Ontology-based Specific and Exhaustive User Profiles for Constraint Information Fusion for Multi-Agents

Xiaohui Tao, Yuefeng Li, [†]Raymond Y. K. Lau, and Shlomo Geva
Faculty of Science and Technology, Queensland University of Technology, Australia
[†]Department of Information Systems, City University of Hong Kong, Hong Kong

Abstract

Intelligent agents are an advanced technology utilized in Web Intelligence. When searching information from a distributed Web environment, information is retrieved by multi-agents on the client site and fused on the broker site. The current information fusion techniques rely on cooperation of agents to provide statistics. Such techniques are computationally expensive and unrealistic in the real world. In this paper, we introduce a model that uses a world ontology constructed from the Dewey Decimal Classification to acquire user profiles. By search using specific and exhaustive user profiles, information fusion techniques no longer rely on the statistics provided by agents. The model has been successfully evaluated using the large INEX data set simulating the distributed Web environment.

1. Introduction

One of the major problem domains served by Intelligent Agent Technology is improving Web search performance in a distributed information environment. A distributed Web information gathering system uses a single interface to search information from entire searchable Web corpus available. Such a system simplifies administration and restricts search to the best part of corpus. A Web distributed information gathering system consists of a broker and a set of search agents. The broker gets queries from a user, and forwards the queries to agents. Agents search the accessible Web corpus and return the results to the broker. The broker then fuses the results and returns to users the re-ranked, better results. Figure 1 illustrates a classic distributed Web information gathering system, where the empty arrows are way passing over queries and the solid arrows are way returning the results. Because the broker and agents rely on each other, their collaboration heavily affect the performance of the Web information gathering system.

Information fusion performed by the broker can be categorized into two types: the cooperative and non-cooperative fusion [5]. Information fusion relies on the statistics of corpus searched by agents. Usually, the results returned from agents are not consistent because agents employ different ranking methods and may search in different corpus. Hence, when the broker fuses the results returned from agents, the statistics of the searched corpus are in need. The cooperative fusion methods rely on the collaboration of search agents to provide the explicit statistics, for example, the size of the searched corpus, accessed documents, etc. However, in the real world agents may not like to provide such information due to, say, commercial intelligence. Sometimes although agents do collaborate, the communication between the broker and agents costs expensively in computation. Aiming at solving this problem, the non-cooperative fusion methods have been developed to ascertain the statistic information [5, 12, 17]. However, the ascertained information is not as accurate as that directly provided by agents. Thus, information fusion techniques need to be improved for application to the real world.

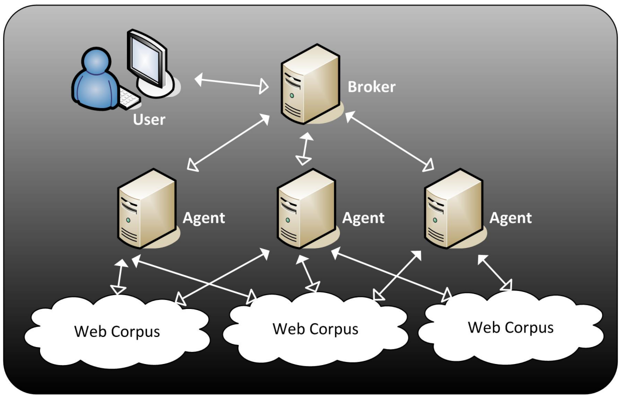

Figure 1. Distributed Web Information Gathering

Information gathering is moving from keyword-based towards concept-based. On this journey, ontologies play an important role [11]. Ontologies are a formal description and specification of concepts. They provide a well-defined and -constructed knowledge base to being shared by different systems. The brokers and agents in distributed Web information gathering systems, thus, can also share a same knowledge base for their common understanding of user information needs when performing a Web search task. This scenario raises a hypothesis: information fusion may not require the agent-provided statistics if the broker and agents can have an agreement for the searching concepts. Such an agreement can be enforced and constrained by applying an ontology to the distributed Web information gathering system.

Motivated by evaluating the above hypothesis, the work reported in this paper proposed an information fusion method using ontology-based user profiles. A subject ontology was first constructed based on the library system of Dewey Decimal Classification[1]. The semantic meaning of a users' information need was captured and extracted from the subject ontology. Two different versions (specific and exhaustive) of user profiles were acquired based on these extracted topic-relevant subjects. An improved information fusion method then used these specific and exhaustive user profiles to fuse results returned from agents without the agent-provided statistics. The proposed method was evaluated on the large INEX 2004 data set [6], and the evaluation result was promising.

The aim of this study was to promote the understanding of user profiles and to reduce the broker's heavy dependance on agents. With these aims, three contributions were made by the proposed model:

- A subject ontology constructed on the basis of the Dewey Decimal Classification (see Section 3). Abstracting and categorizing topics into a world knowledge taxonomy. Such an ontology provides a world knowledge base for share by different concept-based systems and applications;

- Exhaustive and specific user profiles that describe a user's information need at different concept levels (see Section 4). Theorems were proposed to constrain the use of the subject ontology for user profile acquisition. The specific and exhaustive user profiles provide a better understanding of user information needs, and will improve the design of personalized Web information gathering systems;

- An improved information fusion method independent from the agent-provided statistics (see Section 5). This method not only improves the performance of distributed Web information gathering systems, but also demonstrates the power of specific and exhaustive user profiles.

The work is reported in the following structure: Section 2 discusses the related work. Section 3 introduces the subject ontology, and Section 4 presents the ontology-based specific and exhaustive user profiles. The information fusion method using such user profiles is presented in Section 5. The related evaluation and discussions are presented in Section 6 and 7. Finally, Section 8 makes conclusions and addresses future work.

2. Related Work

Ontologies have been used by many groups to describe user background knowledge for Web information gathering. Li and Zhong [11] utilized pattern recognition techniques to discover knowledge from Web content for ontology construction. They also used association rules mining to capture semantic meanings from unstructured data for user information need interpretation. [24] introduced an approach to translate keyword queries to DL (Description Logics) conjunctive queries and used ontologies to describe user background knowledge. Sieg et al. [18] learned personalized ontologies for individual users in order to specify their preferences and interests in Web search. [8] developed an ontology using the Dewey Decimal Classification system for collection selection in distributed information gathering. These works utilized ontologies for user background knowledge discovery in order to improve Web information gathering performance.

