Discussion on the Implementation of Library Resource Discovery Service Systems

Authors: Liu Minjian, Wang Xing, Liu Minjian

Date: 2021-10-05T00:00:00+00:00

Abstract

This article briefly introduces the basic concepts of literature resource discovery service systems, elaborates in detail on the technical framework according to the fundamental principles of literature resource discovery systems, and provides a brief description of the key technologies employed.

Full Text

Discussion on the Implementation of Literature Resource Discovery Service Systems

Liu Minjian, Wang Xing
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract

This paper briefly introduces the basic concepts of literature resource discovery service systems, elaborates on their technical framework based on fundamental principles, and provides a brief overview of the key technologies employed.

Keywords: Resource Discovery, Digital Library, Technical Framework, Key Technology

In the digital library environment, users face numerous challenges when searching for desired resources. Traditional library retrieval systems are often too complex for ordinary users to master quickly. Users typically lack clarity about various metadata standards for library documents and find it difficult to distinguish their characteristics. Moreover, locating and obtaining full-text documents efficiently remains a significant hurdle. Consequently, users strongly prefer simple search interfaces like those offered by Google and Baidu. To address these issues, some database vendors have proposed federated search solutions. These

systems convert a single search request into appropriate syntaxes for multiple independent databases, merge the retrieved results, display them in a concise and unified format with minimal duplication, and provide automatic or user-selected sorting options for the result set [**?**]. However, this approach suffers from several major drawbacks. Most critically, response speeds vary across databases, and according to the barrel principle, the overall system response time is determined by the slowest search server. Under distributed network conditions, the failure of a single search server can cause the entire search to fail, making system reliability quite fragile. Additionally, federated search systems face challenges including lack of unified metadata standards for resource integration, difficulty in deduplicating search results, and the fact that relevance ranking is not always truly based on relevance [**?**].

Due to these problems, federated search provides a poor user experience. As a next-generation library retrieval system, resource discovery systems have emerged. These platforms deeply integrate various types of library resources and provide a single-entry academic resource discovery service. They help readers quickly and accurately locate needed documents within massive information resources, offer the most appropriate access service integration, and deliver an optimal search experience. To accommodate modern user search habits, resource discovery systems provide a simple search box similar to Google, greatly simplifying the relatively complex interfaces of traditional library retrieval systems. Currently, internationally renowned resource discovery systems include Summon, Primo, and EDS.

1. Characteristics of Literature Resource Discovery Service Systems

1.1 Single Search Interface Requirement

In today' s Internet era, users have grown accustomed to search engines with a single entry point. Providing a search engine with a single input box similar to Google or Baidu enhances user experience. Users can simply input a search term to retrieve various types of documents and select those that meet their needs.

1.2 Rapid and Accurate Location of Document Metadata

Users need to quickly locate required documents within massive literature databases. This requires discovery systems to deliver fast search speeds, comprehensive and accurate results, and reasonable ranking algorithms.

1.3 Convenient Full-Text Access

The ultimate purpose of using a resource discovery service system is to obtain needed full-text documents. Therefore, the system must provide the most appropriate full-text access service integration and ensure users have the best possible

experience during the acquisition process.

2. Technical Framework of Resource Discovery Systems

The technical framework of literature resource discovery service systems can be broadly divided into three layers: presentation, application, and data. As shown in [Figure 1: see original paper]:

[Figure 1: see original paper]

The data layer comprises several underlying databases that support system operations, including user databases, source information databases, user review databases, resource evaluation databases, full-text link databases, and literature databases. These databases respectively support the business logic modules in the application layer, with the literature database being the most critical component. The following sections detail the processing workflow for the literature database.

2.1 Electronic Resource Import into the Merge Database

Electronic resource providers may supply data in various heterogeneous formats. To enable proper import into the merge database, we must analyze these heterogeneous formats and establish extraction and mapping relationships to the merge database structure. Specialized tools can directly import heterogeneous data into the merge database, or dedicated software can be developed for specific resources when mapping relationships are complex. Another method for importing electronic resources into the merge database is harvesting. For instance, specialized harvesting programs can be developed to periodically capture electronic resources from specific websites based on update dates. If the quality or accuracy of electronic resources is poor, data cleaning programs may be required to filter out non-compliant data, with manual processing employed when necessary.

