Source Selection in a Distributed Search System

V.V. KLUEV
The Core and Information Technology Center
The University of Aizu
Tsuruga Ikki-machi Aizu-Wakamatsu City
Fukushima 965-8580
JAPAN

Abstract: This paper offers a new method of source selection in a distributed search environment. An integrated collection description on the base of most important single terms and word collocations makes possible efficient selection of a potentially useful collection for any given user query. Experimental results indicate good retrieval accuracy in the case of search for scientific documents.

Key-Words: - Search Engine, Distributed System, Metasearch, Query propagation

1 Introduction

Presently the Internet has hundreds of millions of sites, and its growth is exponential. Because there is not any control over the content of the net, potentially any piece of information can be found there. To help users find appropriate information, there are many free search tools available on the Internet. But searching is still inefficient. Usually formulating queries is some kind of art because the user has to avoid large numbers of useless documents while at the same time he must not lose necessary records.

Existing search systems can be classified in a number of ways. From one site, these kind of systems can be identified as:

- Centralized,
- Distributed.

From another site, there are two main approaches to construct search systems:

- General-purpose,
- Domain-specific.

The classification of information retrieval systems can be done on the base of models they use. Presently the following models are popular:

- Probabilistic,
- Vector space.

Every approach has its advantages and disadvantages. There is a competition between centralized and distributed searchers and between general-purpose and domain-specific systems as well. Most of the famous general-purpose systems like *AltaVista, Excite, Google,* etc. have centralized architecture. General-purpose and domain-specific systems continuously offer a new opportunity for end users to improve the quality of a search. What it still not decided is which model is more sophisticated: probabilistic or vector space.

The idea of a distributed domain-specific search system is promising, and it is getting more popular because:

- Results of a search have to be more accurate,
- Administration of a system should be easier,
- The index of a whole system should be larger when compared to other approaches.

When we are talking about such an approach, the following major tasks should be resolved:

1) How to construct domain-specific collections (databases) semi automatically.
2) How to accurately determine a small number of potentially useful collections to invoke for each user query.
3) How to search inside this kind of collections consisting of topic identical documents for relevant ones.

A preceding solution of the first task is presented in [3]. A method described in that citation works well to compile scientific text documents from the Internet. A vector space model is one of the most powerful tools to resolve the third task [4]. It should be noted, additional work needs to be done to draw more satisfactory conclusions. The second task from the aforementioned set is very important and incredibly difficult to resolve. Different approaches concerning this problem are discussed in the citations [4], [5], [6], [7], [9]. In this paper, we propose a new topic-specific collection selection method applied to the OASIS system

(Open Architecture Server for Information Search and Delivery)., which is a distributed domain specific search system in the Internet [1], [10].

2 Related Work

Source selection is the process of determining which of the distributed document collections are more likely to contain relevant documents for the current query [9]. This task is common for distributed search systems and metasearch engines. The main difference between these kinds of systems is as follows.

A distributed search system consists of several (up to several thousand) search engines, located in different places. Usually each server has its own local database (collection) where indexed data from the Internet are stored. When a user query is processed, servers communicate with each other. Communications are also possible during database construction and updates. The different approaches are applied to build databases, to divide the net between search engines. These kind of systems usually include one or several components, which are responsible for selection of servers to process a user query. Figure 1 illustrates this approach. The bold dotted line in this figure represents communication between systems' servers. A detailed description of the aforementioned systems can be found in [9].

