INFORMATION FUSION WITH SUBJECT-BASED INFORMATION GATHERING METHOD FOR INTELLIGENT MULTI-AGENT MODELS

Xiaohui Tao, John D. King, Yuefeng Li
x.tao@student.qut.edu.au, {j5.king, y2.li}@qut.edu.au

Abstract

This paper addresses the problem of information fusion using a multi-agent information gathering system. We present a hierarchical subject-based query expansion method, followed by a cooperative fusion algorithm for unstructured documents. We evaluate the performance using the traditional methods of precision and recall. The results show that the subject-based fusion method is promising and efficient.

1. Introduction

The astonishing growth of the internet has provided a vast amount of electronic information distributed across almost every country in the world. Many people think that large search engines like Google[1] cover most of the information available on the internet. However, a comparison of deep web collections such as PubMed[2] or IEEE Explore[3] with a popular search engine reveals that most of the database contents are not in the search engine index. Much of this information is only available through the *deep web*. The *deep web* is a massive collection of electronic data accessible only through a search interface. Examples of deep web content are scientific journals, patent collections, news articles, and medical information. There are many deep web collections containing a rich source of information that is not available to traditional search engines. Why is this rich and authoritative information not available to search engines? The simple answer is that there are no hyperlinks contained in these databases. The only practical way to access the information contained in these databases is to enter a query into a search form and then to browse the results. However, there are hundreds of thousands of these databases available, making it difficult for a user to find the right collection to use. It is also difficult to gather relevant information from the deep web because the information is vast, heterogenous, changing, and distributed.

It is well known that most search engine users only view the first page of search results. Retrieving and ranking information so that the user gets the best possible results within the first page of results is the subject of much research. Another problem is that users often do not know an exactly what they are searching for, and search is often made more difficult if the user does not know the exact terms to use in a query. The *Query Expansion* (QE) mechanism has been researched for decades to address this challenge. QE automatically expands the query by adding equivalent terms that are related to the words supplied by the user in order to retrieve more precise results [20]. We test the

[1] http://www.google.com

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[3] http://ieeexplore.ieee.org/Xplore/login.jsp?url=/search/advsearch.jsp

hypothesis that reformulating the search query using subjects instead of terms may improve the results of a query. Compared to traditional QE methods [4,5,17,20] subject based query expansion may have better performance since subjects cover broader conceptual areas than terms.

The notion of *Information Gathering* (IG) is also proposed to address the challenge of deep Web information retrieval [9]. The goal of IG is to obtain only the information that users need (no more and no less). There are two related research fields for IG: Information Retrieval (IR) and agent-based information systems. Information Retrieval works best with a query that covers an exact subject area. However many information retrieval queries are ambiguous and may be mapped to many different subject. Moreover, information is becoming highly distributed in an environment where collections and services are constantly changing. This has lead to researchers applying intelligent agent-based techniques to design new generation information systems.

Typically, a search broker is used for deep web information retrieval. A query is entered by the user, and the search broker selects a candidate set of collections to best answer the query. After the set of collections has been selected, the key problem for the broker is to fuse the resulting information together. This problem in database systems community is described as *Information Fusion* [7,21,22]. The problem of Information Fusion has attracted significant attention in the artificial intelligence and database systems communities [13]. In this paper we discuss the problem of fusion in multi-agent environments with subject-based information retrieval. We present a hierarchical subject-based information retrieval method combined with cooperative fusion algorithm for unstructured text documents to extend the functionality of "structured data (solutions) synthesis".

Our system uses a higher level approach than other information fusion systems. The benefits are that terms do not have to occur in both the query and the document for a match to be made. This "off-the-page" ranking method [1] has some highly desirable qualities such as reducing the effect of polysemy and term ambiguity. Previous research has used methods such as Latent Semantic Analysis for these benefits, however these systems suffer from the problem of limited scalability due to the computation resources to analyse large matrices. Another benefit is that our system does not require each collection to return local document weights which means that it can be used in the deep web. As such, we do not have to concern ourselves with the problem of combining and normalising local and global weights.

