grid with group-privacy access privilege. This method enables a research group to keep track of the latest research information from a particular researcher or research group. As a complement, the GruDexer automatically crawls on the Web to fetch documents and store in the Knowledge Grid with public access privilege. As some websites only provide abstracts of documents, the meta-information of a document and abstracts are fetched for later consult.
3. *Meta-search engine*. Limitations of the single search engines have led to the introduction of meta-search engines [17,18]. GruDexer acts as a meta-search engine based on multiple search engines such as CiteSeer and Scirus. Users can input the keywords to search for relevant documents by means of GruDexer. Fig. 2 shows the returned

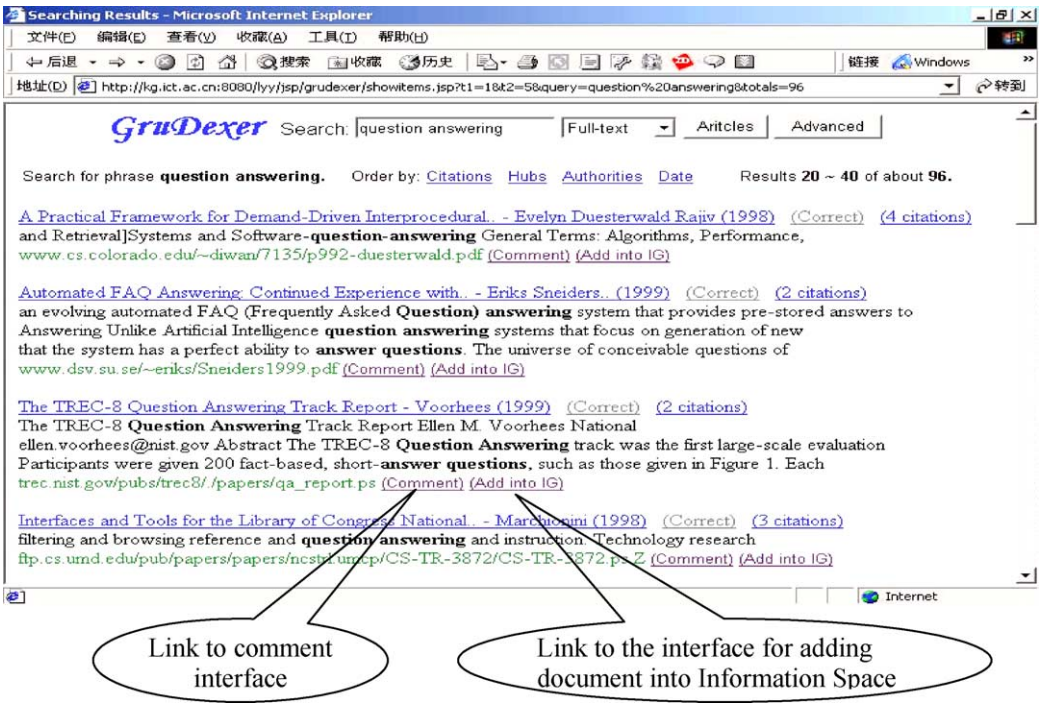



Fig. 2. Search results of meta-search engine GruDexer.

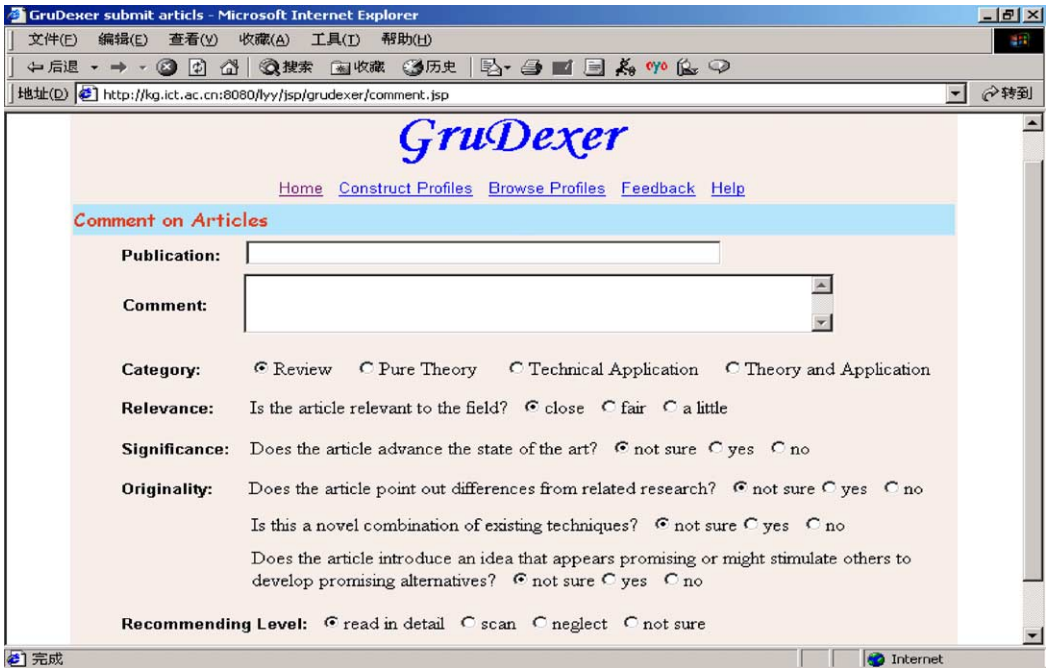



Fig. 3. Interface for commenting on document.