User profiles are playing a more and more important role in Web information gathering and recommendation systems [22]. A user profile may be represented by a set of documents revealing the user's interest [21], a set of terms specifying the interesting topics of the user's information need [11], or a set of concepts or subjects referring to the user's interests and preferences [18–20,22]. Li and Zhong [11] categorized user profiles into two diagrams: the data diagram and information diagram. The data diagram profiles are usually acquired by analyzing a set of transactions, like [11, 18]. The information diagram profiles are acquired by manual techniques like questionnaires and interviews or by using information retrieval and machine-learning techniques like [18]. Interestingly, Some work like [4] and [23] used the collection of a user's desktop text documents, emails and cached Web pages, to explore user interests and acquire user profiles. Makris et al. [15] comprised user profiles by a ranked local set of categories and then utilized Web pages to personalize search results for users. These works attempted to discover user background knowledge for user profile acquisition.

Intelligent agent systems simplify administration and restrict searches to the best part of collections. An agent system consists of two parts, a search broker and a set of search agents. Information fusion is a task performed on the broker site, aiming to re-rank the results returned from agents in order to achieve better performance. To better do this, the statistics for the collections searched by the agents are important, because the searched collections are different. Based on the techniques of acquiring the collection statistics, information fusion methods can be categorized into two groups: cooperative fusion methods and non cooperative fusion methods [5].

The cooperative fusion (also called integrated information fusion) methods rely on the cooperation of search agents. One simple approach is that agents broadcast their collection statistics to public. This approach requires the statistic propagation protocols to constrain the agents for collaboration [5, 17]. Alternatively, another cooperative fusion mechanism fuses results by negotiation with search agents. The relevance of a document to a given query relies on the number of agents who recognize the document as relevant [12]. The cooperative fusion methods are computationally expensive, as relying on the cooperation from agents.

The non-cooperative information fusion methods do not rely on the agents' cooperation for collection statistics. Interleaving [17, 26] and uneven interleaving [3] methods fuse results in a round-robin fashion. Their effectiveness relies on the similar ranking methods used by agents. However, this is not realistic because agents hardly use the same ranking methods. Attempting to solve this problem, the rank position method [17] normalizes results before re-ranking them. Alternatively, the raw score merge, normalized raw score merge, and the weighted score methods attempt to fuse results by using the local document scores assigned by search agents [3]. However, the efficacy of these methods still relies on the similarity level of ranking methods utilized by agents [17]. In terms of collection statistics, MetaCrawler [25] and the shadow document method [28] consider the overlapping results more relevant. Proposed by [13] and [17], the semi-supervised learning method fuses results based on the estimation of collection statistics. Proposed by [5], the reference statistics method fuses results based on the statistics extracted from a reference collection rather than the searched collections. In summary, the current non-cooperative methods ascertain the collection statistics for fusion. Their performances are usually not as good as cooperative meth-

[1] http://www.oclc.org/dewey/.

ods because the estimated statistics are not adequately accurate.

3. Subject Ontology

3.1 Ontology Construction

The subject ontology is constructed based on the Dewey Decimal Classification (DDC) system, and is first introduced by [8]. The DDC system is one of the largest, most well-developed and widely used library classification systems. Around the world, the DDC system has been translated into more than 30 languages and serves library users in over 200,000 libraries in over 135 countries [27]. It is a human and intellectual endeavor covering all disciplines of human knowledge, and has been undergoing continuous revising and editing over more than a century represents a natural growth and distribution of human intellectual works. Wang and Lee [27] pointed out that the DDC system is ideal for knowledge engineering researches, because not only it is classified by professionals and quality guaranteed, but also it has a standard format allowing experiments under various controlled conditions. The subject ontology is thus constructed based on the DDC system. We encode the hierarchical structure of DDC system into the ontology structure and the subject headings in DDC system into concept classes in the ontology. The references connecting subject headings in the DDC system specify the semantic relationships held by any pair of classes in the ontology.

Before the explanation of the subject ontology utilization, we first give the formal definitions to the subject ontology and its related concepts.

Definition 1 *A subject $s \in \mathbb{S}$ is formalized as a 3-tuple $s := \langle code, \sigma, \hat{s}\rangle$, where*

- *code is the unique code assigned to s in the DDC system;*
- *σ is a signature mapping defining the direct neighbor subjects of s, and $\sigma(s) \subseteq \mathbb{S}$;*
- *$\hat{s}$ is a set of instances referred to by s, and each instance is a term t.* □

Definition 2 *A relation $r \in \mathbb{R}$ is a 2-tuple $r := \langle r_\tau, r_\nu\rangle$, where*

- *r_τ is a type of hierarchical relations, type $\in$ {is-a, part-of};*
- *$r_\nu \subseteq \mathbb{S} \times \mathbb{S}$. For each pair $(s_1, s_2) \in r_\nu$, s_2 is the subject who holds the r_τ relation to s_1, e.g. s_2 is-a s_1 or s_2 is part-of s_1.* □

Definition 3 *Let $\mathbb{KB}$ be a hierarchical knowledge base, which is formally defined as a 2-tuple $\mathbb{KB} := \langle \mathbb{S}, \mathbb{R}\rangle$, where*

- *$\mathbb{S}$ is a set $\mathbb{S} := \{s_1, s_2, \cdots, s_f\}$, in which each element is a subject;*
- *$\mathbb{R}$ is a set $\mathbb{R} := \{r_1, r_2, \cdots, r_g\}$, in which each element defines the relationship held by a pair of subjects in $\mathbb{S} \times \mathbb{S}$.* □

After defining the knowledge base, subjects and relations, we finally define the subject ontology:

Definition 4 *Let $\mathcal{O}$ be the subject ontology, which is a 4-tuple $\mathcal{O} := \langle \mathcal{S}, \mathcal{R}, \mathcal{H}(\mathcal{S}), \mathcal{T}, \rangle$, where*

- *$\mathcal{S}$ is a set of subjects and $\mathcal{S} \subset \mathbb{S}$;*
- *$\mathcal{R}$ is a set of relations and $\mathcal{R} \subset \mathbb{R}$;*
- *$\mathcal{H}(\mathcal{S})$ is the hierarchical structure of the ontology, and $\mathcal{H}(\mathcal{S}) \subseteq \mathcal{S} \times \mathcal{S}$;*
- *$\mathcal{T}$ is the set of instances referred to by all $s \in \mathcal{S}$ in the ontology.* □

In respect to a subject $s \in \mathcal{S}$, its child subjects $child(s)$ refer to the subjects in $\sigma(s)$ that are located at a more specific level in the $\mathcal{H}(\mathcal{S})$ than s, and its parent subject $parent(s)$ refers to a subject set with just a single element, which is in $\sigma(s)$ and located at a more abstractive level in the $\mathcal{H}(\mathcal{S})$ than s. The $desc(s)$ refers to the set of subjects in $\mathcal{S}$ that have direct and indirect links to s and are located at more specific levels in the $\mathcal{H}(\mathcal{S})$ than s. Finally, the $ance(s)$ refers to the set of subjects in $\mathcal{S}$ that link to s directly and indirectly and are located at more abstractive levels in the $\mathcal{H}(\mathcal{S})$ than s. Thus, we can have $child(s) \subseteq desc(s)$ and $parent(s) \subseteq ance(s)$.