2. Agent-Based Information Gathering (IG)

The goal of fusion in Distributed Information Retrieval is to effectively combine the retrieval results from multiple, independent collections into a single result. Its effectiveness should approximate the searching considering multiple collections as a centralized single collection [21]. This implies that the search broker knows all collections, assumes collections are independent, and maintains a standard communication protocol between collections and the search broker. For deep web-based information system that aims at covering all uncertain and constantly changing information sources this assumption is not realistic. For this reason, we now consider the case of multi-agent environments, in where each agent knows some collections (which may overlap), and the search broker uses the techniques of agent cooperation to solve problems. We call this kind of system the Agent-Based Information Gathering (IG) system.

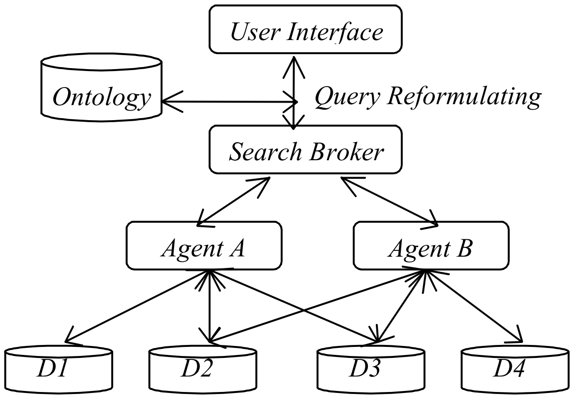

Figure 1: Agent-Based Information Gathering System

The following process describes a strategy for agent-based IG, which extends the descriptions of [14, 15]:

```
(1) Reformulate the user information need using ontology
(2) Send the reformulated query to each agent $\in \Theta$;
(3) While not ($t$) // $t$ is the time point of fusion
        Get new retrieved documents;
(4) /* Here we use $\Theta_t$ to represent the set of cooperated agents
     * which have sent back the retrieved documents at time $t$.*/
    Fuse the retrieved documents sent by $\Theta_t$;
(5) Assign the relevant documents to the user.
```

In the above process, there may be a waiting time for the search broker to get the retrieved documents. This period will be decided at the time of fusion, and is usually specified by the user of the system. During this period, the cooperating agents would typically send back the retrieved documents to answer the user information need. Failure would occur if a particular agent was unavailable (e.g. network failure), or where results could not be obtained. At the time of fusion, the search broker will evaluate the retrieved documents by using its knowledge about the cooperating agents and subject matter of the query and documents. Lastly, it will return the relevant documents to the user.

3. Subject-Based Query Expansion

As in any IR systems, how to expand the given query that can improve information retrieval performance is also a challenge in Agent-based IG system. Users do not always give a query consisting of easily categorical terms [18]. Often a query will contain ambiguous terms. We observe that terms that are in common use are often multi-purpose and used in many different senses. One example is the term "*system*", which is frequently used in computing, education, engineering, and science among others. Instead of expanding the query by traditional methods such as user logs, user feedbacks, or synonymy [4,5,17,20], we reformulate query by subjects. Subjects are extracted from ontologies such as WordNet [3,20], by using the terms contained in the given query. Subjects are identified as the highest hierarchical node in ontology covering connected topographic area that contains all, and only the family members of the given term. We hypothesise the joint set of extracted subjects will suggest the accurate subject, which is the one that best describes users need.

For example, *base, bat, glove,* and *hit* may point to multiple subjects respectively. However joint them together, they will lead to the subject of baseball game [19].

3.1. Subject Connections

Subjects are extracted by terms contained in the given query. The relation between subject and term is *m*:*n*, which means one term may point to multiple subjects, and one subject may cover multiple terms in the given query. Given a query $q{:=}\{t_1, t_2, t_3\}$, the query may be expanded as in Figure 2. Terms t_1, t_2, t_3 point to subject S_1 (solid lines). Since all terms under subject S_1 have been contained in the given query, there is no need to expand the query with S_1. Terms t_2, t_3 point to subject S_2, which expands the query with t_4 contained in S_2 (dash lines). Apart from S_1 and S_2, term t_3 also points to subject S_3, which results the query being expanded with t_5 and t_6 contained in subject S_3. Based on the given query, we can then have a set of subjects $S := \{S_1, S_2, S_3\}$ and expended terms as $S_1{:=}\{t_1, t_2, t_3\}$, $S_2{:=}\{t_2, t_3, t_4\}$, $S_3{:=}\{t_3, t_5, t_6\}$.