To ensure smooth and efficient data harvesting and import operations, a standardized import process should be established. A recommended approach involves defining a unified import metadata specification that stipulates a consistent data structure and content standards for each field. Each electronic resource provider converts their data into formats compliant with this specification, typically XML files. The system then provides a unified import interface allowing providers to import data themselves or designates specific sources for harvesting.

2.2 Data Merging and Deduplication

Since the merge database receives metadata from multiple providers, duplicate data is inevitable. To enhance user experience, merging and deduplication are crucial tasks. Duplicate metadata can be identified using several groups of key identifier fields. For example, the DOI field can serve as a standalone

deduplication criterion, while title, author, year, and page number can form another deduplication standard. These groups of criteria determine duplicate relationships among different metadata records.

After identifying duplicate relationships among multiple metadata records, the next step involves merging them into a single record. A simple strategy selects the metadata from the most standardized provider as the merged record. A more sophisticated approach uses the metadata from the most standardized provider as a baseline, then compares each field of duplicate records to incorporate content from more standardized fields into the baseline record, with manual processing when necessary. While this method improves the quality of merged metadata, it reduces efficiency and increases system complexity.

2.3 Data Content Standardization

The system includes a dedicated subsystem for data content standardization, containing standardized tables for author names, institutions, keywords, funding sources, subject classifications, journal titles, languages, publishers, and other fields. These tables maintain mappings between standardized and non-standardized content. When new data is imported, the standardization subsystem can normalize non-standard content based on these tables. The standardization tables should be continuously updated to ensure data accuracy and validity.

2.4 Indexing and Index Service Provision

To enable users to quickly retrieve desired literature metadata, indexes must be created for merged and standardized data. Based on user search requirements, metadata fields requiring indexing should be identified. Generally, a universal single search box requires full-text indexing of all searchable fields. If the system provides an advanced search mode, non-full-text indexing of entire field contents is needed to meet precise field-specific search requirements. After index files are created, full-text search engines can provide retrieval services. Functionally, these engines can be divided into search modules and storage modules. When users submit search queries, the search content undergoes appropriate word segmentation based on language before being submitted to the search module. The search module then accesses index files through the storage module and returns the results to users.

The application layer is divided into several subsystems based on functionality:

2.5 Standardization Subsystem

This subsystem contains various standardization tables responsible for normalizing content in specific fields of the literature database.

2.6 User Review Subsystem

This subsystem manages user comments and ratings for literature.

2.7 Unified User Authentication Subsystem

This subsystem handles user authentication and permission management, such as restricting access to certain resources to specific user groups.

2.8 Context-Aware Resource Subsystem

This subsystem determines resource access permissions based on user location (e.g., within the library). Some resources are only available in the library environment. It can also provide information about holdings in nearby libraries based on user location to facilitate physical borrowing.

2.9 Holdings Subsystem

This subsystem records the holdings of various electronic resource providers and performs standardization and deduplication at the source level. When users locate a document with multiple holdings, all options are displayed for user selection.

2.10 Source Management Subsystem

This subsystem catalogs all source information for electronic resources and records arrival information.

2.11 Resource Evaluation Subsystem

Unlike the user review subsystem, this subsystem evaluates sources or documents based on scientific metrics and objective indicators using various tools or plugins. A notable analysis tool is EBSCO' s PLUMX.

2.12 Full-Text Location Subsystem

This subsystem locates network addresses for requested full-text documents using static or dynamic methods. If a document has multiple holdings, users can select their preferred source for full-text access.

2.13 Search Subsystem

This subsystem provides literature search services—the most fundamental and important service in the system and the core function of resource discovery services.

The presentation layer consists of web services composed of front-end pages, which can be grouped by function:

2.14 Literature Search

This includes single-input-box search, multi-condition search for professional users, advanced expression search pages, and search results pages.