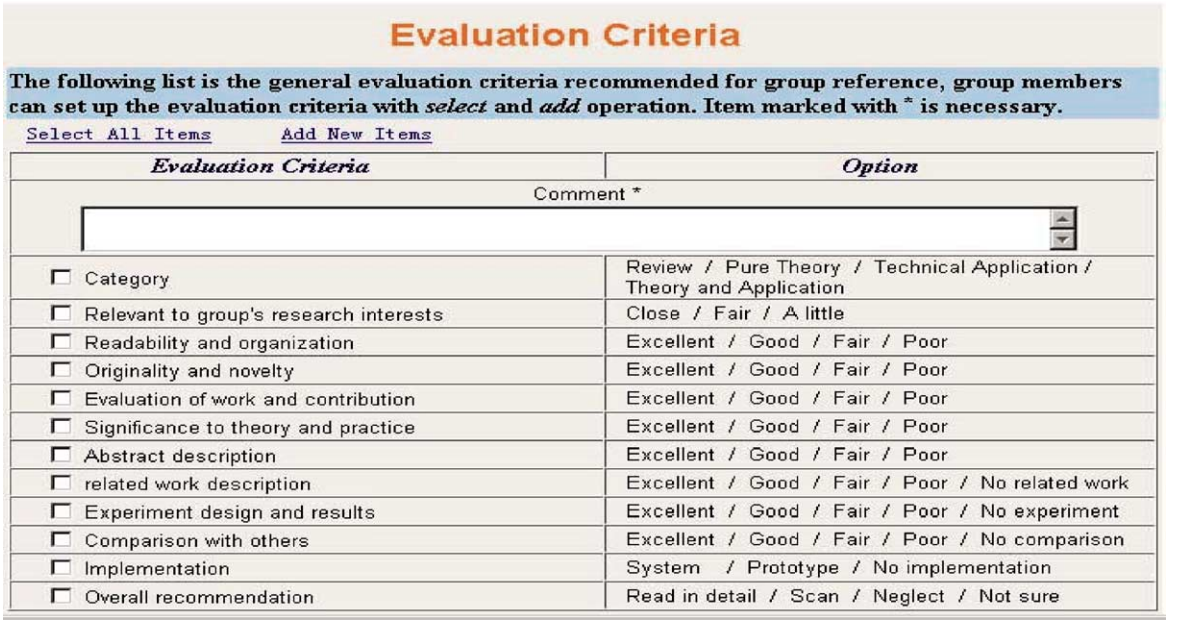
documents related to the phrase "question answering". Different from the interface for displaying search results of traditional search engines, users can choose the *Comment* link to evaluate the document or choose the *Add into IG* link to add the document into the information space with specified personal-privacy or group-privacy access privilege. GruDexer will display the interface shown in Fig. 3 for accepting comment on the document when the user clicks the comment link.

Inspired by the approach of extracting the classification knowledge of Web pages by mining term correlation [13], we make use of the profiles to assist document classification. Each research group can specify the semantic categories according to their research interests, and the assigned concepts are generalized as representative classification knowledge. Furthermore, associated concepts in the profiles are applied to refine the classification knowledge. So, the classification knowledge is a set of concepts with semantic associations. By using the Cosine method, all the collected documents are classified and put into the appropriate category of information space.

## 5. Document evaluation based on user feedback

### 5.1. Dynamic generation of group-oriented evaluation criteria

Peers' evaluations on documents are useful for researchers' reference. We adopt a dynamic mechanism for research groups to generate the evaluation criteria according to their own research goals. By consulting experts and research students, we firstly generate a set of general evaluation criteria. Whenever a new research group registers successfully, the general evaluation criteria are provided for reference, the research group can either inherit the evaluation criteria and make some modifications or generate a new set of evaluation criteria. Furthermore, the research group can specify the importance weight of each evaluation criterion. Fig. 4 shows the interface for generating evaluation criteria where a group can modify their evaluation criteria anytime. According to the evaluation criteria made by a group, the system will then automatically generate a new feedback interface for the group members to evaluate the document they read.