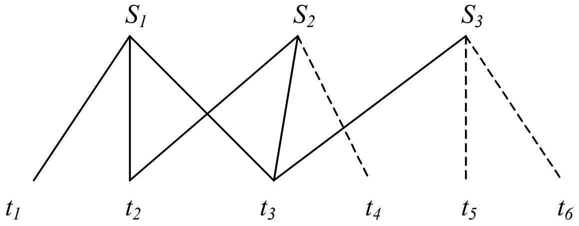

Figure 2: Query expansion using multiple subjects

Considering that the terms in the given query are those the users consider most important, we treat these terms as having higher priority than others. The connections of a subject are defined as the number of terms that point to it. For each connection pointed from original terms to the subject (solid lines), we count it as 1 in order to scale the subjects. The weights of $SC_{(S_k,\ q)}$ can then be normalized as

$$SC_{(S_k,\ q)} = \frac{1}{\sum_{j=1,\ \dots,\ n} \#Connections_{(S_j,\ q)}} * \#Connections_{(S_k,\ q)} \qquad (1)$$

$SC_{(S_k,\ q)}$ is the assigned weight of subject S_k respect to the query q. $\#Connections(S_k, q)$ is the number of connections the subject S_k having with query q. In respect to the given query q in Figure 2, the weights of subject connections can then be calculated as $SC_{(S_1,\ q)}$=0.5, $SC_{(S_2,\ q)}$=0.33, $SC_{(S_3,\ q)}$=0.167 where higher value indicates higher priority.

3.1. Subject Distance

In order to weight subjects it also deals with the distance from terms to the subject node in ontology. The closer the subject node is with term nodes, the more directly the terms are relevant to the subject. Figure 3 shows the scenario in topography.

In Figure 3, the query $\{t_1, t_2\}$ travels three levels to reach subject node *S*, the distance is counted as 3. The query $\{t_3, t_4\}$ for the same subject *S* travels two levels, and the distance is then counted as 2. In this case, we conclude that subject *S* is more relevant to $\{t_3, t_4\}$ than $\{t_1, t_2\}$, since $\{t_3, t_4\}$ is closer to it.

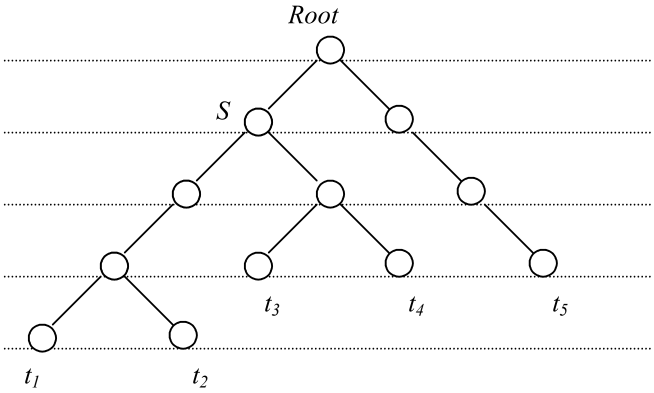

Figure 3: Subject Distance in Ontology

Another situation is terms being on different levels, eg $\{t_1, t_3\}$. In this case, we consider the longest path as the distance. For $\{t_1, t_3\}$ the distance to subject S is counted as 3. However, in case of term t_5, it has to travel through the root to reach subject S. Then its distance to subject S will be considered as infinite, and the relevance of subject S to t_5 can be ignored [11].