2.15 Resource Navigation

This enables navigation of different literature types by journal title, conference name, classification number, and other fields to help users quickly browse desired documents.

2.16 Search Result Faceting

This provides faceted display of search results based on multiple criteria such as publication year, classification, document type, and journal title.

2.17 Full-Text Location and Access

This provides a set of pages for users to obtain desired full-text links, with document delivery service pages available when full-text links cannot be found.

2.18 User Evaluation

This allows users to comment on or rate specific documents through relevant pages.

2.19 Context-Aware Resources

This recommends nearby library holdings based on user location to facilitate physical borrowing.

2.20 User Authentication and Permission Management

This includes user login, registration, and personal information management pages. Display and access to certain resources also depend on current user permissions.

3. Key Technical Aspects of Resource Discovery Service Systems

The construction of resource discovery service systems requires support from several key technologies, which are introduced below.

3.1 Full-Text Retrieval Technology

As analyzed above, the most frequently used module in resource discovery service systems remains the literature search function, which can be considered the core determinant of system success. Therefore, selecting a full-text search engine

that can accurately and quickly retrieve needed documents from massive literature databases is particularly important. Based on overall user requirements, full-text search engines should have the following characteristics:

First, retrieval response speed must be fast. User experience directly depends on the time required to return search results; shorter times yield better experiences. In the current big data era, literature retrieval systems contain extremely large data volumes, making response speed particularly critical when searching massive datasets.

Second, excellent Chinese word segmentation algorithms are essential. Word segmentation is fundamental to full-text retrieval. Unlike English and other Western languages where words have clear separators, Chinese lacks formal delimiters between words. While Chinese characters, sentences, and paragraphs can be easily segmented through obvious delimiters, words present a much more complex and challenging problem than in English [**?**]. As Chinese word segmentation is a foundational technology in natural language retrieval, considerable research has been conducted. Reference [**?**] categorizes Chinese word segmentation algorithms into three types: string-matching-based methods, understanding-based methods, and statistics-based methods, and discusses the current state of Chinese word segmentation and its application in search engines.

Third, the utilization of thesauri in Chinese word segmentation. Unlike general internet search engines, scientific literature contains numerous specialized terms. Generic Chinese word segmentation algorithms do not perform well in this context. A thesaurus, also known as a subject heading list, is a semantic dictionary composed of terms and their relationships that reflects semantically related concepts in a discipline [**?**]. The term list from a thesaurus can be incorporated as a specialized lexicon into the custom dictionary for full-text word segmentation, significantly improving segmentation accuracy. Reference [**?**] proposes adding Medical Subject Headings (MeSH) terms to general word segmentation dictionaries and utilizes MeSH vocabulary combined with word length and position weighting to implement automatic keyword extraction strategies for medical news webpages.

3.2 Cloud Computing Technology

Due to the enormous volume of literature data in resource discovery service systems, the resulting index databases are also extremely large and often cannot be handled by a single machine. The indexing solution involves dividing the entire document collection into several subsets and establishing a distributed cluster, where each machine maintains a portion of the overall index and multiple machines jointly complete index construction and query response [**?**]. When users submit search requests, a distribution server forwards the request to multiple search servers. Each search server returns its results to the distribution server, which then merges and sorts them before returning the final results to users.

Currently, mainstream full-text search engines support this parallel computing

expansion. As the number of users and literature data volume in resource discovery service systems continues to grow, required server resources also increase. The search module can be deployed on cloud computing platforms, dynamically adjusting the number of servers used for parallel search computation based on changes in system traffic and data volume. This approach offers advantages over traditional data centers, including reduced operational costs, dynamic scalability, and simplified maintenance.

3.3 Full-Text Location and Access Mechanism

The ultimate purpose of accessing a resource discovery service system is to obtain needed full-text documents. Satisfying this final requirement is critical to system success. Multiple mechanisms can be employed to locate and access full-text documents.

For in-house electronic resources, which are the easiest to access, full-text documents can be located and retrieved directly based on their addresses.