For each subject $s \in \mathcal{S}$, to learn its $\hat{s}$ we first extract a training set from the catalogue of Queensland University of Technology (QUT) library[2]. A training document is generated by using the descriptive information about an item in the library catalogue. Such information includes the title, table of content, call number and summary. The call number is an unique DDC code assigned to the item by librarians. The expert knowledge underlying from the librarian assigned DDC codes and librarian summarized information, as pointed out by [27], are quality guaranteed. Hence, we attempt to mine these information for expert knowledge and use the knowledge to populate the subject ontology.

For a subject $s \in \mathcal{S}$, a set of training documents are retrieved from the library catalogue, based on the unique $code(s)$. The text pre-processing is performed on the training documents, including stopword removal, word stemming and term grouping. After that, A training document d is represented by $d = \{\langle t_1, w_{t_1}\rangle, \langle t_2, w_{t_2}\rangle, \ldots, \langle t_n, w_{t_n}\rangle\}$, where w_t is the weight of term t calculated using the $tf \cdot idf$ method. Treating a term as an instance, a subject $s \in \mathcal{S}$ refers to an instance set $\hat{s} = \{t | t \in \mathcal{T}\}$ and thus $\mathcal{T} = \bigcup_{s \in \mathcal{S}} \hat{s}$. Based on these, we can have an instance-subject matrix, which is formalized as follows:

Definition 5 *The instance-subject matrix $\mathcal{M}(\mathcal{O})$ is a 4-tuple $\mathcal{M}(\mathcal{O}) := \langle \mathcal{T}, \mathcal{S}, \mathcal{TS}, \eta\rangle$, where*

- *$\mathcal{T}$ is the entire instance set as defined in Definition 4;*
- *$\mathcal{TS}$ is a $m \times n$ zero-one matrix, where $n = |\mathcal{T}|$ and $m = |\mathcal{S}|$. $\mathcal{TS}(t_i, s_j) = 1$ means $t_i \in \hat{s_j}$; $\mathcal{TS}(t_i, s_j) = 0$ means $t_i \notin \hat{s_j}$;*
- *η is called reference, a mapping $(\eta : \mathcal{T} \rightarrow 2^{\mathcal{S}})$, that defines a set of subjects referring to the instance t:*

$$\eta(t) = \{s \in \mathcal{S} | \mathcal{TS}(t, s) = 1\}. \quad (1)$$

and its reverse is a set of instances referred to by a subject:

$$\eta^{-1}(s) = \{t \in \mathcal{T} | \mathcal{TS}(t, s) = 1\} = \hat{s}.\Box \quad (2)$$

Based on the matrix, given a subject, a set of instances can be extracted from $\mathcal{T}$; vice versa, given an instance, a set of subjects can also be extracted from $\mathcal{S}$.

The subject ontology is populated, adopting the agglomerative hierarchical clustering algorithm and the instance-subject matrix defined in Definition 5. Based on the hierarchical structure $\mathcal{H}(\mathcal{S})$ of $\mathcal{O}$, from leaf subjects ($\{s | child(s) = \emptyset\}$) we group the $\hat{s}$ of child subjects for their parent subject. Let $\hat{s}_{tr}$ be the instance set

[2] http://www.library.qut.edu.au.

of s extracted from the set of training documents of s. We have $\hat{s}$ defined:

$$\hat{s} = \hat{s}_{tr} \cup \bigcup_{s' \in child(s)} \hat{s'}. \tag{3}$$

As a result of ontology population, the leaf subjects are the most specific subjects that only refer to their own instance sets. The root subject in $\mathcal{H}(\mathcal{S})$ is the most general subject whose instance set is $\mathcal{T}$ covering all the instances in $\mathcal{O}$.

The agglomerative ontology population constrains the instance sets referred to by the subjects in the ontology. From Eq. (3), we can infer the following theorem:

Theorem 1 *Let $s_{(s_1 \vee s_2)} \in (ance(s_1) \cap ance(s_2))$, $s_{(s_1 \wedge s_2)} \in (desc(s_1) \cap desc(s_2))$, and $\{s_1, s_2\} \subseteq \mathcal{S}$ in $\mathcal{O}$, we have:*

1. $\hat{s}_{(s_1 \vee s_2)} \supseteq (\hat{s}_1 \cup \hat{s}_2)$;
2. $\hat{s}_{(s_1 \wedge s_2)} \subseteq (\hat{s}_1 \cap \hat{s}_2)$, *if* $\exists s_{(s_1 \wedge s_2)}$.

Proof 1

1. *From Eq. (3), we can have* $\hat{s}_{(s_1 \vee s_2)} \supseteq \bigcup_{s \in child(s_{(s_1 \vee s_2)})} \hat{s}$;
   based on the agglomerative clustering algorithm, we have:
   $\hat{s}_{(s_1 \vee s_2)} \supseteq \bigcup_{s \in desc(s_{(s_1 \vee s_2)})} \hat{s}$;
   $\because s_{(s_1 \vee s_2)} \in (ance(s_1) \cap ance(s_2))$;
   $s_1 \in desc(s_{(s_1 \vee s_2)})$ *and* $s_2 \in desc(s_{(s_1 \vee s_2)})$;
   $\therefore \hat{s}_{(s_1 \vee s_2)} \supseteq (\hat{s}_1 \cup \hat{s}_2)$.
2. *Assume* $\exists s_{(s_1 \wedge s_2)}$, *also from Eq. (3), we can have:*
   $\hat{s'} \subseteq \hat{s}$ *where* $s' \in child(s)$;
   based on the agglomerative clustering algorithm, we have:
   $\hat{s''} \subseteq \hat{s}$ *where* $s'' \in desc(s)$;
   $\because s_{(s_1 \wedge s_2)} \in (desc(s_1) \cap desc(s_2))$;
   $s_{(s_1 \wedge s_2)} \in desc(s_1)$ *and* $s_{(s_1 \wedge s_2)} \in desc(s_2)$;
   $\therefore \hat{s}_{(s_1 \vee s_2)} \subseteq \hat{s}_1$ *and* $\hat{s}_{(s_1 \vee s_2)} \subseteq \hat{s}_2$
   $\Rightarrow \hat{s}_{(s_1 \vee s_2)} \subseteq (\hat{s}_1 \cap \hat{s}_2)$. □

3.2 Semantic Study of Subjects

In this section, we introduce a multidimensional method, exhaustivity and specificity, to study the semantics of subjects in the ontology.