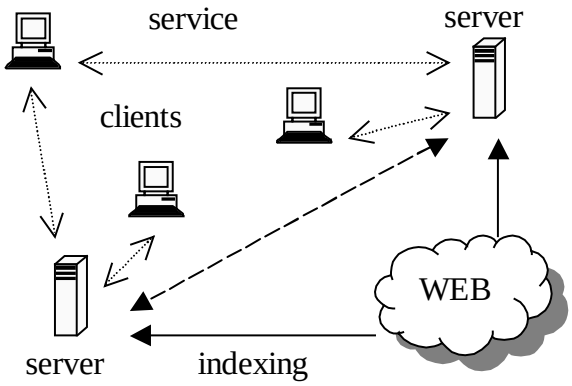

Fig.1 Distributed Search System's Architecture

As it noted in [7], the main feature of a metasearch system is to support unified access to multiple local search engines. These search engines are usually domain-specific. The metasearch system does not maintain its own index on web pages, but it compiles characteristic information about each underlying search engine. Usually a special component is responsible for this task. The architecture of the metasearch system is shown in Figure 2. The bold dotted line there represents interaction between the metasearch server and underlying domain specific servers.

Tasks to select a set of distributed search engines from a distributed system and a set of underlying search engines are similar to each other. From this point of view, these systems are virtually alike. Methods used in one kind of system can be applied in another one. Results obtained before 1999 are discussed in [4], [5], [6], [9]. Review of latest works is presented in [7], [13], [14], [15], [16]

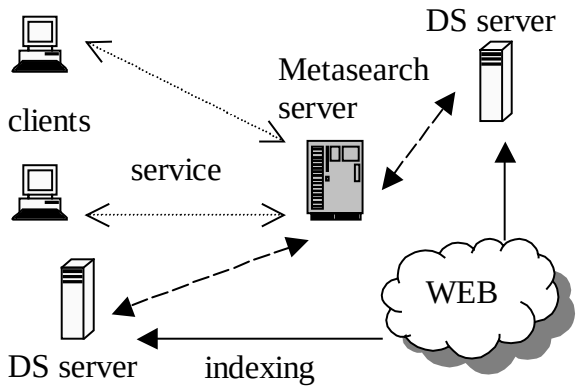

Fig.2 Metasearch System's Architecture

3 Collection Selection

We use the following approach to select which servers should receive a particular search query. The objective of collection selection is to improve efficiency by sending each query to only potentially useful collections; network traffic and the cost of searching useless collections can be reduced. From this, our method never assumes that every collection is equally likely to contain relevant documents and never broadcasts the query to all collections.

The main points to be discussed here are:

- What kind of information must characterize each collection (collection description)?
- How should this information be used in selection of a set of appropriate collections?
- What kinds of tools are used to speed up communication between different parts of the system?
- What should be done to improve the quality of a collection description?
- How often should the collection description be updated?

We are going to give a necessary explanation concerning these questions. A requirement to the part of the system responsible for collection selection is as follows: It should select a set of necessary collections very quickly and accurately.

3.1 Collection description

Each collection is responsible for compiling its own description. Authors of the research [8] found the average length of a user query equal to 2.2 words. This fact is one of essential points of our study. Good results of filtering topic-specific scientific documents from the Internet were obtained in [3]. The method described there proposes to construct a filter using single terms, two and three word collocations. The main thrust of our proposal is to put in the collection description the aforementioned components: single terms, two and three word collocations. The number and composition of these components are up to each collection. A collection description is a set of the following pairs: <term, *tf* weight > and <word collocation, *tf* weight >. A *tf* (term frequency) weight is the number of a term occurrence in the collection. The number of documents in the collection is also a part of description. On the base this information *tf***idf* (term frequency / inverted document frequency) weight is calculated according the standard formula, used in the vector space model [4].

3.2 The Ranking Score

As a basis, we used an approach proposed in [7] to calculate the ranking score of collections. They tested the following measure using TREC collections, articles published in the Financial Times, the Los Angeles Times, the Foreign Broadcast Information Service, Congressional Records of the 103rd Congress and the collection of Federal Register from 1994. (It should be noted, all of these documents are not real Internet data. This is a weakness of their tests.) The ranking score of collection C_j with respect to a query q is as follows:

$$rs(q, C_j) = \max_{1 \le i \le k} \{q_i * gidf_i * mnw_{i,j}\}$$

Here q is a query; C_j is a collection; q_i is a weight of a corresponding term in the query; $gidf_i * mnw_{i,j}$ is the normalized weight of term t_i in collection C_j; $gidf_i$ is a global idf weight of term t_i calculated over all collections; $mnw_{i,j}$ is a maximum from the normalized weight of term t_i in all documents of collection C_j; and k is the number of terms in the query.