For literature metadata with DOIs (Digital Object Identifiers), which identify content objects in digital environments, full-text data can be obtained through the DOI. By entering http://dx.doi.org/ in a browser address bar and inputting the document DOI in the "Resolve A DOI Name" prompt box, clicking "Go" directs the DOI system to automatically link to the document' s URL and display the corresponding page. Users who have purchased access rights to the document' s database can download the full text directly; otherwise, individual purchase may be required.

For literature IDs or full-text URLs provided by electronic resource providers, the provided URL addresses can be used directly to obtain full-text documents. Alternatively, full-text URLs can be constructed by following specific formats using the provider' s literature ID. The disadvantage of this method is that URL addresses may change, leading to access failures. To prevent this, the system must periodically update invalid full-text links.

Dynamic full-text acquisition involves using program code to search within resource providers'systems based on key document fields (title, author, publication date, etc.) and return the URL address of the most matching document. This is a dynamic full-text address resolution process that can be completed offline through periodic background resolution and updates or performed in real-time upon user request. The drawback of this method is accuracy; since searches do not always return the correct document metadata, incorrect results may be retrieved.

Literature resource discovery service systems should comprehensively employ these various mechanisms to help users locate and access full-text documents while providing the best possible user experience throughout the process.

3.4 Search Result Ranking Algorithm Design

When users search massive datasets, they may receive large result sets. Determining how to rank the documents users truly need at the top of results is another critical success factor. Ranking algorithms calculate a ranking index for each document based on two main aspects.

First, relevance between documents and search conditions must be calculated. If a document' s title exactly matches the user' s search terms, relevance is highest. For partial matches, search terms should be segmented and the term frequency of each word in document metadata fields (including title, subject terms, keywords, abstract, etc.) should be calculated. A total weight is then computed based on certain weighting schemes to represent the relevance degree between the document and search conditions.

Second, various document or source metrics should be considered, such as publication date, document type, citation count, source citation count, whether it is a library-owned resource, and the document' s browsing and full-text request frequency within the system. Academic or peer reviews can also be referenced, combined with the system' s own user evaluation functions.

The final ranking index is obtained by summing these various metrics with specific weighting coefficients, which may need adjustment based on system performance or user feedback.

4. Summary and Outlook

Library retrieval systems have continuously evolved with the times, progressing from traditional systems to federated search, and then to resource discovery systems. Current literature resource discovery service systems adapt to the Internet age by fully considering user search habits and providing maximum convenience for users requesting needed full-text documents. Future resource discovery systems will continue to prioritize user needs and experience as their primary principle.

While current retrieval systems are essentially still based on search terms, future semantic search technologies may be applied to resource discovery systems as search engine technology advances. With the popularization of mobile Internet, literature resource discovery systems should be ported to mobile devices, enabling users to access the system anytime, anywhere. Users should also be able to perform more personalized customization, such as customizing ranking algorithms or adjusting weighting systems for ranking indicators when dissatisfied with search result ordering.

[?] Ma Hua. The Rise, Current Status, and Development Trends of Major Foreign Federated Search Systems[J]. Library Construction, 2009, (3): 1-5.
[?] Chen Jiacui. Federated Search Mechanisms and Their Existing Problems[J]. Library and Information Service, 2006, 50(6): 87-89, 103.
[?] Wang Xing. Technical Architecture of the National Engineering Technology

Digital Library[J]. Digital Library Forum, 2013, (10): 14-19.
[?] He Shen, Wang Wangu. Research Progress and Application of Chinese Word Segmentation Technology in Natural Language Retrieval[J]. Information Science, 2008, 26(5).
[?] Li Jing, Qian Ping. The Differences and Connections Between Thesauri and Ontologies[J]. Journal of Library Science in China, 2004, 30(1): 36-39.
[?] He Xiaoyang, Zhang Jingli, Ding Ting. Automatic Keyword Extraction Strategy for Medical News[J]. Chinese Journal of Medical Library and Information Science, 2014, (4): 13-.
[?] Wang Hao, Zhang Zhengfeng, Feng Wei. Research and Implementation of Library and Information Resource Discovery Systems[J]. Digital Library Forum, 2013, (6): 51-56.

Note: Figure translations are in progress. See original paper for figures.