

The Research of Distributed and Cooperative Information Retrieval on the World Wide Web

Wang Jicheng (王继成), Jin Xiangyu (金翔宇)
Yang Xiaojiang (杨晓江), and Zhang Fuyan (张福炎)

*State Key Laboratory for Novell Software Technology, Nanjing 210093, P.R. China
Department of Computer Science and Technology, Nanjing University, Nanjing 210093 P.R. China*

Abstract: A mass of heterogeneous, distributed and dynamic information on the World Wide Web (the Web) has resulted in "information overload". It's an important and urgent research issue to provide users with effective information retrieval service on the Web. Web search engines attempt to solve this problem, yet their effect are far from satisfying. In this paper, we first propose a distributed and cooperative strategy for information retrieval on the Web to substitute the centralized mode adopted by current search engines. Then we present a new information retrieval system model IRSM, which supports the retrieval of metadata about Web documents and uses Z39.50 standard protocol to unify the heterogeneous interfaces of different systems. Based on that, we design a distributed and cooperative information retrieval framework, called DCIRF, to help users in fast and effective information retrieval on the Web.

Key words: Web, information retrieval, search engine, cooperative, Z39.50

1 Introduction

Since its origin in 1991, the Web has evolved into a cyberspace of information, which bears some important characteristics. (1) Large volume: it's estimated that there are at least 300 millions of Web pages in 1998 [1]. (2) Distributed: the information on the Web is distributed over 4 million servers. (3) Dynamic: old Web pages are being updated (the modification of content, the move of location) and deleted, while new pages are being published. (4) Heterogeneous: the types of information on the Web are different, including HTML pages, multimedia files and databases. These characteristics result in "information overload" on the Web. Therefore, It's an important and urgent research issue to provide users with effective Web information retrieval service.

Since 1960 there have been many achievements in the field of information retrieval. These achievements were applied on the Web successfully, which gave rise to search engines such as Altavista, Yahoo. However, the effect of search engines is far from satisfying because they always adopt centralized mode and do not support the retrieval of metadata about Web documents.

In this paper, we first propose a distributed and cooperative strategy for information retrieval on the Web to substitute the centralized mode adopted by current search engines. Then we present a new information retrieval system model IRSM, which supports the retrieval of metadata about Web documents and uses Z39.50 standard protocol to unify the heterogeneous interfaces of different systems. Based on the above, we design a distributed and cooperative information retrieval framework DCIRF to help users in fast and effective information retrieval on the Web.

2 Distributed and cooperative information retrieval strategy

2.1 Centralized information retrieval on the Web

Now the major Web search engines provide retrieval service of Web documents. The common process of search engines includes: using Robot to gather Web documents, creating inverse index of documents, analyzing user's query, calculating the similarity between the query and documents, sorting the retrieval results and relevance feedback from user [2]. Search engines

adopt the typical centralized mode, which attempt to traverse the Web and create tremendous and integrated full-text index for all documents on the Web. Centralized mode brings on serious limitations as follows.

- (1) Huge consumption of resources including network bandwidth and server load.
- (2) Limited coverage: no search engine can index more than 1/3 of the whole Web pages [1].
- (3) Difficulty of maintenance: search engines update index databases not frequently enough that some pages returned to users are no longer valid [3].

Meta-search engines, such as MetaCrawler, combine the results of several search engines to contend with the problem of limited coverage. However, they are unable to solve the above problems completely for their dependence on search engines. With the rapid development of the Web, centralized mode cannot meet the needs of information retrieval service. On the one hand, the information resource to be managed is so tremendous that none of the centralized information retrieval system can fit completely. On the other hand, each centralized system that goes its own way is constructed repeatedly and wastefully.