**Evaluation Criteria**

**The following list is the general evaluation criteria recommended for group reference, group members can set up the evaluation criteria with *select* and *add* operation. Item marked with * is necessary.**

Select All Items  Add New Items

| *Evaluation Criteria* | *Option* |
|---|---|
| Comment * | |
| Category | Review / Pure Theory / Technical Application / Theory and Application |
| Relevant to group's research interests | Close / Fair / A little |
| Readability and organization | Excellent / Good / Fair / Poor |
| Originality and novelty | Excellent / Good / Fair / Poor |
| Evaluation of work and contribution | Excellent / Good / Fair / Poor |
| Significance to theory and practice | Excellent / Good / Fair / Poor |
| Abstract description | Excellent / Good / Fair / Poor |
| related work description | Excellent / Good / Fair / Poor / No related work |
| Experiment design and results | Excellent / Good / Fair / Poor / No experiment |
| Comparison with others | Excellent / Good / Fair / Poor / No comparison |
| Implementation | System / Prototype / No implementation |
| Overall recommendation | Read in detail / Scan / Neglect / Not sure |

Fig. 4. The interface for generating evaluation criteria.

### 5.2. Evaluation synthesis

As Fig. 4 illustrates, the feedback interface includes a comment textbox and multiple evaluation criteria. Each evaluation criterion has several options that indicate the interestingness of a document. For example, options to the *overall recommendation* item reflect the document relevance degree, and their corresponding scores are fixed in the current system. At the present stage, we simply combine the comments inputted by group members to form one text, and compute the overall evaluation score of a document by considering all the evaluation criteria.

Credibility is used to indicate a group member's reliance degree for his offered evaluation information. A member's credibility is increased if other members take the corresponding behaviors (such as view, download and ignore) according to his recommendation. In addition, a member's credibility is also increased if his evaluation on a document is consistent with that of most of members. On the contrary, his credibility is decreased if his evaluation on a document is inconsistent with that of most of members. Therefore, based on the statistics method, we use the following formula to compute the credibility of the $i$th member denoted as $\mathrm{CR}_i$ by tracking his browse behaviors

$$\mathrm{CR}_i = \frac{E_i}{T_i} \times W_E + \frac{P_i}{E_i} \times W_P - \frac{N_i}{E_i} \times W_N, \tag{3}$$

where $E_i$ denotes the number of documents that are evaluated by the $i$th member, $T_i$ denotes the total number of documents browsed by the $i$th member, $P_i$ denotes the number of documents on which the $i$th member's evaluation is confirmed by others, $N_i$ denotes the number of documents on which the $i$th member's evaluation is negated by others. $W_E$, $W_P$ and $W_N$ are respectively the assigned score with respect to each case.

Based on formula (3), we use the following two formulas to respectively compute the evaluation score corresponding to the $k$th evaluation criterion ($S_k$) and the overall evaluation score of the $j$th document ($E_j$)

$$S_k = \frac{\sum_{i=1}^{n} \mathrm{CR}_i \times V_{jk}^i}{n}, \tag{4}$$

$$E_j = \frac{\sum_{k=1}^{m} ew_k \times S_k}{m}. \tag{5}$$

Suppose $n$ is the number of members who give evaluation for the same document, $V_{jk}^i$ is the score assigned

to the $k$th evaluation criterion of the $j$th document according to the $i$th member's option, $m$ is the total number of evaluation criteria, $ew_k$ is the importance weight assigned to the $k$th evaluation criterion and the default value is 1.

## 6. Document provision mechanism

### 6.1. Management operations

The Knowledge Grid provides users with a set of operations including *put*, *get*, *browse*, *delete*, etc. With these management operations, users can cooperatively manage the research documents in the Knowledge Grid by following three steps: (1) select suitable operation; (2) locate the correct category; (3) set the parameter and submit to the execution engine. The execution engine is responsible for explaining and executing the operation, and finally the execution results are returned to the users. We herein mainly illustrate the document retrieval approach in terms of *get* operation.

By making use of the keyword-based approach and the PageRank method [14], we propose a profile-based matching approach to make an estimation of document relevance, which considers two factors: the semantics associative keywords and the citation times that reflect the quality of a paper to a certain degree. Whenever a user inputs keywords, the system firstly determines the appropriate profile according to the keyword matching method. Since the keywords inputted by a user may be too abstract or ambiguous, the system adopts the following two strategies to search for the matching documents: (1) use the keywords provided by the user; (2) use the keywords provided by the user as well as the expanded keywords in the appropriate profile. After locating the matching documents, the system computes the similarity score between each document and the appropriate profile with the Cosine method, and then selects the documents whose similarity scores are bigger than the threshold. Following that, the system re-ranks the selected documents by further considering their average citation times per year. Finally, the documents are displayed to the user in the order of the documents' scores. As a complement, authorities can be used directly to determine a document's relevance degree with the user's interests in some cases (e.g. track the specified author's latest research direction). Formula (6) computes the similarity score ($S_{ik}$) between the $i$th document and the $k$th profile, and formula (7) computes the score ($R_i$) of the $i$th document:

$$S_{ik} = \frac{X_i P_k}{|X_i||P_k|}, \tag{6}$$

$$R_i = \tau \times S_{ik} \times \frac{\sum_{u \in B_i} (R_u / N_u) + \varepsilon}{(t - t_i) + 1}, \tag{7}$$