We applied this approach in the following style. Because word collocations have greater distinguishing power than single terms, we consider this knowledge in our calculations. The following formula for computing ranking score is proposed:

$$rs(q, C_j) = \sum_{1 \le i \le p}' q_i * gidf_i * mnw_{i,j} \qquad (*)$$

Here p is equal to the sum of terms and word collocations in the query.

It is not necessary to take all collections into account. We agree with authors [7]: it is worth using only first r collections with highest $gidf_i * mnw_{i,j}$ in calculations. For all practical purposes the value 20 for r is good enough.

3.3 Communication Tool

Our approach has been applied to the OASIS system. Each OASIS server can create and support one or more topic specific collections. Servers in the OASIS world are domain specific. This is a main principle, used in the system to divide the net.

OASIS includes a special component called the OASIS Directory. (See Figure 3.)

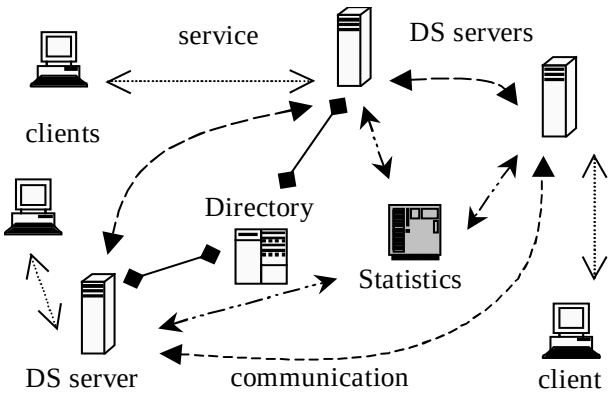

Fig.3 OASIS Architecture

This subsystem is responsible for storing collection description. It assists each OASIS server in selection of a set of collection for query propagation. The aforementioned formula is used by this subsystem in these calculations. The directory is implemented on the base of LDAP servers. The corresponding software is in the public domain [12]. Communication between the Directory and collections is realized via LDAP (Lightweight Directory Access Protocol). A formal description of the LDAP schema was designed to qualify a collection description entry in the directory.

Each entry consists of a set of attributes. An attribute is a type with one or more associated values. The schema concerning [11] is a set of attribute type definitions, object class definitions

and other information. The example of attribute type definition taken from [1] is presented in Figure 4. Accordingly, OASIS has been registered with the Internet Assigned Numbers Authority for a MIB/SNMP enterprise code. The official OASIS OID code is 3284.

(1.3.6.1.4.1.3284.101.120.9
NAME 'NumberDocuments'
DESC 'Number of documents'
EQUALITY integerMatch
SYSTAX 1.3.6.1.4.1.1466.115.121.1.27
SINGLE-VALUE
)

Fig.4 Document Number Attribute

Communication between OASIS servers and the Directory is as follows. The OASIS server sends *p* requests to the directory according to the user query. See Formula (*). The OASIS Directory sends a list of pairs <collection, rank> as a response to each request of the OASIS server. This list consists of *r* elements. Then collections are arranged according sum of ranks: $C_1, C_2, \ldots, C_r, \ldots, C_n$. After that only *r* most appropriate collections are selected for query propagation.

3.4 Updating a Collection Description

The OASIS architecture includes a special subsystem called a Server of Statistics. See Figure 4. It collects a different kinds of statistical data received from OASIS servers. The main purpose of this server is to provide data necessary to improve the quality of the service and the performance of the system. Each OASIS server may regularly send (once per month) a log file to the Server of Statistics. This file includes most frequently used queries submitted to the system by users. Queries consisting of two and three terms are most important. The Server of Statistics stores them. Each OASIS server can request global statistics concerning these kinds of queries to improve a description of the local collection. The aim of this operation is to discard "dead" word collocations from the collection description and to include only collocations found in the statistics. After that, a new collection description can be submitted to the OASIS Directory. The second reason why the collection description should be updated regularly is the alternation of the Internet. The collections are usually not static; they are permanently growing and changing. The collection description should reflect the current state of its content. The period of updating a collection description is once per month on average. The OASIS Directory recalculates rank of collections according new descriptions once per month as well.