2.2 Distributed and cooperative information retrieval on the Web

The origin of the Web lies in its usage as an internal cooperative environment at CERN [4]. Now it has become an infrastructure for global information sharing, on which people can publish, retrieve and browse information with no limit of time and location. There exists inevitable mismatch between the distributed and cooperative essence of the Web and the centralized retrieval mode. Thus, we propose a distributed and cooperative strategy for information retrieval on the Web and think it holds promise in future.

In order to analyze the strategy in details, we first introduce some concepts as follows.

Definition 1. Web information space R is the set of all indexable documents on the Web. Note that some documents, such as those created by Web server dynamically, which can not be accessed or retrieved directly, are not contained in R .

Definition 2. Web information retrieval is the process of finding a subset C composed of appropriate number of documents relevant to a certain query q from large volume of Web documents. It can be denoted by a mapping $\xi : (R, q) \rightarrow C$.

Definition 3. A partitioning of Web information space R is a set $S = \{S_1, \dots, S_n\}$, where $S_i \subseteq R$, $S_i \neq \Phi (i = 1, \dots, n)$ and $\bigcup_{i=1}^n S_i = R$.

Definition 4. A partitioning S is consistent, iff $\exists i \exists j, (i \neq j) \wedge (1 \leq i, j \leq n) \wedge (S_i \cap S_j \neq \Phi)$. Otherwise, S is inconsistent.

Definition 5. An element S_i of S is called a Web information sub-space. S_i can be partitioned further to obtain hierarchical partitioning $S = \{\{S_{1,1}, \dots, S_{1,i}\}, \dots, \{S_{n,1}, \dots, S_{n,j}\}\}$ of R .

Definition 6. A partitioning operator ∇ of R is a mapping between R and S , which can be denoted by $\nabla : R \rightarrow S$. For example, let ∇_1 be industrial community and ∇_2 be geographical region, then we can obtain an inconsistent partitioning of R as follows: $R \xrightarrow{\nabla_1} \{\langle .com \rangle, \dots, \langle .edu \rangle\} \xrightarrow{\nabla_2} \{\{\langle .com.cn \rangle, \dots, \langle .com.fr \rangle\}, \dots, \{\langle .edu.cn \rangle, \dots, \langle .edu.fr \rangle\}\}$. Moreover, other factors such as subject of discipline, etc. can also be used as partitioning operator to obtain complicated consistent partitioning of R .

The distributed and cooperative information retrieval on the Web is to partition R , and get a partitioning S . For each S_i , an information retrieval system IRS_i is constructed specifically to provide retrieval service. These systems, which are distributed over the Web, form an Information Retrieval Community IRC , as shown in figure 1. Users can submit queries to corresponding systems according to their needs. When more than one IRS are involved in q , it is required that

q should be decomposed to several sub-queries for each IRS , namely $(R, q) \Rightarrow \bigwedge_{i=1}^k \Omega(S_i, q_i)$, where Ω are Boolean functions, and the retrieval results returned by each IRS should be combined to get result set, namely $\bigcup_{i=1}^k C_i \Rightarrow C$, where Ψ are Set functions.

Once this strategy is adopted, the information managed by each IRS will decrease relatively, consumption will be reduced and maintenance will be easy to carry out. Meanwhile, each IRS can increase its coverage by cooperating with one other. In addition, when one IRS fails, others are still available. It is obvious that this strategy can make up for the weakness of centralized one and improve information retrieval quality. At present, distributed computing community has provided several inter-operation models such as CORBA [5] to integrate autonomous and heterogeneous systems. So IRS can cooperate and supplement each other in the distributed computing environment.