The exhaustivity of a subject refers to the extent of the semantic space dealt with by the subject. If the locality of a subject is toward an upper level in the $\mathcal{H}(\mathcal{S})$ of $\mathcal{O}$, the subject tends to be more general and has a greater exhaustivity value. Because $\mathcal{O}$ is populated using the agglomerative hierarchical clustering algorithm, the exhaustivity of a subject can be measured based on the matrix defined in Definition 5:

$$exhaustivity(s) = |\bigcup_{t \in \eta^{-1}(s)} \eta(t)|. \tag{4}$$

The specificity of a subject refers to the subject's focus on the referring-to semantic space. A subject located towards a lower level in the $\mathcal{H}(\mathcal{S})$ of $\mathcal{O}$ tends to be more specific, and thus has a higher specificity value. Specificity can also be measured based on the instances referred to by subjects:

$$specificity(s) = \frac{path(s)}{|\hat{s}|}. \tag{5}$$

where $path(s)$ refers to the shortest path travelling from the root to s in $\mathcal{H}(\mathcal{S})$.

Equation (4) and (5) scale the semantic extent and focus of a subject, and constrain the applicability of the subject for problem solving. A highly exhaustive subject holds a weak focus on its referring-to semantic space, whereas a highly specific subject refers to only a limited extent of semantic space.

Also based upon Equation (4) and (5), we can infer the following theorem:

Theorem 2 *Let $\{s_1, s_2\} \subseteq \mathcal{S}$ in the $\mathcal{H}(\mathcal{S})$ of $\mathcal{O}$, if $\exists (s_1, s_2) \in r_\nu$, then:*

1. $exhaustivity(s_1) \geq exhaustivity(s_2)$;
2. $specificity(s_1) \leq specificity(s_2)$.

Proof 2 *Assume* $\exists (s_1, s_2) \in r_\nu$, *we have* $s_2 \in child(s_1)$, *thus:*

1. *from Eq. (3), we have:* $\hat{s}_1 \supseteq \hat{s}_2 \Rightarrow \bigcup_{t \in \hat{s}_1} \eta(t) \supseteq \bigcup_{t \in \hat{s}_2} \eta(t)$;
   from Definition 5, we have: $\hat{s} = \eta^{-1}(s)$;
   $\Rightarrow \bigcup_{t \in \eta^{-1}(s_1)} \eta(t) \supseteq \bigcup_{t \in \eta^{-1}(s_2)} \eta(t)$;
   $\Rightarrow |\bigcup_{t \in \eta^{-1}(s_1)} \eta(t)| \geq |\bigcup_{t \in \eta^{-1}(s_2)} \eta(t)|$;
   $\because$ *from Eq. (4), we have:* $exhaustivity(s) = |\bigcup_{t \in \eta^{-1}(s)} \eta(t)|$;
   $\therefore exhaustivity(s_1) \geq exhaustivity(s_2)$.
2. *in the $\mathcal{H}(\mathcal{S})$ of $\mathcal{O}$ we have a path existing:* $s_2 \longrightarrow s_1 \longrightarrow \cdots \longrightarrow s_{root}$;
   based on the definition of $path(s)$, we have: $path(s_2) > path(s_1)$;
   from Eq. (3), we have $\hat{s}_2 \subseteq \hat{s}_1 \Rightarrow |\hat{s}_2| \leq |\hat{s}_1|$;
   $\because$ *from Eq. (5), we have:* $specificity(s) = \frac{path(s)}{|\hat{s}|}$;
   $\therefore specificity(s_1) \leq specificity(s_2)$. □

4. Ontology-based User Profiles

4.1 Acquiring User Profiles

In this presented work, a user profile is represented by subjects extracted from the subject ontology via the analysis of the user's given query. A query is first defined as:

Definition 6 *A query can be described as a 2-tuple $\mathcal{Q} := \langle termset, coverset \rangle$, where*

- *$termset$ is the set of terms in $\mathcal{Q}$ and $termset(\mathcal{Q}) = \{t_1, t_2, \ldots, t_n\}$;*
- *$coverset$ is the semantic space referred to by $\mathcal{Q}$ and $coverset(\mathcal{Q}) = \{s | s \in \eta(t), t \in termset(\mathcal{Q})\}$.* □

$coverset(\mathcal{Q})$ can be extracted from $\mathcal{S}$ using $termset(\mathcal{Q})$ based on the mappings defined in Definition 4. The extent of $coverset(\mathcal{Q})$ can be measured by $|coverset(\mathcal{Q})|$.

The subjects in $coverset(\mathcal{Q})$ require further observation for specificity and exhaustivity. As previously discussed in Section 3.2, subjects have varying focus and extent of the referring-to semantic spaces. Thus, their belief to the user profile is also varying.

To scale a subject's belief to the user profile, we first measure the extent of semantic space overlapped by the subjects and query. Sometimes, the semantic space referred to by a subject is within the shadow of the query's space. In such a case, the semantic meaning of the subject is important to the query and thus, the subject has strong belief to the profile. In some other cases, the semantic space referred to by a subject is overlapping with that of the query. In this case, the subject and the query have overlapping meanings and thus, the subject holds only partial belief to the user profile. This type of influence to the belief can be measured by the size of $\hat{s} \cap termset(\mathcal{Q})$.

Another influence to the belief held by a subject to the user profile is from the focus of the subject. Subjects with a great specificity have strong focus on their semantic meanings. Hence, the influence from a subject's semantic focus can be measured by the subject's specificity.

Also influencing the belief of a subject to the user profile is the semantic extent possibly shadowed by the profile. More subjects shadowed by the profile will result in weaker belief of each subject to the profile because everyone shares less. Such an influence can be measured by the inverse of $|coverset(\mathcal{Q})|$.