The similarity of a query to a subject will be normalized by using 1 divided by the distance value, as formula (2). $SS_{(S_k,\ q)}$ indicates the similarity of subject S_k to query q. The $distance_{(q,\ S_k)}$ indicates the distance of query q travelling to reach subject S_k. For the examples in Figure 3, the similarity of subject S to query $\{t_1, t_2\}$ is counted as 0.33, and to query $\{t_3, t_4\}$ is counted as 0.5. Higher similarity value means more relevant the subject to the query.

$$SS_{(S_k,\ q)} = \frac{1}{distance_{(q,\ S_k)}} \qquad (2)$$

Based on the Connection and Similarity values of subject to query, the final weight of a subject can be calculated as follows, in which SW_{S_k} is the final score we assign for the weight of subjct S_k respect to query q:

$$SW_{(S_k,\ q)} = SC_{(S_k,\ q)} * SS_{(S_k,\ q)} \qquad (3)$$

4. Cooperative Fusion Algorithm

Another difficulty arises when collections may be accessible by more than one agent [6]. For example, as scenario in Figure 1 we have the case that collections $D1$, $D2$, and $D3$ are served by agent A, and collections $D2$, $D3$, and $D4$ are served by agent B. The return documents retrieved by agent A could belong to $D1$, $D2$, and $D3$; and the return documents retrieved by B belong to $D2$, $D3$, and $D4$. The problem for the search broker is to evaluate all of the retrieved documents which may contain duplicate information.

A clue for solving this problem is given by Fuhr in [7]. The search broker can ask agent A to revise its retrieved documents in case $D2$ as well as $D3$ are ignored. After that the problem will become the case where different agents access disjoint sets of collections. To use this approach, we give a negotiation method to select agents which are willing to revise their retrieved documents. One problem is that this approach cannot use multiple opinions for the same question (one kind of cooperation in multi-agent environments).

Considering many agents (rather than 3) which may have different efficiencies in information retrieval, the problem looks complex and the evaluation for this problem is not easy. However the following algorithm provides a cooperative approach for this problem:

```
(1) for $i$=1 to $n$          // initial the relevant array
        $R(d_i) = 0$;
(2) for each $\theta \in \Theta_t$     // support evaluation
        for each $d \in \Gamma_t(\theta)$
            $R(d) = R(d) + Precision_\theta(| \Gamma_t(\theta) |)$;
(3) for $j$=1 to $m$          //potential support evaluation
        for each $\theta \in \Theta_t$
            if ($C^t_j$ is not served by $\theta$)
                for each $d \in D_{t_j}$
                    $R(d) = R(d) + Precision_\theta(| D_t |)$;
(4) Resort $D_t$ based on multi-keys.
```

In this algorithm, the inputs are the set of cooperating agents at time t (Θ_t), the set of retrieved documents at time t ($D_t = \cup_{i=1,\ldots,m} D_{t_i}$, | D_t | = n), and the associated collections C^t_1 , C^t_2 , . . . , C^t_m , where, $D^t_i \in C^t_i$. We expect the outputs are the relevant degree function R, and the sequence of the retrieved documents.

The multi-key in step 4 for a document d consists of two keys. The first key is $R(d)$, the second key is $\sum_{\theta \in \Theta t} Weight_\theta(d)$. Treating each subject in the query as a sub-query, IR agent θ provides the cosine similarity of the document to the particular sub-query. The $weight_\theta(d)$ is then calculated as the formula presented below, where d is a retrieved document, m is the number of subjects, $SW_{(S_k, q)}$ is the weight of that particular subject respect to the query, $Sim_{(d_n, s_k)}$ is similarity value of the document regarding to the subject.

$$Weight_\theta(d) \;=\; \sum_{i=1,\ldots,m} (Sim_{(d_n, s_k)} * SW_{(S_k, q)}) \qquad (4)$$

If the document is not relevant, then the weight of the document from that agent will be counted as zero. The final weight of the document is then calculated by the sum of weights provided by each agent. In the case of agent θ provides a set of retrieved documents $\{d_1, \ldots, d_{|\Gamma_t(\theta)|}\}$ with the corresponding weights $\{W_1, \ldots, W_{|\Gamma_t(\theta)|}\}$, the retrieved documents will be divided into some classes firstly based on the first key, in which every document has the same R, and then are sorted by the second key.

5. Performance of Subject-Based IG Method

For testing we use the *ACM Portal*, *IEEE*, and *ScienceDirect* deep web collections for information retrieval with a set of queries. The search broker (BROKER) first extracts subjects of original query from WordNet. For each subject, a sub-query is formulated by the terms covered by that subject. The original query then ends with an expanded set of sub-queries.

