Cross-Search: A Web-based system for integrated access to multiple heterogeneous information sources

Li Guangjian

School of Management, Beijing Normal University, Beijing 100875, P.R. China.
Library of Chinese Academy of Sciences, Beijing 100080, P.R. China.
ligj@mail.las.ac.cn

Liu Xiaojuan

Library of Chinese Academy of Sciences, Beijing 100080, P.R. China.
Graduate University of Chinese Academy of Sciences, Beijing 100049, P.R. China.
liuxj@mail.las.ac.cn

Huang Yongwen

Library of Chinese Academy of Sciences, Beijing 100080, P.R. China.
Graduate University of Chinese Academy of Sciences, Beijing 100049, P.R. China.
huangyw@mail.las.ac.cn

Abstract

The Cross-Search is a Web-based system providing simultaneous searching across local and remote diverse information sources through a uniform, user-friendly interface. This paper mainly discusses the design idea of the Cross-Search, as well as its objective, architecture, modules, functions and applications. The key problems solved in the system are the connection with heterogeneous, distributed and autonomous information sources; and the merging of results from multi-sources. It gives a solution and implement for retrieving multi-databases simultaneously. The Cross-Search has been successfully applied in the information integration solution of 7 libraries.

1 Introduction

With the rapid development of Web, there has been an explosive and rather chaotic growth in the amount of information available online during the last decades. On the one hand, all those information is highly distributed and heterogeneous. On the other hand, end users prefer to high efficiency and low cost in utilizing this information. In order to solve the contradiction, there is a need to integrate information from multiple distributed, heterogeneous and autonomous information sources [8][11]. Nowadays, there are some famous Cross-Search systems overseas, including ENCompass, MetaLib, NLM Gateway and Muse, etc. The Cross-Search [2] is one of the ways to integrate information sources on Web. It is a Web-based system providing simultaneous searching across local and remote diverse information sources through a single, user-friendly interface. This system can merge, sort and de-duplicate the results from different information sources into a single result set. Results from every information source can also be presented as individual result set as well. In this paper we describe the Cross-Search in detail, including its objectives, architecture, modules and functions.

2 Objectives and Considerations

The Cross-Search aims to integrate the heterogeneous information sources transparently and seamlessly, and form a dynamic information access environment, which offers one-stop searching of the sources. Without learning how to use so many individual sources, users can get what they need through one click or one search by means of the Cross-Search. To this end, during the design and implement of the Cross-Search, we work in accordance with the following designing considerations:

(1) Completeness: The Cross-Search should not only cover all functions of each distributed source, but also assure that the amount and types of information in each integrated source should not be lessened. In other word, the Cross-Search should retain all functions and information of integrated sources.

(2) Standardization: The Cross-Search should support concurrently searching across the heterogeneous information sources, using protocols appropriate to each source. Commonly supported protocols and specifications include, but are not limited to, HTTP, Z39.50, OAI, SQL, OpenURL, Lotus Notes, ODBC, JDBC or JTDS, and API specific to a single source. Meanwhile, XML should be used as a standard of query processing, information exchange and integration, information representations, etc. in the Cross-Search.

(3) Personalization: The Cross-Search should consider carefully the user's preference and provide personalized services to meet their particular need. A user can customize his searching interface and result display.

(4) Expansibility: In the face of information sources emerging in an endless stream, the Cross-Search should have a scalable infrastructure to accommodate or integrate more information sources.

(5) Independence: The Cross-Search should not depend on the information sources to be integrated.

Particularly, The Cross-Search should be independent of the architecture and data structure of each integrated information source. It should make no impact on the