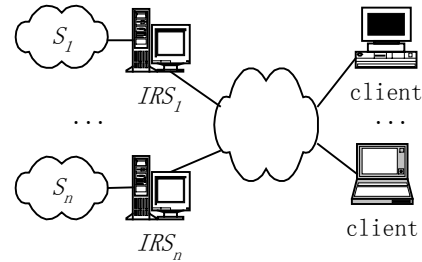


Fig. 1. Distributed and cooperative Web information retrieval Community

3 Web information retrieval system model IRSM

In IRC , each IRS acts as a component whose performance has great impact on the quality of service in IRC as a whole. The retrieval means and representation measures of search engines are monotonous and the accuracy is low. Sometimes users want to search information by metadata such as author and topic. Users also hope that retrieval results contain metadata. Nevertheless, search engines usually provide full-text retrieval that doesn't support retrieval of metadata and the retrieval results are only list of titles and abstracts about Web documents. Besides, the interfaces of different search engines are heterogeneous, which makes it difficult for user to retrieval several search engines simultaneously. Next, we present a new information retrieval system model IRSM, which supports retrieval of metadata about Web documents and adopts Z39.50 standard protocol to unify the heterogeneous interfaces of different systems, as shown in Figure 2.

3.1 Metadata about Web documents

Metadata is the data about the information of Web documents, which can be categorized into two types: descriptive metadata and semantic metadata. Descriptive metadata includes title, date, size, type and so on, while semantic metadata includes author, organization, topic, etc. Metadata reflect the properties of Web documents to great extent. Therefore, it not only can serve as retrieval means to improve the accuracy, but also can be visualization means of retrieval results.

Some elements such as <Title> attribute and <Meta> mechanism are defined in HTML 4.0 [6] to describe the metadata about Web documents. W3C recently issued XML [7] and RDF [8,9] specifications to propose language and framework for richer machine-readable resource description on the Web, which provided the foundation for applying metadata to support information retrieval on the Web. We can design an extractor to extract metadata from Web documents. It is easy to obtain descriptive metadata, while it is relatively difficult to obtain semantic ones. Research achievements in many fields such as computer linguistics, statistics and informatics are helpful in extracting semantic metadata.

3.2 Web documents gathering and processing

On the background, *IRS* gathers documents distributed over multiple Web servers, creates full-text index and extracts metadata, as shown in the top of Figure 2.

It is impossible and unnecessary to gather all documents on the Web for each *IRS*, which is only responsible for managing a corresponding information sub-space. Therefore, administrators can specify a list of URLs to be gathered. A more flexible measure can allow administrators to present gathering strategies consistent with partitioning operator, such as subject of discipline or network domain. These strategies will be stored in the profile and the URL manager will automatically create a list of URLs. Then the scheduler will build plans for gathering and create Robots to download documents on the Web. The indexer will create full-text index for the documents already downloaded. At the same time, the extractor will extract metadata from documents and filter out URLs embedded in the hyperlinks. The URL managers will check these URLs and add new URLs in accord with gathering strategies to the list of URLs. Full-text index and metadata will be stored in database for future retrieval. Systems have to iterate the above process periodically to update databases for the dynamic information on the Web.