$X_i = \langle x_1^i, x_2^i, \ldots, x_n^i \rangle$ is a feature vector of the $i$th document where each component indicates the importance degree of a concept in the document. $P_k$ is the selected $k$th profile that can be denoted as a vector $P_k = \langle w_1^k, w_2^k, \ldots, w_t^k \rangle$, $t$ and $w_j^k$ respectively denote the number of concepts and the weight of the $j$th concept in the $k$th profile, $B_i$ is the set of documents that cite the $i$th document, $N_u$ the number of citations of the $u$th document, $\varepsilon$ is a adjustment factor to avoid the numerator is zero when the $i$th document has not been cited, and here $\varepsilon$ is initially assigned 0.01, $t$ and $t_i$ are respectively the current year and the publication year of the $i$th document, and $\tau$ is a factor used for normalization.

The documents can also be listed in the order of their evaluation score (as introduced in Section 5.2), publishing date, or citation times. In addition, the metadata (the URL from which each document is linked, the publication, the author name, etc.) is a descriptive tag associated with a document, so it may provide useful information about the relevance of a document. Fig. 5 shows the results of a *get* operation with the keyword as "meta-search". In this example, the documents are shown in the order of evaluation score. In order to get a view of evaluations on this document, group members can simply click the *Evaluation* link. The member can also add new evaluation through the same interface.

### 6.2. Recommendation

In order to effectively support cooperative research, three types of recommendation are provided as follows:

1. *Document recommendation*. The system recommends the top $N$ documents based on user's evaluation feedbacks, where $N$ denotes the number of recommended documents, which can be specified by the users themselves. On the other hand,

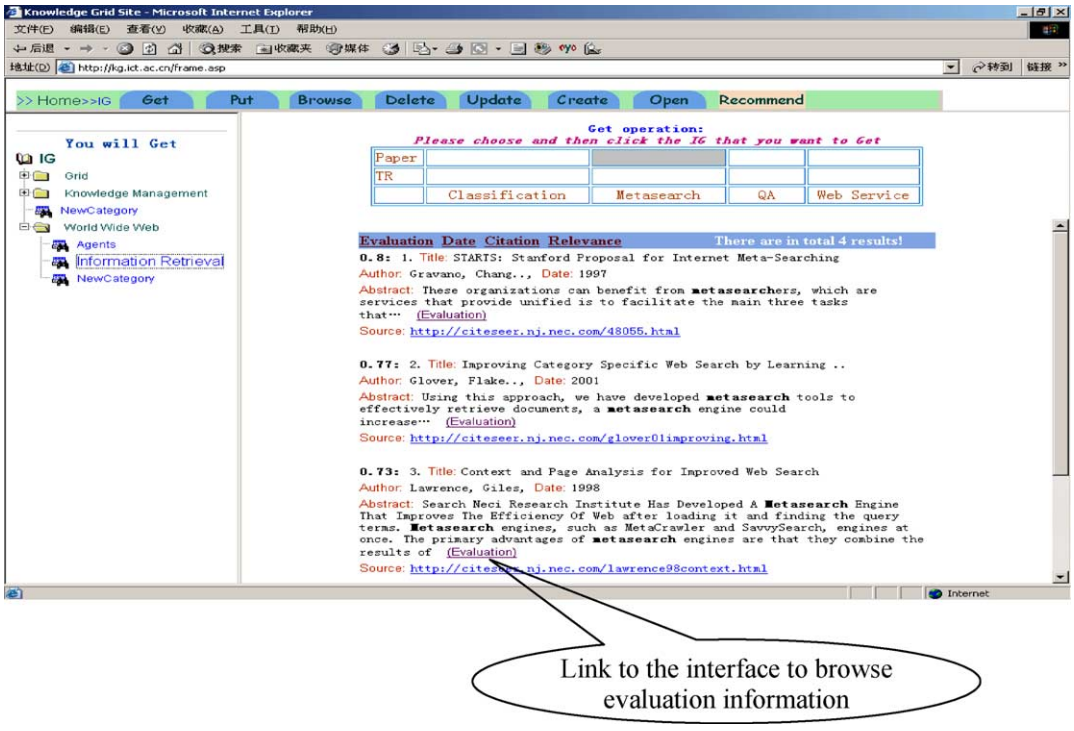



Fig. 5. Interface for displaying the results of *get* operation.

the newly fetched documents are recommended based on the profile-based matching approach. Whenever a user begins a new session of using the Knowledge Grid, he or she can be alerted to the existence of new recommendations. If the user chooses to display the recommendation page, the new recommended documents are displayed along with the related materials. The system can also check all of the existing user profiles daily for new matches, and inform users of new relevant research documents by email.