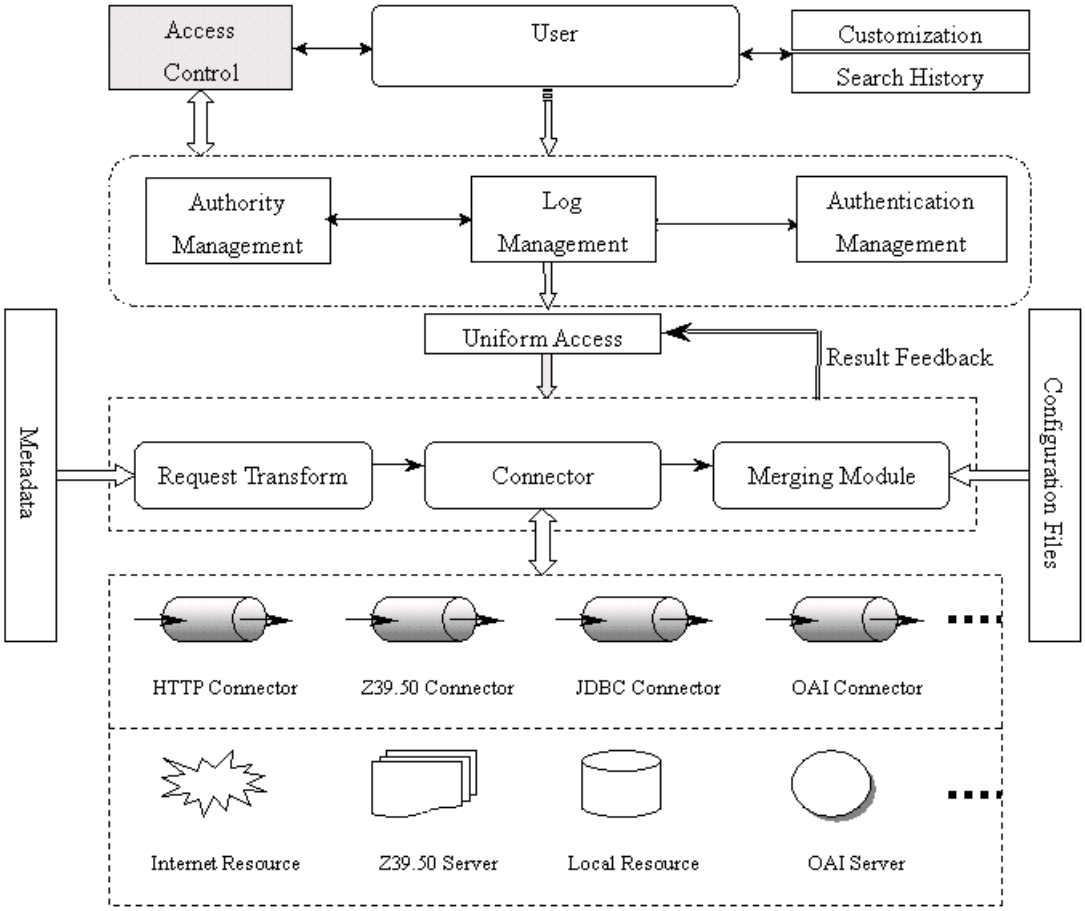

Fig. 1. The Cross-Search System Architecture

integrated information sources. Any information source on Web can be simply added in the Cross-Search as an integrated component without modification of the existing code.

3 Architecture and Realization

3.1 Architecture

The Cross-Search functionally consists of two parts: service-oriented application and management-oriented application. The former includes access control module, user interface module, search engine and merging module etc. The latter includes information source management module, user management module, statistics module, document delivery module and system customization module etc.

Figure 1 sketchily shows general architecture of the Cross-Search.

3.2 Realization

The Cross-Search is implemented in Java. The database server is MS SQL Server and the Web server is Tomcat. The user can access the Cross-search by using standard Web browsers. More particularly, JSP (including JavaScript, JavaBean) and XML are main technologies used in the implement of the Cross-Search. As one of the popular script languages at present, JSP provides a simplified, fast way to create dynamic web applications and enables rapid development procedure. The Cross-Search written by JSP are server- and platform-independent and run smoothly under multi-operation system and multi-web server. XML, considered as a meta-language: a language for defining markup languages, can describe all kinds of information with different structures. Because the heterogeneous Web sources have different structures and formats, XML is adopted as a basis of query processing and result integration in the Cross-Search. By means of its strong ability of data definition and management, XML is better to realize the sharing of heterogeneous sources in the Cross-Search.

4 Modules and functions

The following is the brief introduction about main modules and functions of the Cross-Search.

4.1 Access control module

The access control module is responsible for the whole control of the run of the Cross-Search. This module

includes the functions of authority management, authentication management and log management.