Based upon these discussion, the belief of a subject to the user profile can be measured by:

$$belief(s, \mathcal{Q}) = \frac{|\hat{s} \cap termset(\mathcal{Q})| \times spcificity(s)}{|coverset(\mathcal{Q})|}. \tag{6}$$

Corresponding to a query $\mathcal{Q}$, the user profile denoted by $\wp(\mathcal{Q})$ can then be acquired by:

$$\wp(\mathcal{Q}) = \{\langle s, belief(s, \mathcal{Q})\rangle | s \in coverset(\mathcal{Q})\}. \tag{7}$$

4.2 Specific and Exhaustive User Profiles

The subjects extracted for the user profile need to be pruned. The subjects in the subject ontology are populated using the hierarchical agglomerative clustering algorithm, as defined by Eq. (3). The instances referred to by a child subject are also referred to by the child's parent subject. When using $termset(\mathcal{Q})$ to extract the relevant subjects ($coverset(\mathcal{Q})$), if extracting a child subject, its parent subject would also be extracted as well. Eventually, all subjects on the path from the most specific relevant subject to the root of ontology would be extracted. This raises a problem that if using all the subjects in $coverset(\mathcal{Q})$ to represent the user profile, some semantic spaces referred to by the user profile are duplicately counted. This problem is solved in this section by utilizing the concepts of specificity and exhaustivity.

The semantic meanings referred to by the user profile can be clarified in two different versions using the concepts of specificity and exhaustivity: *specific user profile* and *exhaustive user profile*. Two algorithms are introduced to refine the acquired user profile for these multidimensional versions.

```
input : $coverset(\mathcal{Q})$: subjects in initial $\wp(\mathcal{Q})$
output: $coverset_{spe}(\mathcal{Q})$: subjects for specific user
        profile $\wp_{spe}(\mathcal{Q})$
$coverset_{spe}(\mathcal{Q}) = \emptyset$ //initialize the specific subject set;
foreach $s_i \in coverset(\mathcal{Q})$ do
    forall $s \in coverset(\mathcal{Q})$, where $s \neq s_i$ do
        if $\neg\exists(s_i, s) \in r_\nu$ then
            $coverset_{spe}(\mathcal{Q}) = coverset_{spe}(\mathcal{Q}) \cup \{s_i\}$;
    end
end
```

Algorithm 1: Specific User Profile Acquisition

The specific user profile emphasizes the semantic focus of relevant subjects in the profile. Thus, only the subjects with solid specificity remain in the user profile. Treating the relevant subjects extracted in Eq. (7) as a subgraph of $\mathcal{H}(\mathcal{S})$ in $\mathcal{O}$, we keep only the leaf subjects for the specific user profile because they have the greatest specificity values, as defined by Eq. (5). Algorithm 1 presents the specific user profile acquisition. Constrained by $coverset_{spe}(\mathcal{Q})$, we revise Eq. (7) and define the specific user profile $\wp_{spe}(\mathcal{Q})$ as:

$$\wp_{spe}(\mathcal{Q}) = \{\langle s, belief(s, \mathcal{Q})\rangle | s \in coverset_{spe}(\mathcal{Q})\}. \tag{8}$$

Systems utilizing specific user profiles are meant to have high precision performance because the chosen subjects are specific and focused.

```
input : $coverset(\mathcal{Q})$: subjects in initial $\wp(\mathcal{Q})$
        $coverset_{spe}(\mathcal{Q})$: subjects in $\wp_{spe}(\mathcal{Q})$
output: $coverset_{exh}(\mathcal{Q})$: subjects for exhaustive user
        profile $\wp_{exh}(\mathcal{Q})$
1  $coverset_{exh}(\mathcal{Q}) = \emptyset$//initilize the exhaustive subject
   set;
2  foreach $s \in coverset_{spe}(\mathcal{Q})$ do
3      forall $s' \in coverset(\mathcal{Q})$, where $s' \neq s$ do
4          if $\exists(s', s) \in r_\nu$ then
           $coverset_{exh}(\mathcal{Q}) = coverset_{exh}(\mathcal{Q}) \cup \{s'\}$;
5      end
6  end
7  foreach $s_i \in coverset_{exh}(\mathcal{Q})$ do
8      foreach $s_j \in coverset_{exh}(\mathcal{Q}), s_j \neq s_i$ do
9          if $s_j \in ance(s_i)$ then
           $coverset_{exh}(\mathcal{Q}) = coverset_{exh}(\mathcal{Q}) - \{s_i\}$;
10     end
11 end
```

Algorithm 2: Exhaustive User Profile Acquisition

The exhaustive user profile emphasizes the semantic extent of subjects in the profile. Only the subjects of solid exhaustivity remain in the user profile. Paying a price of minimum loss in semantic focus (specificity), we choose the direct parents of leaf subjects in $\wp(\mathcal{Q})$ for exhaustive user profile $\wp_{exh}(\mathcal{Q})$. Algorithm 2 presents the exhaustive user profile acquisition. Note that at Step 10 s_i is removed instead of s_j because some leaf subjects in $child(s_j)$ may not be in $child(s_i)$. Removing s_j may be risky in losing valuable semantic meanings. Constrained by $coverset_{exh}(\mathcal{Q})$, we revise Eq. (7) and define the exhaustive user profile $\wp_{exh}(\mathcal{Q})$ as:

$$\wp_{exh}(\mathcal{Q}) = \{\langle s, belief(s, \mathcal{Q})\rangle | s \in coverset_{exh}(\mathcal{Q})\}. \tag{9}$$

Systems utilizing exhaustive user profiles are meant to have high recall performance because the chosen subjects cover broad semantic extent.