We believe the aforementioned mechanism is powerful enough to make necessary improvements in ranking collections and to provide more accurate search results for users of the system.

4 Results of the Tests

Preliminary tests were conducted using real Internet data from our test collections. Two OASIS servers were used in our experiments. One of them acted as the Directory and the Server of Statistics at the same time. Table 1 describes collections of documents installed on servers. More details concerning these collections can be found in [2], [3]. Because it is allowed that collections can cover the same topic, collections of *Programming Languages* on both servers were intersected. They differ from each other in their volume. We needed to know how well our mechanisms could make selection appropriate collections in these situations. As results of the tests have shown, the system selects appropriate collections very accurately in the case of a search for scientific documents. These kinds of queries usually consist of two or some times three word collocations. It is not true for topics like *Museums, Card Games or Travelogues*. In the case of using in queries, words which are scientifically highly precise, corresponding collections were usually selected. For example, queries of "network games" and "optimal solution" generated results of the search about algorithms using in game theory, but not about card games.

Vocabularies used to express knowledge in different areas are usually intersected. Figure 5 reflects the situation with vocabularies from different topics.

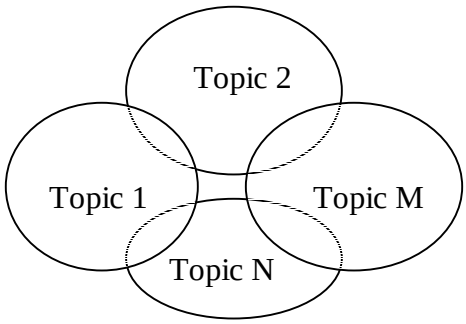

Fig.5 Alignment of topics' vocabularies

There are several possibilities to express a query:

1. Using keywords from a unique part of a topic vocabulary (this kind of search can be

done by very high professional in the area of search).

2. Using common and "powerful" keywords for several vocabularies (this kind of search is done by an experienced user).
3. Using keywords, which are far from the topic of interest (these users are not experienced).

Table 1 Location of the collections

Server	Collections	Number of documents
First	Programming Languages	7659
	Algorithms	7775
Second	Programming Languages	445
	Cars	427
	Travelogues	226
	Linux & Unix	488
	Information Retrieval	202
	Research Groups	811
	Physics	467
	Card Games	798
	Museum	444
	Monitors	70

In case 1, our method selects necessary collections very well, but the percentage of these queries is low.

In case 2, a selected set of collections usually includes ones, which can contain requested information. Exemplary of relevant documents from those collections are usually on the top of the list presented to the user. Feedback from the user can help the system to make selections more accurate. The search process can be done in two or three steps (feedback iterations).

It is incredibly difficult to help the user in case 3. The probability to guess the topic is low. The search for relevant information can take a long time.

Results concerning the accuracy of collection selection are presented in Table 2. They show, that the longer a query is, the more accurate the propagation. Table 3 accumulates the accuracy of topic selection. The percentage of the right selection of collections is high enough for scientific topics: *Programming Languages*, *Algorithms*, *Information Retrieval*, etc.

Our method works well with intersected collections (collections of *Programming Languages* in our tests). Source selection is independent from the volume of collections.

The volume of collected files with queries is not big enough to make significant changes in collection descriptions. A similar task was discussed in [17]. To test more carefully the notion of changing the quality and volume of collection description using the log file with queries, we need to compile many more collections from the Internet and incorporate them into OASIS.