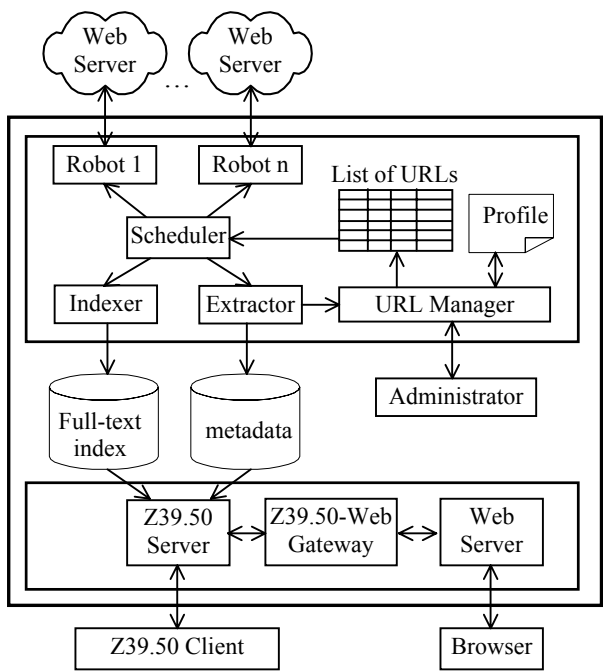


Fig. 2. Web information retrieval system model IRSM

3.3 Retrieval interface based on Z39.50

As a network information retrieval standard, Z39.50 [10] specifies data structure and exchange formats involved in the process of client's requesting and obtaining results from servers. These formats have little to do with the implementation of the information resource and retrieval system, which enables users to retrieve several servers simultaneously through uniform interfaces to obtain part of or all records relevant to the queries. At present, Z39.50 is adopted mostly by libraries. We had developed the first bibliographic retrieval system based on Z39.50 in China last year [11]. We found in practice that Z39.50 is flexible and scalable enough to be used as an interface standard for *IRS*, as shown in the bottom of Figure 2. Consequently, the heterogeneous interfaces of different system are unified.

Table 1. The mapping between metadata and Z39.50 Bib-1/Use attributes

HTML elements	metadata	Bib-1/Use attributes
<TITLE>...</TITLE>	Title	att 4 Title
<META name="subject" content="...">	Subject	att 21 Subject-heading
<META name="date" content="...">	Date	att 30 Date
<META name="abstract" content="...">	Abstract	att 62 Abstract
<META name="author" content="...">	Author	att 1003 Author
<META name="keywords" content="...">	Keyword	att 1016 Any
<META name="content-type" content="...">	Type	att 1034 Content-type
...

After building a mapping table, as shown in table 1, we can use Bib-1/Use attribute set to express retrieval query on metadata about Web documents. GRS-1 is used to return retrieval

results. In addition, we extend Z39.50 in following aspects to express richer retrieval semantics on Web documents.

- (1) Add new attributes in Bib-1/Use, such as 'score', 'rank', 'language', etc. of document.
- (2) Add new attributes in Bib-1/Position, such as 'DocumentBodyText', 'DocumentBodyTable', 'DocumentBodyPicture' and so on.
- (3) Add new attributes type 'Modification' in Bib-1 to describe whether the item in query is case sensitive, or stop-word list will be used and so on.
- (4) Add new attributes type 'Item' in Bib-1 to describe the features of items in query, such as 'language', 'weight', 'count', etc.

To make use of retrieval services based on Z39.50, clients should use tools supporting Z39.50. Due to the popularity of Web browsers, IRSM comprises a Z39.50-Web gateway, which is capable of transform the queries submitted by users through HTML tables into formats recognizable to Z39.50 servers and vice versa. Thus, users can access Z39.50 retrieval servers through Web browsers.

4 Distributed and cooperative information retrieval framework

There exist two types of cooperation in *IRC*. One is task/sub-task relationship between *IRC* and each *IRS*, the other is provider/consumer relationship between user agent (browser, intelligent agent, etc.) and *IRS*. Cooperative relationships are dynamic in open environment like the Web. Each *IRS* can dynamically participate or withdraw and user agents need to dynamically detect *IRS* with certain service. Therefore, we designed a framework, called DCIRF, for distributed and cooperative information retrieval on the Web.