Based on Algorithm 1 and 2, we have the following theorem:

Theorem 3 *Given a query $\mathcal{Q}$, let $coverset_{exh}(\mathcal{Q})$ be the subject set extracted for the exhaustive user profile $\wp_{exh}(\mathcal{Q})$, and $coverset_{spe}(\mathcal{Q})$ for the specific user profile $\wp_{spe}(\mathcal{Q})$, from Theorem (1), we always have:*

1. *each $s \in coverset_{exh}(\mathcal{Q})$ is a parent of at least one $s' \in coverset_{spe}(\mathcal{Q})$;*
2. *each $s \in coverset_{spe}(\mathcal{Q})$ is an element of $desc(s')$, where $s' \in coverset_{exh}(\mathcal{Q})$;*
3. *$\bigcup_{s \in coverset_{spe}(\mathcal{Q})} \hat{s} \subseteq \bigcup_{s' \in coverset_{exh}(\mathcal{Q})} \hat{s'}$.*

Proof 3

1. *from Step 3 to 7 in Algorithm 2, we have:*
   $\forall s' \in coverset_{exh}(\mathcal{Q})$, there always $\exists(s', s) \in r_\nu$, where $s \in coverset_{spe}(\mathcal{Q})$;

from Definition 2, we have: $(s', s) \in r_\nu \Rightarrow s \in child(s')$;
$\therefore \quad \forall s' \in coverset_{exh}(\mathcal{Q})$, *there always* $\exists s \in child(s')$, *where* $s \in coverset_{spe}(\mathcal{Q})$.

2. *based on Prove 3(1), after Step 7 in Algorithm 2 we have:*
   for $\forall s \in coverset_{spe}(\mathcal{Q})$, *we have* $s \in child(s')$, *where* $s' \in coverset_{exh}(\mathcal{Q})$;
   from Step 8 to 12 in Algorithm 2, we have:
   the kept subjects are the ancestor subjects of the removed subjects;
   $\therefore s' \in coverset_{exh}(\mathcal{Q}) \in ance(s)$, *where* $s \in coverset_{spe}(\mathcal{Q})$; *and*
   $s \in coverset_{spe}(\mathcal{Q}) \subseteq desc(s')$, *where* $s' \in coverset_{exh}(\mathcal{Q})$.

3. *based on Prove 3(2), we have: for* $\forall s \in coverset_{spe}(\mathcal{Q})$, *there always* $\exists s \in desc(s')$, *where* $s' \in coverset_{exh}(\mathcal{Q})$;
   from Eq. (3) and Theorem 1, we have: $\hat{s} \subseteq \hat{s'}$;
   $\therefore \bigcup_{s \in coverset_{spe}(\mathcal{Q})} \hat{s} \subseteq \bigcup_{s' \in coverset_{exh}(\mathcal{Q})} \hat{s'}$. □

5. Ontology-based Web Information Fusion

In this section, we apply the subject ontology and specific and exhaustive user profiles to distributed information gathering (DIG) systems.

Cooperative fusion is an information fusion method used in distributed Web information retrieval systems. It merges the results retrieved by agents based on their negotiations. The search broker ranks a document higher than others if the document is judged relevant by more agents. The related evaluations [10, 12] reported that the cooperative fusion method had successfully achieved remarkable performance.

However, this method has a drawback. Agents return only the list of relevant documents but not the accessed documents. Thus, two reasons may cause the absence of a document: the agent judged it non-relevant; the agent did not access it. For the latter case, we have no evidence of that the agent believes either non-relevance or relevance of the document. To clarify this, the cooperative fusion method requires agents to provide the statistics of their searching collections. Such a mechanism is computationally expensive and impractical. In the real world, search agents may not tend to provide such information because of commercial reasons. Consequently, this drawback weakens the contribution of the cooperative fusion method.

By using the ontology-based specific and exhaustive user profiles, we can solve the above problem. Based on Theorem 3, we have that the subjects in $\wp_{exh}(\mathcal{Q})$ are the parents of those in $\wp_{spe}(\mathcal{Q})$, and $\bigcup_{s \in coverset_{spe}(\mathcal{Q})} \hat{s} \subseteq \bigcup_{s \in coverset_{exh}(\mathcal{Q})} \hat{s}$. Thus, when retrieving using the query expanded with the terms in $\hat{s}$ (where $s \in coverset_{spe}(\mathcal{Q})$ or $s \in coverset_{exh}(\mathcal{Q})$), the results gathered by using the specific user profile are always the subset of that using the exhaustive user profile. We can then treat the specific results as the relevant documents judged by the agent, and the exhaustive results as the documents accessed by the agent. The agent collaboration problem previously discussed can then be solved by the following rules:

$$(d \in \mathcal{R}(\wp_{exh}(\mathcal{Q}))) \wedge (d \notin \mathcal{R}(\wp_{spe}(\mathcal{Q}))) \Rightarrow \neg Relevant(d, \mathcal{Q})$$
$$(d \notin \mathcal{R}(\wp_{exh}(\mathcal{Q}))) \wedge (d \notin \mathcal{R}(\wp_{spe}(\mathcal{Q}))) \Rightarrow \neg Accessed(d, \mathcal{Q});$$

where $\mathcal{R}$ denotes a result set; $Relevant(d, \mathcal{Q})$ denotes the judgement made by an agent that d is relevant to $\mathcal{Q}$, and $Accessed(d, \mathcal{Q})$ denotes that d was accessed when retrieved for $\mathcal{Q}$.

6. Empirical Experiments

The experiments were designed to evaluate the hypothesis that using the ontology-based user profiles can improve the effectiveness of distributed Web information gathering systems.

6.1 Experimental Models

The **ONTO** model is the implementation of the proposed ontology-based user profiles information fusion approach. A subject ontology was first built (see Section 3). For a given query, a specific and an exhaustive ontology-based user profiles were acquired (see Section 4). Two expanded queries, one with the terms generated from the subjects in the specific profile and one with those from the exhaustive profile, were used to perform search in a distributed information environment via multi-agents. The results returned by the agents were finally fused using the cooperative fusion method with the "accessed or non-relevant" problem solved using the specific and exhaustive user profiles (see Section 5).

The **Baseline** model includes four sub-models implemented from typical information fusion methods in distributed information retrieval:

Interleaving (IL) [26] This method fuses the results in a round-robin fashion, which takes one document in turn from each ranked list;

Normalized raw scored weight (NRSW) [3] The NRSW fuses the results based on the documents' normalized relevance scores assigned by agents;

Rank Position (RP) [17] This method fuses the results based on the documents' normalized index positions ranked by agents;

Shadow Document Method (SDM) [28] If a document appears on one result list but not the others, the SDM assigns a shadow value to the document, and then sums the values of each document to fuse the results.

Another experimental model, namely the **ONTO-Sens**, was implemented for sensitivity analysis of the proposed approach. Different from the ONTO model, this model used only the initial user profile (see Section 4.1) instead of specific and exhaustive versions, and employed the baseline fusion methods instead of the ontology-based cooperative fusion. Table 1 shows the differences of three experimental models. Thus, the experimental hypotheses for ONTO-Sens were that (i) if the ONTO-Sens model outperformed the baseline models, the ontology-based user profiles could be proven successful, because they were the only difference between the ONTO-Sens and the baseline models; (ii) if the ONTO outperformed the ONTO-Sens, the ontology-based cooperative fusion could be proven successful, because this was the only difference between the two models.