(1) Authority management

Authority management is implemented based on the concept of the roles. Each role has different rights. Each user can belong to one kind of roles. The role of Administrator, with all privileges, is assigned to system administrators and is not allowed to be deleted or modified. Three different rights are included: full-text search, with which users can perform full text search of integrated sources; document download, with which users can download the original document from the integrated sources; and document delivery, with which users can request interlibrary loan and document supply to the Cross-Search. Administrator can create new roles and assign the different rights to the roles. He also can delete or modify the other roles. Now there are two roles, Advanced User and Common User, in the Cross-Search except Administrator. Advanced User has the three rights mentioned above. However, the role of Common User as the default value only has the rights of full-text search and document delivery. Certainly Administrator can modify the rights of the role. As long as a user registers, he or she will have the role of Common User. If needed, the user can have another role. This method makes it so convenient to manage authority.

(2) Authentication management

In the Cross-Search, authentication involves two levels [5]: ① to make use of controlled services provided by the Cross-Search, such as full-text search and customization of information sources; ② to download the original ("full text") documents from subscribed information sources, such as commercial databases and other information sources with controlled access. The Cross-Search has two authentication methods: username/password and IP address.

Username/password: The Cross-Search can verify the identity of the user by the files held on the server.

IP address: Access can be granted or denied based on the network address of the client. It requires configuration of the user's browser.

(3) Log management

System logs contain logging time, IP address, search formula and searched information sources etc. All these information is mainly from the session objects. Logs serve several purposes. Firstly, they help system administrator to troubleshoot all kinds of system and application problems. Secondly, system administrator can check and analyze particular log according to some options, such as information source, user ID, date, time and so on. Thirdly, by means of examining users' log file, the system can get users' information like search history. Then it is helpful for further data mining.

4.2 User interface module

User interface module aims to provide personalized service for each user. At present the Cross-Search supports the following functions.

(1) Information sources selection

The Cross-Search provides the navigation of information sources from the point of category, such as type (Full Text, Abstract and Index, E-book, OPAC, …), subject (Physics, Chemistry, Biology, …), and language (Chinese and English). All these are arranged in a two–level tree of category. Every information source name links to its home web site or Fact Sheet. Along with clicking different first-level category, information sources will change their display. If clicking subcategory, all information sources of corresponding subcategory will change their status (checked or unchecked) [4]. It is convenient for the users to select appropriate information sources. An authenticated user can customize his or her preferred information sources from all integrated sources and form one or several subject groups, which can be named by the users' preference. The user can modify or delete his or her any subject group. To improve the efficiency, the user may directly search in the information sources of a subject group after logging in.

(2) Search display

The Cross-Search provides flexible and multiple search screens, such as simple search, advanced search and second search. The search fields contain title, author, keyword, abstract, ISSN, full text, etc. Title is the default value. The user can also limit the time of search process, 30 seconds as default. If the actual search time of one information source goes beyond the limit, the Cross-Search will terminate the corresponding search thread for the sake of efficiency. Furthermore, a user can configure the rule of sorting (title as default, author and databases) and de-duplicating (title as default, title and author, and without de-duplication) either before or after searching. The interface also allows for any number of alternative views with varying levels of detail to be selectable by a user.

(3) Search history saving

The Cross-Search saves all search history for each user, including search formula, search time, hits and so forth, and allows a user to save it to his or her local disk or email [7]. At the same time, the Cross-Search provides storage space, which is only open to the users with this jurisdiction and privately owned, for a user to hold result records. Preserved records may be kept in a period of time set by the system. By means of personalized space a user can continue the same search at different place and different time.

4.3 Information sources management module

The information sources management module is the underlying support of integrating all kinds of heterogeneous information sources in the Cross-Search. Through this module administrator registers metadata of integrated sources, such as host URL, protocol, search parameter, language and so on. Then a new source can be integrated into the Cross-Search. This module provides a simple web interface, so the task of managing information sources becomes easy and efficient.

(1) Registration

This module supports the registration of new information sources by simple and friendly interface. One information source has many items to record. These items are divided into four groups: basic information, particular information, search parameters and hit extracting parameters. Some items, such as name, port, and basic URL address, need be inputted by administrator. Other items are allowed to be selected from list through radio, checkbox or pull-down menu, such as category, protocol and language. This module provides individual registration interface according to different protocol. All items, which are presently saved in SQL SERVER, are used to realize the process of searching. We plan to adopt XML files to keep the configuration of the information in the next version.