2. *Helper recommendation*. By comparing the personal profiles of different users, a helper with the similar interest is recommended for further discussion and learning.
3. *Summarization recommendation*. In order to speed up the learning process of a newly joined member, a summarization is provided to give a summary of related works and related concepts (such as keywords, author name, etc.) in accordance with a query. We currently adopt simple heuristics and synthesized evaluations to locate the documents with good introduction of related works. One heuristic is that a document has a separate section of "related works" or "background", and another one is that there are many citations in the section of "related works" or "introduction". Then, the specific paragraph is extracted from the selected document while the related concepts are extracted from the corresponding profiles. Fig. 6 shows the interface of a recommendation page, where the related concepts and related works about "Semantic Web" are shown. Users can click the *resource paper* link to read details or click the *related concepts* links to track more information.

## 7. Experiments and comparisons

### 7.1. Experiment 1

The goal of the first experiment is to construct a collective profile of a research group by discovering the expanded concepts and corresponding weights in the

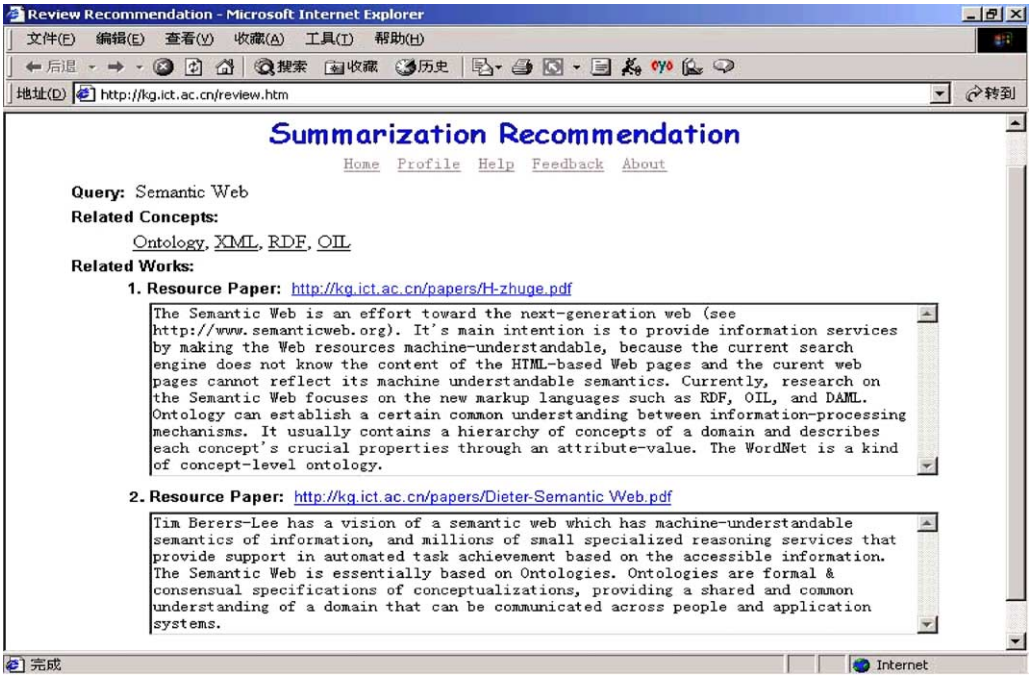



Fig. 6. Summarization recommendation interface.

Table 1
Expanded concepts

| Concept | Expanded concepts | | | | | | |
|---|---|---|---|---|---|---|---|
| Web information retrieval ($C_1$) | Search engine ($C_{11}$) | SIGIR Conference ($C_{12}$) | Classification ($C_{13}$) | Query processing ($C_{14}$) | Link analysis ($C_{15}$) | Steve Lawrence ($C_{16}$) | http://www.haifa.il.ibm.com/wenor/ ($C_{17}$) |
| Personalized service ($C_2$) | Profile construction ($C_{21}$) | Web service ($C_{22}$) | Web log records ($C_{23}$) | Web source discovery ($C_{24}$) | User behavior ($C_{25}$) | Recommendation ($C_{26}$) | |
| Similarity measure ($C_3$) | Euclidean distance ($C_{31}$) | Vector space ($C_{32}$) | Classification ($C_{33}$) | Relevance score ($C_{34}$) | | | |
| Question answering ($C_4$) | TREC Conference ($C_{41}$) | Redundancy elimination ($C_{42}$) | NLP ($C_{43}$) | | | | |
| Meta-search ($C_5$) | Search engine ($C_{51}$) | Semi-structure data integration ($C_{52}$) | Relevance feedback ($C_{53}$) | Text combination ($C_{54}$) | | | |
| Text extraction and summarization ($C_6$) | Term selection ($C_{61}$) | Redundancy elimination ($C_{62}$) | Text mining ($C_{63}$) | Text combination ($C_{64}$) | Semantics ($C_{65}$) | | |