Methods	ONTO	ONTO-Sens	Baseline
Ontology-based user profile	Yes	Yes	No
Ontology-based cooperative fusion	Yes	No	No
Baseline fusion methods	No	Yes	Yes

Table 1. Experimental Models

The experiment design is illustrated in Fig. 2. Note that the subject ontology used by the ONTO and ONTO-Sens was constructed (see Section 3) but is not illustrated in Fig. 2.

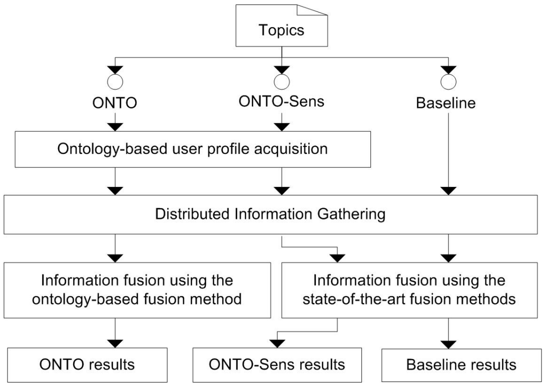
Topics
ONTO
ONTO-Sens
Baseline
Ontology-based user profile acquisition
Distributed Information Gathering
Information fusion using the ontology-based fusion method
Information fusion using the state-of-the-art fusion methods
ONTO results
ONTO-Sens results
Baseline results

Figure 2. Experiment Design

Collection	A	B	C	INEX
# of documents	5,984	3,994	2,389	12,107
# of terms	131,774	106,417	83,406	190,013
# of unique stems	99,828	79,478	61,590	147,167

Table 2. Statistics for the Data sets

6.2 Experimental Environment

The standard data set and queries provided by *Initiative for the Evaluation of XML Retrieval* (INEX) 2004 [3] were used in our experiments. The data set is a large volume of total 12,107 XML documents in a size of 494 megabytes, covering a great range of topics [16]. The data set was first pre-processed by stopword removal, word stemming, and term grouping.

For IR evaluations, INEX 2004 also provided a set of topics that were designed by the INEX linguists to simulate the behavior of general users. Each topic consisted of a title, a description, a narrative, and some keywords. In our experiments only the titles of topics were used, based on an assumption that in the real world users often use short queries. A total of 29 such queries were used in the experiments.

In order to simulate the distributed information environment, we randomly partitioned the INEX data set into three collections by using the following mechanism:

$$\begin{cases} A = \{d | id_d\%2 = 0\} \\ B = \{d | id_d\%3 = 0\} \\ C = \{d | id_d\%5 = 0\} \cup (Corpus - A - B) \end{cases} \qquad (10)$$

where id_d is the document identity assigned by the INEX and $\%$ denotes the modulus operator in math that returns the remainder. The documents in three collections were overlapping. Also because the document IDs have no connection to the contents, Eq. (10) ensured the random partition. Table 2 displays the statistics for the corpus and the partitioned collections.

Corresponding to these partitioned collections, three intelligent agents were also implemented, employing different ranking methods: Cosine Similarity [14], *bm25* [1], and the query-focused similarity measure [9], respectively. Figure 3 illustrates the implemented distributed information environment.

7. Results and Discussions

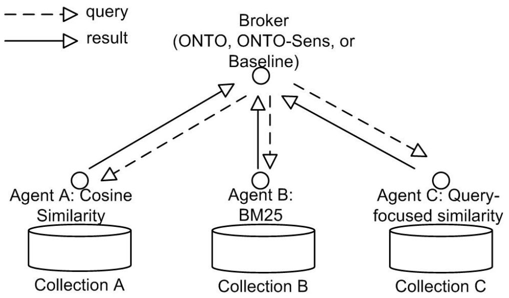
query
result
Broker
(ONTO, ONTO-Sens, or Baseline)
Agent A: Cosine Similarity
Agent B: BM25
Agent C: Query-focused similarity
Collection A
Collection B
Collection C

Figure 3. Experimental System

The performance of the experimental models was measured by precision and recall, the modern and standard methods in information gathering evaluations for effectiveness tests [2]. Precision is a measure of the ability of a system to retrieve only relevant documents, and recall is of the ability to retrieve all relevant documents. Only the $strict$ documents, those highly focused on and exhaustively covering the concepts requested by a query, were recognized as relevant documents to the query (see [16] for details). The experimental results are presented in Fig. 4.

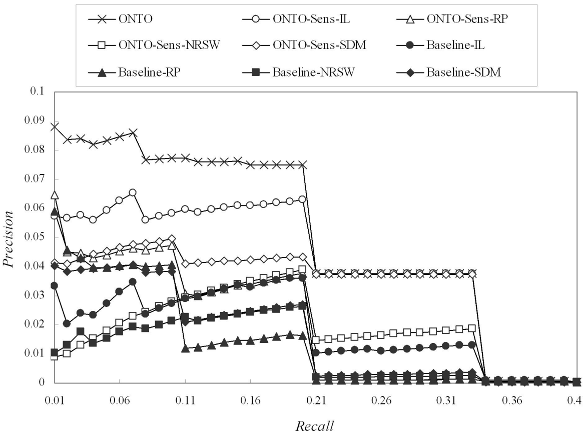

Figure 4. Experimental Results

The ONTO and ONTO-Sens models were designed to evaluate the ontology-based cooperative fusion method. The ONTO model used the specific and exhaustive profiles and the ontology-based cooperative fusion, whereas the ONTO-Sens model used only the initial user profiles and the baseline fusion methods. Thus, the experimental design here was to evaluate the ontology-based cooperative fusion using the specific and exhaustive user profiles.

As the results shown on Fig. 4, the performance achieved by the ONTO model was better than those of the ONTO-Sens model using Rank Position, NRSW and SDM methods. After the recall cutoff point 0.2, the ONTO-Sens using the Interleaving method matched the performance achieved by the ONTO. However, the ONTO model's performance was still much better than that of ONTO-Sens before reaching recall cutoff point 0.2. Based on the experimental results, the ONTO model using the ontology-based fusion method has achieved better performance than the ONTO-

[3] http://inex.is.informatik.uni-duisburg.de:2003/

Sens using state-of-the-art fusion methods.