(2) Delete and modification

Administrator can delete or modify integrated information sources. The basic information of registered sources, including category, name and protocol is listed. Administrator can delete one or several information source(s) at the same time. The process of modification is similar to registration. System will guide administrator to finish the modification by four steps.

(3) Source sequencing

Administrator has the ability of customizing the sequence of information sources on the search screen. The adjustment of sequence is implemented by an up arrow button or a down arrow button.

(4) Category management

The Cross-Search provides the functions of managing type category and subject category. For example, administrator can add, delete, modify, sort and hide the type category. As long as providing the name and description, new type will be added. Similar with information source sequencing, administrator can also change the sequence of types by arrow buttons. Any selected type can be hided through a “hide” button. Then the hidden information sources will be absence on the search screen. Subject category makes use of the same method of management.

(5) Source detection

Information sources change frequently. Once the searching interface of an information source is changed, the corresponding metadata will be modified along. The Cross-Search has the function of automatically or manually detecting information sources’ connectivity by sending some search requests. The feedback helps the system to decide whether the information source is available or not. Administrators can get a report of information sources variation, which is helpful for system maintenance.

4.4 Search engine

In the Cross-Search, the search engine is a module of getting search result from distributed and heterogeneous information sources, i.e. a search agent of each information source. Due to the variation and complexity of information sources integrated in the Cross-Search, the search engine must meet the following requirements. Firstly, it can simultaneously search the local and remote information sources. Secondly, it has the ability of searching multi-type information sources, namely Web pages, databases, electronic book and others. Thirdly, it has the ability of searching multiple content types of information sources including abstract, citation, full text and so on. Finally, it can support multi-protocols.

The search engine is adaptable, easily extensible and maintainable. This means putting as much information as possible in a configuration layer. That is easily managed and extended to handle new situations. This separates the mechanics of generating search queries from the details of the query language [6].

The type and amount of protocols supported by the search engine depend on integrated objects. Commonly supported protocols and specifications in the Cross-Search include the following: HTTP, Z39.50, OAI, SQL, OpenURL, Lotus Notes, ODBC, JDBC or JTDS, and API specific to a single source.

The following is main parts of the search engine.

(1) Query transformer

The Cross-Search provides the uniform search interface. Because heterogeneous information sources have different query syntaxes, the search engine must perform a syntactic reformulation of a user query, translating it into queries that have been optimized for the native query language of individual evidence sources [10]. The foundation of transformation is the registration information of each source. As mentioned in section 4.3, the information includes basic URL, port, search fields and other search parameters. When the search request is transferred to query transformer, it will be transformed into the corresponding expression through the mapping between the Cross-Search and the search target. Then the transformed query, which conforms to the query syntax of search target, is formed and sent to connector.

(2) Connectors

Every kind of connector, including HTTP connector, Z39.50 connector, SQL connector, API connector (such as Google) , encapsulates the way of connecting and accessing the corresponding information source. It builds the connection with local/remote information sources and performs the search. A connector is a reusable piece of code that can “federate” within a common protocol (for example HTML)-where the request/response ”language” is the same and the data retrieved has the same general format but differs in its detailed structure. Connectors, with the design pattern of modularization, can track the status of search and return the results. When adding a new information source with a new protocol, the new connector supporting this protocol can be conveniently embedded, only with a relative small piece of code. This kind of open design pattern makes the Cross-Search easily extend the type of information sources.

In addition, in order to improve the efficiency of system response, multithread technology is used in a connector. Meanwhile the parallel operation mechanism is adopted among the connectors. The integration and

display of results is also parallel with search, so the results can be displayed while searching.

4.5 Merging module

The merging module realizes the integration of results from heterogeneous information sources, so users can get uniform result set. Data in various formats is merged into the uniform XML document, in which the elements are mainly based on Dublin Core. In addition, XSLT [3] is used to de-duplicate and sort the records in the XML document. The merging module consists of two sub modules: extractor and display. The structure is shown in Figure 2.