The search broker agent (BROKER) sends the set of sub-queries to three agents: *ACM Portal*, *IEEE*, and *ScienceDirect*. At fusion time, each agent provides a set of retrieved documents, in which each document has a collection name, document name and weight (similar value).

The search broker, however, does not have any idea about which agents are better. So it assumes that every agent has the same trustworthiness (same precision curve). By using the cooperative fusion algorithm the broker agent divides the retrieved documents into 3 classes, the first class includes documents with the support from every agent (3 agents); the second class has documents with the support from 2 agents, the third class has documents with the support from one agent. By re-sorting the retrieved documents with weights of each subject (sub-query), the search broker can select the top part documents to answer its users based on how many documents users want. Figure 4 shows the comparative precision and recall of the original and expanded queries.

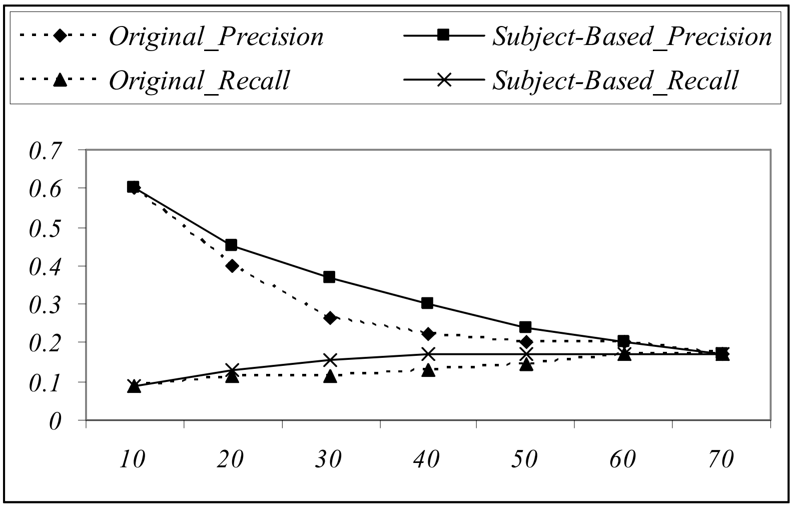

Figure 4: Precision/Recall Performance

By comparing with search agents of *ACM Portal*, *IEEE*, and *ScienceDirect* using original query set, the BROKER can obtain higher precision and recall with our method and algorithm. In summary, this experiment shows that the retrieval and fusion result is better than any single agent with traditional query expansion mechanism. The effectiveness of the combination in this subject-based Distributed Information Retrieval approach is enhanced rather than only approximated to the effectiveness of searching the entire set of documents as a single collection with traditional QE methods by any agent (the goal of collection fusion in Distributed IR).

6. Conclusion

The importance of subject knowledge with query expansion and text retrieval was presented by Vakkari in [18]. He points out that knowing more about a subject can generate better query expansion. Voorhees [19,20] proposes a method of using synonyms and hyponyms to expand query and disambiguate terms in query, which results in the QE performing better. Joho et. al. [10] present a hierarchical approach to query expansion, and they conclude that using a hierarchical approach to query expansion results in a significant shorter time to complete the retrieval task.

The problem was first identified with respect to collection fusion in [21]. This approach approximates the relevant document distribution over the collections by learning from the results of past queries. Another approach [2] uses an inference network to rank not only documents at the provider sites. Much of collection fusion literature focuses on using local and global weighting and the difficulties of normalising the weights with respect to each other. Various solutions have been presented for this, however in the context of the deep web few collections return local weights, making these solutions suitable only in cases where these is a standard communication protocol between the search broker and the collections. The collection fusion problem can also be solved by either using globally valid document frequencies (if the same cosine-based ranking method is used at each site) or by re-ranking selected (retrieved) documents at the broker site [16]. However, past research found that the performance of QE critically depends on user's subject knowledge [18].

Metacrawler [6] and Savvy Search [8] are agents that operate at a higher abstraction level by utilizing existing search engines. BIG [12] performs information fusion by retrieving documents, extracting attributes, converting unstructured text to structured data, and then integrating the data. In this paper we talk about unstructured text rather than structured information (data).