The ONTO-Sens and Baseline models were designed to evaluate the ontology-based user profiles, because the profiles were their only difference. Demonstrated by the results, the performance achieved by the ONTO-Sens outperformed those achieved by all four fusion methods in the Baseline models. The ontology-based user profiles was proven successful by the experiments.

The ONTO-Sens model investigated the semantic concepts of given queries by using the subject ontology and acquired the ontology-based user profiles. In contrast, the Baseline models went into the information gathering tasks straight way and did not attempt to interpret the concepts from the queries. The out-performance of the ONTO-Sens model over the Baseline demonstrates that performance of IR systems can be improved by utilizing the subject ontology and ontology-based user profiles.

Note that Fig. 4 shows only the experimental results before reaching the recall cutoff point 0.41. As since that on, there is no significant difference between the experimental models. Considering that most of Web users are interested in only the topic few pages of results (as reported by [7]), the presented experimental results are still meaningful.

Also as shown on Fig. 4, the results produced by all experimental models have achieved only a limited level of precision-recall performance. After an investigation conducted on the insight of INEX data set, we found that the set was highly sparse, and only a very limited number of positive documents (only 40 per topic on average) available in the testing set in terms of $strict$ relevance. However, the proposed model and baselines were tested on the same data set and they suffered from the same limitation of sparseness. Thus, this limitation did not affect the stability of evaluation results.

8. Conclusions and Future Work

The aim of the work presented in this paper was to improve the performance of distributed Web information gathering systems by using ontologies. For this aim, an ontology-based model has been proposed and made three contributions: (i) a subject ontology constructed on the basis of the Dewey Decimal Classification system; (ii) a method to acquire a user's specific and exhaustive user profiles; and (ii) an improved information fusion method using the specific and exhaustive user profiles. Theorems have also been proposed to constrain the use of the subject ontology for acquisition of specific and exhaustive user profiles. Aiming to evaluate the proposed model, empirical experiments have been performed on a distributed information environment simulated using the INEX 2004 data set. The proposed model was compared with baselines employing state-of-the-art information fusion methods and using or not-using the subject ontology. The proposed model outperformed all the baselines in the experiments. The evaluation result was satisfactory and promising.

Considering the limitations (sparse data set) in the current evaluation, further experiments will take place in the near future to evaluate the proposed ontology-based model on more data sets. The presented work will also be extended to applying the subject ontology and exhaustive/specific user profiles to solving more problems in distributed Web information gathering.

Capturing user information needs is so hard and gathering accurate Web information is so challenging. By the way we are approaching towards the solutions to these difficulties and challenges, the contributions made by this presented work will largely extend and become increasingly significant.

Acknowledgements

The work presented in this paper was partially supported by Grant DP0988007 from the Australian Research Council.

9. References

[1] B. Billerbeck, *et. al.* RMIT University at TREC 2004. In *TREC 2004*, 2005.

[2] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. of SIGIR'00*, pages 33–40, 2000.

[3] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proc. of SIGIR'95*, pages 21–28, 1995.

[4] P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the Web. In *Proc. of SIGIR'07*, pages 7–14, 2007.

[5] N. E. Craswell. *Methods for Distributed Information Retrieval*. PhD thesis, The Australian National University, 2000.

[6] N. Fuhr, S. Malik, and M. Lalmas. Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003. In *Proc. of the 2nd INEX Workshop*, pages 1–11, 2004.

[7] B. J. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263, 2006.

[8] J. D. King, Y. Li, X. Tao, and R. Nayak. Mining World Knowledge for Analysis of Search Engine Content. *Web Intelligence and Agent Systems*, 5(3):233–253, 2007.

[9] K. L. Kwok. Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information System*, 8(4):363–386, 1990.

[10] Y. Li, C. Zhang, and S. Zhang. Cooperative Strategy for Web Data Mining and Clearning. *Applied Artificial Intelligence*, 17(17):443–460, 2003.

[11] Y. Li and N. Zhong. Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.

[12] Y. Li. Information fusion for intelligent agent-based information gathering. In *Proc. of WI'01*, pages 433–437, 2001.

[13] J. Lu and J. Callan. Merging retrieval results in hierarchical peer-to-peer networks. In *Proc. of SIGIR'04*, pages 472–473, 2004.

[14] A. D. Maedche. *Ontology Learning for the Semantic Web*. Kluwer Academic Publisher, 2002.

[15] C. Makris, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis. Category ranking for personalized search. *Data & Knowledge Engineering*, 60(1):109–125, January 2007.

[16] S. Malik, M. Lalmas, and N. Fuhr. Overview of INEX 2004. *Lecture Notes in Computer Science*, 3493:1–15, 2005.

[17] L. Si and J. Callan. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems*, 21(4):457–491, 2003.

[18] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proc. of CIKM'07*, pages 525–534, 2007.

[19] X. Tao, Y. Li, N. Zhong, and R. Nayak. Ontology Mining for Personalized Web Information Gathering. In *Proc. of WI'07*, pages 351–358, 2007.

[20] X. Tao, Y. Li, N. Zhong, and R. Nayak. An Ontology-based Framework for Knowledge Retrieval. In *Proc. of WI'08*, pages 510–517, 2008.

[21] X. Tao, and Y. Li. A User Profiles Acquiring Approach Using Pseudo-Relevance Feedback. In *Proc. of RSKT'09*, pages 658–665, 2009.

[22] X. Tao, Y. Li, and N. Zhong. A personalized ontology model for web information gathering. *Accepted by IEEE Transaction on Knowledge and Data Engineering*, December 2009.

[23] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proc. of SIGIR'05*, pages 449–456, 2005.

[24] T. Tran, P. Cimiano, S. Rudolph, and R. Studer. Ontology-based interpretation of keywords for semantic search. In *Proc. of ISWC'07*, pages 523–536, 2007.

[25] T. Tsikrika and M. Lalmas. Merging techniques for performing data fusion on the Web. In *Proc. of CIKM'01*, pages 127–134, 2001.

[26] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. The collection fusion problem. In *The Third Text REtrieval Conference (TREC-3)*, pages 1–10, 1994.

[27] J. Wang and M. C. Lee. Reconstructing DDC for interactive classification. In *Proc. of CIKM'07*, pages 137–146, 2007.

[28] S. Wu and F. Crestani. Shadow document methods of resutls merging. In *Proc. of SAC'04*, pages 1067–1072, 2004.