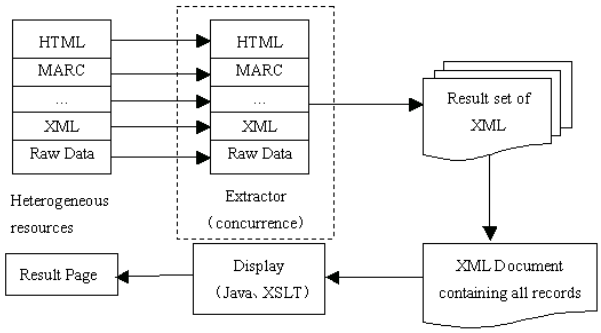

Fig. 2. Merging Module

(1) Extractor

The information sources integrated in the Cross-Search have a variety of data formats. Provided a corresponding extractor, an information source can be integrated through a configuration file (which contains mapping rules between different data formats). Extractor can extract the relevant information in XML format from every information source based on the extraction rules. At present the formats supported by the extractor include MARC, HTML, XML, raw data and so on. The results from existing information sources mainly belong to the formats mentioned above. Because of the design pattern of modularization, a new heterogeneous information source can be added by adding a new configuration file. The methods of the extractor are listed below in detail.

① MARC data

The key point of extraction and transformation depends on the MAP file and DTD file for MARC in different standards, such as CNMARC and USMARC. The MAP file, produced by manually analyzing, is the mapping file between the fields of MARC and the elements of XML document. It can extract obscure machine-readable information into simple and plain XML document. The DTD file, which is used to standardize the XML document, prescribes the elements and the context. New MARC format can be integrated as long as corresponding MAP file and DTD file are added.

② HTML data

The information source responds to the query and returns HTML raw data. Although HTML provides a convenient way to present information, HTML is not data-driven and it is too difficult to automatically extract information from Web page. The HTTP connector passes the raw data to the extractor, which converts it to records in XML format by using the extraction rules for the information source. The Cross-Search firstly extracts Web page with complicated structure into well-formatted equivalent XHTML document and locates citing point of useful information in the document [1]. Then XPATH expression in XSL file can be used to extract useful information. Finally, the results of XML file are exported as predefined format. XSL file is really an extracting and mapping tool. When a new information source is added, the relevant XSL file need be modified.

③ Raw data

Now a majority of databases support XML in a certain extent, especially the familiar RDBMS, such as SQL Server and Oracle. For example, SQL Server 2000 supports XML 1.0, XML Namespaces, Microsoft's XDR (XML-Data Reduced) Schemas, XSLT 1.0, and contains partial support for XPATH 1.0. SQL Server 2000 added an extension to SELECT T-SQL statement by introducing the FOR XML clause. The FOR XML clause specifies that the query results must be returned as an XML fragment, instead of the standard row sets. Without performing the complex database programming, raw data can be extracted into uniform XML document depending on the support of database.

(2) Display

The result sets of XML from each extractor are merged into one XML document, which will contain all searched records, so the records in heterogeneous formats can be integrated seamlessly. Then the XML document is transformed into HTML at the server before the result is presented to the client so that the XML limits of some client browser can be avoided. In the result page, a user can either browse the results from a single database, or sort the results from several databases by author or title. Meanwhile, the user is allowed to specify the modes of de-duplication. The Cross-Search supports detection of duplicates by comparison based on the modes of de-duplication, such as by “title”, by “title and author” or “without de-duplication”, specified by a user. When the user requires de-duplication and a duplicate is detected, The Cross-Search shows one as “primary record” on the screen and the other as “secondary record(s)” by a icon attached to the “primary record”. The user can click the icon to view the duplicate(s), which would come from the different information sources.

4.6 Other modules and functions

(1) Statistics module

The Cross-Search provides statistics functions including database usage statistics, document delivery statistics and user registration statistics, etc. The task of statistics can be appointed by database, date and user role. The statistic result can be exported as Excel file. By checking and analyzing statistic data, the administrator can acquaint himself with the service condition of the whole system to improve this system gradually.

(2) System customization module

Similarly with other Web systems, the Cross-Search has some system information, which frequently needs modification, such as the newest message and system email address. For the sake of convenient maintenance, the interface of system information customization is provided. The administrator can publish and modify system information through this interface at any moment. It increases the transferability of this system.

(3) Document delivery module

The information sources integrated in the Cross-Search include many commercial full text databases. Considering intellectual property and business profit, the data provider limits the use of unauthorized users by IP validation or login. These users can get the basic item of record through the Cross-Search, but they can't get the full text. In order to fully meet the need of users, the Cross-Search provides the service of document delivery by E-mail or mailing. The policy of the document delivery is formulated by the administrator of the Cross-Search.