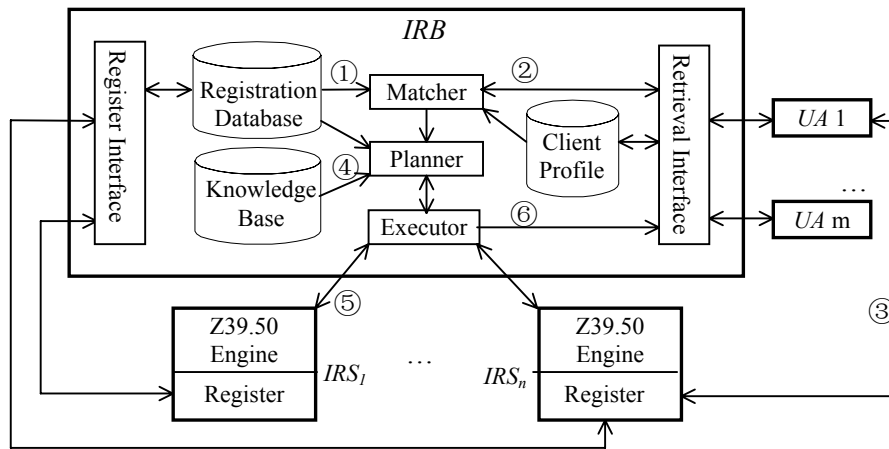


Fig. 3. Distributed and cooperative information retrieval framework DCIRF

4.1 DCIRF Components

DCIRF consists of three components: IRSM-based information retrieval system (*IRS*), user agent (*UA*) and information retrieval broker (*IRB*), as shown in Figure 3. If *UA* has a specific target *IRS*, it can send queries directly to the *IRS*. However, the participation and withdrawal of each *IRS* is dynamic, what's more, *UA* does not have specific target in most cases. *IRB*, as an intelligent middleware between *IRS* and *UA*, can manage the participation and withdrawal of *IRS*, dynamically match or monitor queries sent by *UA* and available services provided by *IRS*.

4.2 Registration of *IRS*

The registration database in *IRB* is used to store information about *IRS*. The registration

information of each *IRS* is denoted by a tuple: $(id, handle, Content, Capability, Constraint)$, where *id* is the unique identifier of *IRS*, *handle* is the handle of *IRS* such as network address, *Content* is service content including subject or region of the information sub-space managed by *IRS*. *Capability* is service capability including metadata attributes, relation (truncation, approximation, etc.) and Boolean functions supported by *IRS*. *Constraint* is service restriction including retrieval syntax and maximum number of results supported by *IRS*.

The registration interface of *IRB* provides *IRS* with three primitives as follows.

- (1) **Register**: *IRS* participates in *IRC* by submitting its service information to *IRB*.
- (2) **Unregister**: *IRS* deletes its service information and withdraws from *IRC*.
- (3) **Modify**: *IRS* updates its registration information in *IRB*.

4.3 *UA*'s query

UA interacts with the retrieval interface of *IRB*. According to the period of effectiveness, *UA*'s query can be categorized into one-off query and durative query. Durative query will be stored in user's profile which will be monitored by *IRB*, and *IRB* will push new result to *UA* periodically. According to the type of result returned by *IRB*, *UA*'s query can be categorized into directory query and detail query. By directory query, *UA* can obtain from *IRB* the handle of target *IRS* for later retrieval. By directory query, *UA* can directly get results from *IRB* (retrieval on target *IRS* is performed by *IRB*, which is transparent to *UA*). Four primitives (**Match**, **Get**, **Monitor** and **Subscribe**) corresponding to these types are offered by retrieval interface of *IRB* to *UA*, as listed in table 2.

Table 2. Type of query and primitive

	one-off query	durative query
directory query	Match	Monitor
detail query	Get	Subscribe

4.4 Flow of distributed and cooperative information retrieval

Now we illustrate the flow of distributed and cooperative information retrieval by a simple example. Assume that the partitioning operator of *R* is network domain. *UA* puts forward a query: find all the documents relevant to "multimedia database" released by universities in China. The query is processed by *IRB* as follows (the number of each step is marked in figure 3).

(1) The matcher searches the registration database for *IRS* which provides retrieval service for the domain of "Universities and Colleges in China". If the partitioning is consistent, the matcher will find more than one target *IRS*, such as *IRS* responsible for the domain of "China" and *IRS* responsible for the domain of "Universities".

(2) If *UA* submits a directory query, the matcher will return the handle of target *IRS* to *UA* and jump to step 3. Otherwise, if *UA* submits a detail query then jump to step 4.

(3) *UA* submits query to each target *IRS* and obtains retrieval results, the process ends.