In summary, we have presented a subject-based information fusion method applying cooperative fusion algorithm. Our system extends the capability of the traditional query expansion and collection fusion methods in Distributed IR. By utilising agent cooperation, this algorithm provides a novel approach for the problem of information fusion on the deep web.

For further research we plan a full evaluation of our system against other fusion systems and a better use of the subject hierarchy using world knowledge to leverage our understating of term usage across various subjects.

7. References

[1] Brin, S., & Page, P. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[2] Callan, J. P., Lu, Z., & Croft, W. B., (1995). Searching distributed collections with inference networks, *in Proceedings of SIGIR'95*. (p.21-29).

[3] Cognitive Science Laboratory, Princeton University. (2005). *WordNet - Princeton University Cognitive Science Laboratory*. Retrieved 02 May, 2005, from http://wordnet.princeton.edu/

[4] Cui, H., et al., (2002). *Probabilistic query expansion using query logs* in *Proceedings of the 11th international conference on World Wide Web* 2002 ACM Press: Honolulu, Hawaii, USA p. 325-332

[5] Cui, H., et al., (2003). Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering, 2003. 15*(4), 829-839.

[6] Etzioni, O., (1995). Results from using the metacrawler, *in Proceedings of 4th WWW Conference*, F. Varela & P. Bourgine (Eds.). MIT Press: Cambridge, MA.

[7] Fuhr. N., (1999). A decision-theoretic approach to database selection in networked IR, *ACM Transactions on Information Systems*, 1999, 17(3): 229-249.

[8] Howe, A. E., & Dreilinger, D., (1997). Savvy Search : a metasearch engine that learns which search engines to query; *AI Magazine*,1997, 18(2): 19-25.

[9] Jennings, H. R., Sycara, K., & Wooldridge, M., (1998). A Roadmap of agent research and development, *Autonomous Agents and Multi-Agent Systems*, 1998, 1(1): 7-38.

[10] Joho, H., et al., (2002). Hierarchical presentation of expansion term*s*. In *Proceedings of the 2002 ACM symposium on Applied computing* 2002. (p. 645-649). ACM Press: Madrid, Spain.

[11] Khan, L., McLeod, D., & Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal, 13*(1), 71-85.

[12] Lesser, V., et al.,(2000). BIG: an agent for resource-bounded information gathering and decision making, *ArtificialIntel ligence*, 2000, 118: 197-244.

[13] Levy, A. Y., & Weld, D. S. (2000). Intelligent Internet systems, *ArtificialIntel ligence*, 2000, 118: 1-14.

[14] Li, Y. (2001). Information Fusion for Intelligent Agent-Based informaiton Gathering. *WI 2001, LNAI 2198 - Conference proceedings*. (pp. 433-437). Springer-Verlag: Berlin Heidelberg.

[15] Li, Y., Zhang, C., & Zhang, S. (2003). *Cooperative Strategy for Web Data Mining and Cleaning*. Applied Artificial Intelligence, 17(433-460). Taylor & Francis.

[16] Meng, W., et. al., (1998). Determining text databases to search in the Internet, *in Proceedings of 24th VLDB Conference*, 1998. Extended version.

[17] Peat, H. J., & Peter, W. (1999). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science, 42*(5), 378-383.

[18] Vakkari, P. (2002). Subject Knowledge, Source of Terms, and Term Selection in Query Expansion: An Analytical Study. *The 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval - Conference Proceedings.* (pp. 110-123). Springer-Verlag: London, UK

[19] Voorhees, E.M., (1993). Using WordNet to disambiguate word senses for text retrieval *in Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval-Conference proceedings. (171-180)*ACM Press: Pittsburgh, Pennsylvania, United States.

[20] Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. *The 17th annual international ACM SIGIR conference on Research and development in information retrieval in Dublin, Ireland - Conference Proceedings.* (pp. 61-69). New York: Springer-Verlag.

[21] Voorhees, E. M., Gupta, N. K., & Johnson-Laird, B., (1995). The collection fusion problem, *in Proceedings of TREC-3*, (p.95-104).

[22] Wu, S., & Crestani, F.(2004). Shadow Document Methods of Results Merging. In *SAC 04-Conference proceedings*.ACM:Nicosia, Cyprus.