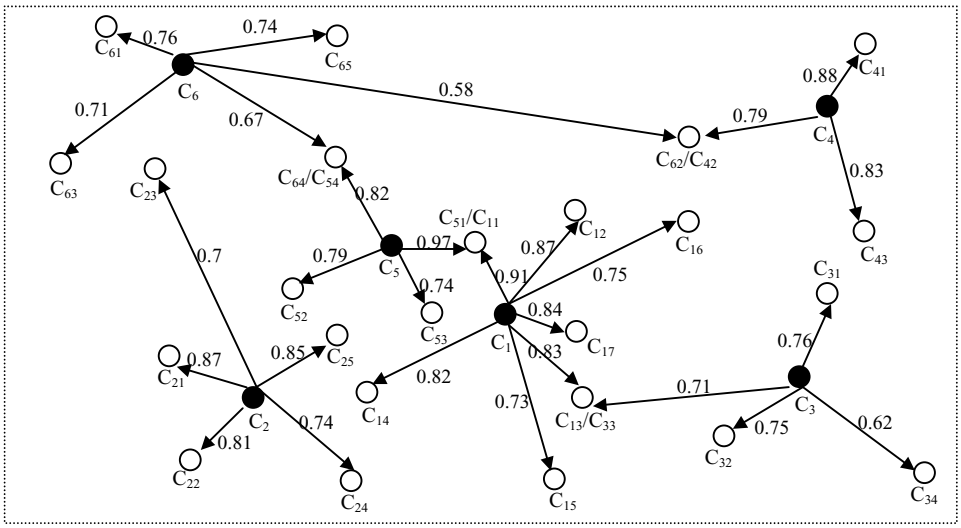



Fig. 7. Concepts association of a collective profile.

specific domain with aforementioned maximum-value strategy. Five groups of domain experts (each group includes three experts) are asked to present concepts in "Information Retrieval" domain and specify the association weights to denote their relationships. Most of the suggested concepts are term-phrases. A small research group whose interest is also "Information Retrieval" takes part in the experiment. The group includes three members who are all research students. When constructing the collective profile, the three members are firstly asked to give several keywords to represent the interests of their group, and the self-weights of these core keywords are assigned 1. Secondly, we collect relevant documents from two sources: journal abstracts (306 documents) and conference proceedings papers (784 documents). Lists of author name and concepts were extracted to compute the association weight between them. Table 1 lists some of the expanded concepts. Due to the space limitation, Fig. 7 mainly illustrates the association weights between the specified concepts and the expanded concepts.

Table 2 shows the comparison between the concepts recommended by the first group of experts and the automatically discovered concepts according to the small-scale experiment data. In this table, the coverage percentage and precision percentage are respectively computed by dividing the *expert-recommended* item and *automatic-generated* item by the *common concepts* item. The result does not provide precise quantitative analysis because it depends on the subjective experience of experts. So, we invite five groups of experts to conduct the same experiment with the aim to show the objective results.

Table 2
Comparison between the automatic-generated profile and the expert-recommended profile

| Object | Automatic-generated | Expert-recommended | Common concepts | Coverage percentage | Precision percentage |
|---|---|---|---|---|---|
| Total concepts | 29 | 27 | 23 | 81.5 | 79.3 |
| Concepts | 25 | 21 | 19 | 90.5 | 76 |
| Author | 1 | 1 | 1 | 100 | 100 |
| Conference | 2 | 2 | 2 | 100 | 100 |
| Paper | 0 | 1 | 0 | 0 | |
| URL | 1 | 1 | 1 | 100 | 100 |
| Community | 0 | 1 | 0 | 0 | |

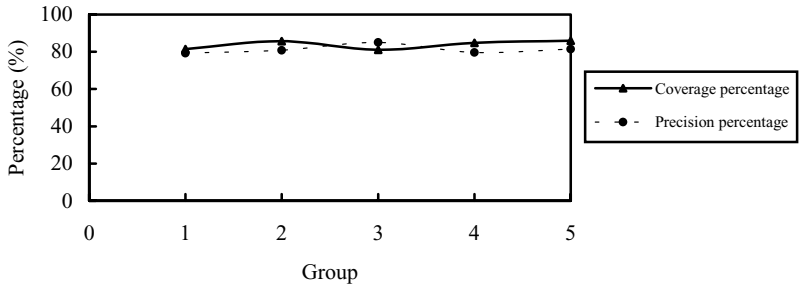



Fig. 8. The comparison results for five groups of experts.

Fig. 8 intuitively shows the comparison results for the five groups of domain experts. Although the coverage percentage and the precision percentage may vary with different groups of experts, we find that the coverage percentages are between 81% and 86% while the precision percentages are between 79% and 85%. The experiment results are relatively precise and steady, which indicates the feasibility and the effectiveness of the proposed approach.