5 Applications

Initiated in 2001 and then funded by Chinese national Science Digital Library (CSDL), the Cross-search has experienced five versions and integrated nearly one hundred databases of six categories: bibliographic databases, library online public access catalogues, electronic journals from academic publishers, online archives of preprints and grey literature, indexes of mathematical Internet resources.

(1) The version 1 was released at year-end of 2002. It only accomplished the simultaneous searching across 8 information sources.

(2) The version 2 was released in August 2003. In this version some abstract/index information sources and Web OPAC were added into the system. Added functionality included sorting and de-duplicating of result, customization of information sources, authority management, and authentication management.

(3) The version 3 was released in April 2004. Electronic books and free web sources were added into the system. This version provided the dynamically registration of information sources and second search.

(4) Version 4 was released in September 2004. On the one hand, the amount of information sources was increased. On the other hand, functions of documents delivery and system statistics were added.

(5) The current version (version 5) was released in July 2005. This version provides the navigation and automatically detection of information sources. All search results can be fetched as long as the user wants. Administrator had the ability of search and management of registered users.

The Cross-Search has been successfully applied in the information integration solution of 7 libraries, including 3 public libraries, 1 university library, 1 scientific library and 2 government libraries. Table 1 lists the application of the Cross-Search.

Table 1. Application of the Cross-Search

Type	Name	Integrated information Sources
Public Library	Guangzhou Library	17 information Sources (Founder Apabi ebook, Wanfang Data, CNKI Databases, …)
	Dongguan Library	20 information Sources (VIP Database, People's Daily, DRCnet, …)
	Panyu Library	17 information Sources (Springer Link, Blackwell, Yahoo, …)
University Library	Hainan University Library	8 information Sources (BP, EBSCO, Shusheng ebook, …)
Scientific Library	CSDL	97 information Sources (Elsevier ScienceDirect, HighWire, CSHL Online Journals, …)
Government Library	Academy of Macro-economic Research	47 information Sources (China InfoBank, CASS, OECD, …)
	Library of National People's Congress	34 information Sources (Superstar Library, People's Daily, CNKI Databases, …)

Figure 3 is the current user interface of the Cross-Search for CSDL.

Fig. 3. User Interface of the Cross-Search version 5

6 Conclusion

The Cross-Search is a fairly complete solution to the heterogeneous information source integration. Although it is proven in exiting application practices that the Cross-Search is a user-friendly and efficient integration tool, there are various problems to be solved in its further development. More research efforts will be focused on the topics of cross-language search, dynamical integration of information sources and result relevancy ranking [9].

References

[1] HTML Tidy Library Project, 2005. http://tidy.sourceforge.net/

[2] The Cross-Search Web site, 2002. http://cross.csdl.ac.cn/metasearch/jsp/index.jsp

[3] The Extensible Stylesheet Language Family (XSL), 2005. http://www.w3.org/Style/XSL/

[4] Fang, Lin. A developing search service: heterogeneous resources integration and retrieval system. D-Lib Magazine. Vol. 10 No. 3/2004.

[5] Fryer, Donna: ”Federated search engines: federated searching aggregates multiple channels of information into a single searchable point.” Online March-April, v28 i2, p16 (4), 2004.
[6] Ghemawat, S., Gobioff, H., et al. Architecture for knowledge-based and federated search of online clinical evidence. Journal of Medical Internet Research, 7: e52. 2003.
[7] Lawrence C. Kingsland III, Maureen F. Prettyman, Sonya E. Shooshan. The NLM Gateway: a Metasearch Engine for Disparate Resources. In: Medinfo. 2004 Sept, pages 52-56.
[8] Paepcke, A., Brandriff, R., Janee, G., Larson, R., Ludaescher, B., Melnik, S., & Raghavan, S. Search middleware and the simple digital library interoperability protocol. D-Lib Magazine 6(3). 2000.
[9] Paula J. Hane. The Truth About Federated Searching， Information Today Vol. 20 No. 10-Nov./Dec. 2003.
[10] Sergey Chernov, Bernd Fehling, et al. Enabling Federated Search with Heterogeneous Search Engines : Combining FAST Data Search and Lucene
Federated Search Project Report. Version 1.0, 22.03.2006
[11] Tennant, R. Cross-database search: one-stop shopping. Library Journal,126(17), pages 29-30. 2001.