Table 2 Results of query propagation

Query length	Number of queries	Accuracy of collection selection
1	100	60%
2	70	80%
3	40	85%

Table 3 Accuracy of topic selection

Topic	Number of queries	Accuracy of collection selection
Programming Languages	30	86%
Algorithms	30	80%
Cars	17	47%
Travelogues	10	60%
Linux & Unix	20	70%
Information Retrieval	30	75%
Research Groups	10	50%
Physics	10	75%
Card Games	25	60%
Museums	20	70%
Monitors	8	50%

5 Conclusion

A new method of source selection in the distributed search system is presented in this paper. This method was tested using the OASIS system. Preliminary results have shown, it works well in the case of a scientific search: topic specific collection for query propagation were selected very accurately.

This method can be used in distributed search systems and metasearch engine environments.

Additional tests need to be conducted to check the dependence of the quality of a collection description from the statistics of queries submitted to the system.

A prototype system based on the proposed method is publicly accessible: http://oasisntc.u-aizu.ac.jp/oasis/.

References:

[1] *OASIS: Distributed Search System in the Internet*, Edited by A. Patel, L. Petrosjan, W. Rosenstiel, St. Petersburg: St. Petersburg State University Published Press, 1999, - 614 p. (ISBN 5-7997-0138-0)

[2] V. Kluev, V. Dobrynin and S. Garnaev, Intelligent Construction of Thematic Collections, *In. N. Mastoracis, editor, Recent Advances in Applied and Theoretical Mathematics*, pp. 103 – 106, World Scientific and Engineering Society Press, Athens, 2000 (ISBN:960-8052-21-1)

[3] V. Kluev, Compiling Document Collection from the Internet, *ACM SIGIR Forum*, Vol. 34, Number 2, pp. 9 – 14, Fall 2000.

[4] Christopher D. Manning and Hinrich Schuetze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 2000 (ISBN 0–262-13360-1)

[5] David A. Grossman and Ophir Frieder, *Information Retrieval: algorithms and heuristics*, Kluwer Academic Publishers, 2000, - 254 p (ISBN 0-7923-8271-4).

[6] Charles T. Meadow, Bert R. Boyce, Donald H. Kraft, *Text Information Retrieval Systems*, Academic Press, 2000, - 364 p. (ISBN: 0-12-487405-3)

[7] Zonghuan Wu, Weiyi Meng, Clement Yu, Zhuogang Li, Towards a Highly-Scalable and Effective Metasearch Engine, *In Proceedings of 10th International World Wide Web Conference*, 2001

[8] S. Kirsch, The Future of Internet Search: Infoseek's Experiences Searching the Internet, *ACM SIGIR Forum*, Vol. 32, Number 2, pp. 3 – 7, 1998.

[9] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999, - 513 p. (ISBN: 0-201-39-829-X)

[10] V.Kluev and Akiyasu Hayashi, OASIS: High-Speed Network Possibilities, *In Proceedings of the WSES/IEEE Multiconference on Modern Information Technologies and Robotics*, Malta, 2001.

[11] M. Wahl, T. Howes, and S. Kille. *RFC 2251: Lightweight Directory Access Protocol (v3)*, December 1997.

[12] LDAP: http://www.openldap.org

[13] Sergey Melnik, Sriram Rafhavan, Beverly Yang, Hector Garcia-Molina, Building a Distributed Full-Text Index for the Web, *In Proceedings of 10th International World Wide Web Conference*, 2001

[14] Allison L. Powell, James C. French, Jamine Callan, Margaret Colnell, and Charles L. Viles, The Impact of Database Selection on Distributed Searching, *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Greece, 2001.

[15] Xiaolan Zhu and Susan Gauch, Incorporating Quality Metrics in Centralized / Distributed Information Retrieval on the World Wide Web, *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Greece, 2001.

[16] Masumi Narita and Yasushi Ogawa, The Use of Query Text in Information Retrieval, *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Greece, 2001.

[17] Shui-Lung Chuang, Hsiao-Tieh Pu, Wen-Hsiang Lu, and Lee-Feng Chien, Auto-construction of a Live Thesaurus from Search Term Logs for Iteractive Web Search, *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Greece, 2001.