### 7.2. Experiment 2

With the encouraging results obtained from the first experiment, we proceed to integrate the profile into our system and conduct a follow-up experiment to test the effectiveness of document retrieval for small-scale research groups.

Two research groups take part in the experiment. Each group includes five research students who have 1-year research experience in the same research domain. There are two hypotheses: one is that each student provides the same number of feedbacks each time while using the system, and the other is that the testing research documents are predefined closed-corpus, not adding new documents during the test period. The first group (denoted as $G_1$) has used our system for 3 months and the second group (denoted as $G_2$) used our system only several times. The document retrieval experiment is conducted based on the test set of 300 research documents among which 60 research documents are identified related to their interests by the two groups. Suppose the two groups input the same keywords, the system uses two methods (keyword matching and profile-based matching) to retrieve relevant documents in the test set. Regarding the profile-based matching, the system uses only the keywords inputted by group members to search for documents (as introduced in Section 6.1). Table 3 records the retrieval results. In this table, $T_i$ denotes the $i$th duration and herein we take 15 days as a duration, *total documents retrieved* denotes the number of all retrieved documents, *relevant documents retrieved* denotes that how many documents are indeed relevant to group's interest among all the retrieved documents, *recall* and *precision* are respectively computed by dividing the *relevant documents* in *test set* and *total documents retrieved* item by the *relevant documents retrieved* item. Fig. 9 illustrates the change of retrieval efficacy with the increment of usage duration. In the figure, Fig. 9(a) shows the change of recall and Fig. 9(b) shows the

Table 3
Document retrieval results

| | Keyword matching ($G_1$, $G_2$) | Profile-based matching ($G_1$) | | | | | |
|---|---|---|---|---|---|---|---|
| | | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ |
| Total documents retrieved | 121 | 121 | 92 | 79 | 70 | 64 | 61 |
| Relevant documents retrieved | 55 | 55 | 53 | 52 | 52 | 51 | 51 |
| Recall (%) | 91.7 | 91.7 | 88.3 | 86.7 | 86.7 | 85 | 85 |
| Precision (%) | 45.5 | 45.5 | 57.6 | 65.8 | 74.3 | 79.7 | 83.6 |

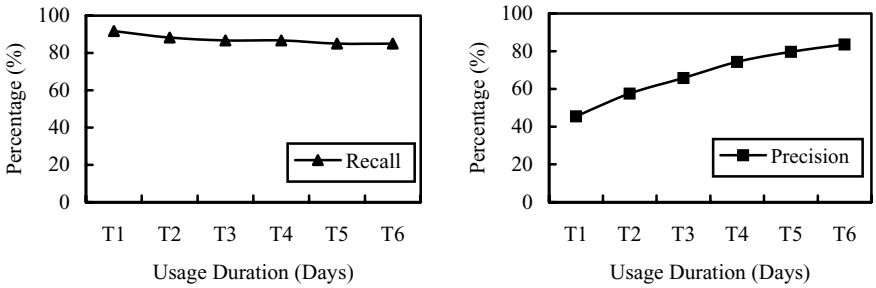



Fig. 9. Change of retrieval efficacy.

change of precision. However, the change speed of the curves may vary with the subjective effort of group members, such as the system usage frequency, the evaluation information, etc.

From this experiment, we can draw the following implications: (1) In the case of performing the document retrieval based on keyword matching, different groups are sure to get the same searching results if they input the identical keywords. (2) In the initial state of using our system to retrieve the documents, the profile is roughly described. Thus, the retrieval effect is as bad as that of the case with the keyword matching approach. (3) Assuming that the research interest of a group is constant in a period of time, there exists a trend that the longer a group uses our system, the more precise a group's interest is described by the profile, and the better retrieval effect is obtained (i.e. the retrieval precision increases obviously despite little decrease of retrieval recall). Nevertheless, the experiment is relatively short-term with a limited set of sample fields. Currently, we are trying to experiment on larger research groups during a long period to test the effectiveness of our proposed approach.

The applications in research group show that group members have different abilities to contribute knowledge and that the experienced members obviously contribute much more than the novices. The experienced members put excellent research documents in the Knowledge Grid with corresponding evaluation information in most cases, while the newly joined group members mostly browse the documents recommended by the other members but provide less evaluation feedbacks.

## 8. Conclusions

The proposed document logistics approach has the following three characteristics. First, it can continuously discover and collect new potentially relevant documents based on the semantic profiles. Second, it allows distributed group members to collaborate on organizing and evaluating shared documents with the support of Knowledge Grid. Third, it provides flexible management operations and recommendation services for group members to efficiently access relevant documents. We have implemented the prototype of document logistics based on the Knowledge Grid platform VEGA-KG (available at http://kg.ict.ac.cn). Experiments show that the approach can promote the effectiveness and efficiency of cooperative research to a certain extent.