(4) The planner creates retrieval plans for the corresponding *IRS* based on its capability, constraint in registration database and network routing information in knowledge base. These plans are then submitted to the executor. When there are more than one target *IRS*, the planner will consider other factors, such as resource consumption, connection cost, etc. in order to create retrieval plans for all *IRS*s or part of them.

(5) The executor submits the query: search documents whose subject, title, or keywords containing "multimedia database", to Z39.50 retrieval engine of the target *IRS*.

(6) The executor returns the retrieval results of *IRS* to *UA*. If there is more than one target *IRS*, the executor should post-process retrieval results, including integrating results of each *IRS* and deleting duplicate ones.

4.5 Features of DCIRF

As an open framework for distributed and cooperative information retrieval on the Web, DCIRF features in the following aspects. (1) Transparency: with the help of *IRB*, it is unnecessary for *UA* to comprehend the concrete information of *IRS*. (2) Extensibility: *IRS* is convenient to

participate in or withdraw from *IRC* using registration mechanism. (3) Interoperation: due to the adoption of Z39.50 as retrieval interface of *IRS*, interoperation between *IRB*, *UA* and each *IRS* is easy to achieve.

5 Conclusion

With the flood of information, information retrieval on the Web is a research issue of great potential. In this paper, we present DCIRF, a framework for distributed and cooperative information retrieval on the Web. DCIRF adopts distributed and cooperative strategy, supports retrieval of metadata about Web documents, bears transparency, extensibility and interoperation and overcomes the limitations of Web search engines. We have successfully developed *IRS* based on IRSM and tested DCIRF in Intranet. The work we will do next is to implement DCIRF in a wider and more practical environment. There still remain many topics for further research, such as designing rational partitioning operator, improving information gathering and processing using mobile agent technology [12], and so on.

References

- [1] Steve Lawrence, C. Lee Giles. Searching the World Wide Web. Science, 1998, 280(5360): 98-100.
- [2] Venkat N.Gudivada, Vijay V. Raghavan, William I.Grosky and Rajesh Kasanagottu. Information Retrieval on the World Wide Web. IEEE Internet Computing, 1997, 1(5): 58-68.
- [3] Steve Lawrence, C. Lee Giles. Context and Page Analysis for Improved Web Search. IEEE Internet Computing, 1998, 2(4): 38-46.
- [4] Tim Berners-Lee. The World Wide Web: Past, Present and Future. 1996. <http://www.w3.org/People/Berners-Lee/1996/ppf.html>
- [5] Object Management Group. The Common Object Request Broker: Architecture and Specification, 1998. <http://www.omg.org/library/c2indx.html>
- [6] Dave Raggett, Arnaud Le Hors, Ian Jacobs. HTML 4.0 Specification. World Wide Web Consortium Recommendation, 1997. <http://www.w3.org/TR/REC-html40-971218>
- [7] Tim Bray, Jean Paoli and C. M. Sperberg-McQueen. Extensible Markup Language (XML) 1.0 Specification. World Wide Web Consortium Recommendation, 1998. <http://www.w3.org/TR/1998/REC-xml-19980210>
- [8] Dan Brickley and R.V. Guha. Resource Description Framework (RDF) Schemas. World Wide Web Consortium Proposed Recommendation, 1999. <http://www.w3.org/TR/1999/PR-rdf-schema-19990303>
- [9] Ora Lassila and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. World Wide Web Consortium Recommendation, 1999. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222>
- [10] ANSI. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. ANSI/NISO Z39.50-1995. Bethesda, MD: NISO Press, 1995.
- [11] Yang Xiaojiang, and Zhang Fuyan. Bibliographic Retrieval Based on Z39.50. Journal of Software, (accepted, in Chinese).
- [12] Zou Tao, Wang Jicheng and Zhang Fuyan. Information Service Model with Mobile Agent Technique Adopted. Journal of Computer Science and Technology (accepted).