Ongoing work includes the following four aspects: add time attenuation factor into the computation of association weight between concepts so as to reflect the change of research trends and the emerging new areas of science; make use of the semantic-link network to construct complex semantic profiles and realize service logistics based on the matching between the semantic profiles and services [22]; use the knowledge flow model to realize knowledge logistics [23]; and make use of new resource model to uniformly describe documents, services and profiles.

## References

[1] T. Berners-Lee, J. Hendler, O. Lassila, Semantic Web, Scientific American, May 17, 2001.

[2] E. Bloedorn, I. Mani, T.R. Macmilan, Representational issues in machine learning of user profiles, in: Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access, Stanford, CA, March 1996.
[3] K.D. Bollacker, S. Lawrence, C.L. Giles, A system for automatic personalized tracking of scientific literature on the Web, in: Proceedings of the Fourth ACM Conference on Digital Libraries, New York, 1999, pp. 105–113.
[4] P. Chan, A non-invasive learning approach to building Web user profiles, workshop on Web usage analysis and user profiling, in: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, 1999.
[5] H. Chen, et al., A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system, J. Am. Soc. Inform. Sci. 48 (1) (1997) 17–31.
[6] C.F. Daganzo, Logistic Systems Analysis, Springer, Berlin, 1999.
[7] C.M. Guelzo, Introduction to Logistics Management, Prentice-Hall, Englewood Chiffs, NJ, 1986.
[8] J. Hendler, Agents and the Semantic Web, IEEE Intell. Syst. 16 (2) (2001) 30–37.
[9] L. Houston, et al., Exploring the use of concept spaces to improve medical information retrieval, Decision Support Syst., Special Issue on Decision Support for Health Care in a New Information Age 30 (2) (2000) 171–186.
[10] M. Klein, XML, RDF, and relatives, IEEE Internet Comput. 5 (2) (2001) 26–28.
[11] S. Lawrence, K. Bollacker, C.L. Giles, Indexing and retrieval of scientific literature, in: Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM99), Kansas City, MO, November 1999, pp. 139–146.
[12] S. Lawrence, C.L. Giles, K. Bollacker, Digital libraries and autonomous citation indexing, IEEE Comput. 32 (6) (1999) 67–71.
[13] S.H. Lin, et al., Extracting classification knowledge of Internet documents: a semantics approach, in: Proceedings of the ACM SIGIR'98, 1998, pp. 241–249.
[14] L. Page, et al., The PageRank citation ranking: bringing order to the Web, Technical Report, Stanford, Santa Barbara, CA, January 1998.
[15] V. Roslin, et al., Concept-based searching and browsing: a geoscience experiment, J. Inform. Sci. 27 (4) (2001) 199–210.
[16] G. Salton, Automatic Text Processing, Addison-Wesley, Reading, MA, 1989.
[17] E. Selberg, O. Etzioni, Multi-service search and comparison using the MetaCrawler, in: Proceedings of the Fourth International World Wide Web Conference, Boston, MA, 1995.
[18] E. Selberg, O. Etzioni, The MetaCrawler architecture for resource aggregation on the Web, IEEE Expert 12 (1) (1997) 8–14.
[19] Y.Y. Yao, Measuring retrieval effectiveness based on user preference of documents, J. Am. Soc. Inform. Sci. 46 (2) (1995) 133–145.
[20] H. Zhuge, A Knowledge Grid model and platform for global knowledge sharing, Expert Syst. Appl. 22 (4) (2002) 313–320.
[21] H. Zhuge, Component-based workflow systems development, Decision Support Syst. 35 (4) (2003) 517–536.
[22] H. Zhuge, Active document framework ADF: concept and method, in: Proceedings of the Fifth Asia Pacific Web Conference, Xian, China, Lecture Notes in Computer Science (LNCS), vol. 2642, Springer, Berlin, April 2003, pp. 341–346.
[23] H. Zhuge, A knowledge flow model for peer-to-peer team knowledge sharing and management, Expert Syst. Appl. 23 (1) (2002) 23–30.

**Hai Zhuge** is a professor at the Institute of Computing Technology, Chinese Academy of Sciences. He is serving as the area editor of the Journal of Systems and Software and on the editorial boards of the Information and Management and the Future Generation Computer Systems. He leads the Key Lab of Intelligent Information Processing and the China Knowledge Grid Research Group (http://kg.ict.ac.cn). His current research interest is the model and theory on future interconnection environment. He is the author of two books and over 60 papers appeared mainly in leading international conferences and journals such as Communications of the ACM, IEEE Intelligent Systems, IEEE Computing in Science and Engineering, IEEE Transactions on Systems, Man, and Cybernetics, Information and Management, Decision Support Systems, Journal of Systems and Software, and Future Generation Computer Systems.

**Yanyan Li** is a research assistant and PhD student of the China Knowledge Grid Research Group (http://kg.ict.ac.cn). Her research interests include cooperative knowledge management, intelligent information processing and e-learning. She has published three papers in international conferences.