ATLANTIS PRESS

*Advances in Intelligent Systems Research (AISR), volume 142*

2017 International Conference on Information Technology and Intelligent Manufacturing (ITIM 2017)

# Construction of A General Academic Search Engine Based on Multi-source and Heterogeneous Data

Qiangkui Leng, Shurui Wang, Qi Yan
College of Information Science and Technology
Bohai University
Jinzhou, China
qkleng@gmail.com, srw@gmail.com, lnyanqi@163.com

Yuping Qin
College of Engineering
Bohai University
Jinzhou, China
jzqinyuping@gmail.com

***Abstract*—For researchers, the retrieval needs on multi-source and heterogeneous academic data are increasingly urgent. However, the existing retrieval systems still have some shortcomings, such as expensive prices, poor integration, loss of information. In this paper, we present a general framework for designing a unified retrieval system, called the multi-source heterogeneous academic literature search engine (Abbr. MH-ALSE). MH-ALSE is constructed into three phases, i.e., information collection, data integration, and engine implementation. Furthermore, the collaborative filtering mechanism is also introduced, for achieving personalized recommendation and early warning service. It is expected that MH-ALSE can provide easy-to-use retrieval functions to help researchers develop novel ideas and improve research efficiency.**

***Keywords—unified retrieval engine; MH-ALSE; heterogeneous data; collaborative filtering***

## I. INTRODUCTION

Academic literature retrieval can help researchers to establish the starting scientific objectives, and avoid invalid repetitive work. The effective capture of academic data is one of the main means of scientific research to achieve results. At present, researchers access to literature through several major academic resources platform, such as CNKI [1], IEEE Xplore [2], Elsevier [3], Springer [4], etc. However, due to the large number of literature and the existence of duplication, researchers have to carefully identify and select, resulting in waste of human and material resources.

The typical feature of academic data is multi-source and heterogeneous. They exist in the form of structured data such as traditional relational database, or semi-structured data such as format files, and unstructured data such as multimedia files, which caused sharing difficulties. In addition, many organizations in the establishment of their own applications and data storage, there is no unified planning and management. It makes the data become "information island", and is difficult to achieve integration and unity.

For researchers, the needs of mutual search on multi-source and heterogeneous data are increasingly urgent. Heterogeneous data integration is not a simple integration on database model or the establishment of a global database view, but makes autonomous systems to cross-platform information collaboration [5-6]. It integrates distributed heterogeneous data sources together so that users can access these data sources in a transparent way [7]. At present, many organizations have adopted public modeling or data exchange tools (UML, XML) to carry out standardization. Moreover, they build the norms and standards, called meta-model for specific area of complex information resources [8]. However, each organization, in the meta-model standardization of only for its scope, takes a specific way to develop their own standards. Thus, there are still double differences of conceptual definition and adopted techniques [9-10].

In recent years, a variety of academic resources unified retrieval system has emerged. These systems have their own characteristics due to the use of different technologies. According to the search style, they can be divided into two categories: metadata-based index centralization and the system based on real-time processing.

Metadata-based index centralization can aggregate multiple heterogeneous sources into service centers through crawling, mapping and importing. Ex Libris's Primo Central [11] maintains and updates data from different resource providers, and achieves a centralized index of integration of these data. Serials Solution's Summon Unified Search Index [12] is also a metadata-based indexing system similar to Primo Central. On the whole, metadata-based system has a faster retrieval speed, and the result integration is relatively easy, but it has the difficulty of updating and maintaining resources.

The system based on real-time processing requires real-time response, instant analysis and feedback of the user's retrieval request. It converts the retrieval request into a expression of the corresponding source, and concurrently retrieves multiple heterogeneous data on the local and Internet. Explorit [13] developed by Deep Web Technology is an instant processing system that retrieves hundreds of repositories at the same time and returns highly relevant search results. Finally, the results are presented in a smart clustering fashion. Ex Libris's Metalib [14], Innovative's Encore [15], Swets' SwetsWise Searcher [16] also uses the similar mechanism. However, the retrieval speed on real-time system is relatively slow, and when the retrieval interface changes, the re-configuration is needed.

Based on the above analysis, the establishment of a simple and practical academic resources unified retrieval system is

particularly significant. It should also provide customized services, and can recommend and push the preference information. Moreover, it will carry out early warning of new research hotspots and major academic events. This paper will construct a unified search framework for academic resources, which we call the multi-source heterogeneous academic literature search engine (MH-ALSE). MH-ALSE is designed for the researchers to provide a unified search platform for academic resources, and in which collaborative filtering mechanism is joined to achieve personalized information recommendation and early warning services.

## II. CONSTRUCTION OF MH-ALSE

The general framework of MH-ALSE is shown in Fig.1. It is divided into three phases.(1) Establishment of collection system for multi-source heterogeneous academic literature. The academic resources are collected through the following ways, such as Z39.50 client configuration, protocol exchanging, web crawler crawling, etc. (2) Integration of data and construction of retrieval platform. The system will provide a variety of different retrieval methods, such as database-based centralized retrieval, open system-based Lucene text retrieval, and real-time source retrieval. Finally, the search results are treated in a uniform form after being processed by removing duplicates, merging, sorting, paging, and clustering. (3) Generation of collaborative filtering mechanism. According to the user's browsing and retrieval behavior, some collaborative filtering algorithms will be applied for recommending the interested resources or content in real time. Meanwhile, it will also provide early warning services for new research hotspots and major academic events.

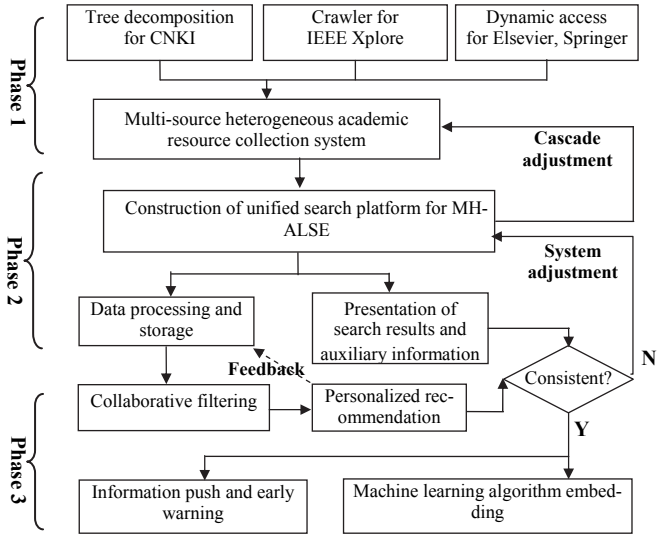


Fig. 1. Flowchart of MH-ALSE schema

### A. Academic Resource Collection

Resource collection is the first step in building a localized repository, playing a key role throughout the system. We will use two kinds of means for gathering data, i.e., protocol-based way and web crawler. In terms of protocol access, we achieve client configuration through standard communication protocols, such as Z39.50. It also includes some open protocols, such as standard XML, or open APIs. Web crawler approach is more difficult than the protocol-based one, since it is mainly through the open page to collect resources, and its performance determines the speed and quality of accessing to data. The crawler collection module is depicted in Fig.2. After gathering through the above means, the academic resources are eventually collated and stored into the local database.

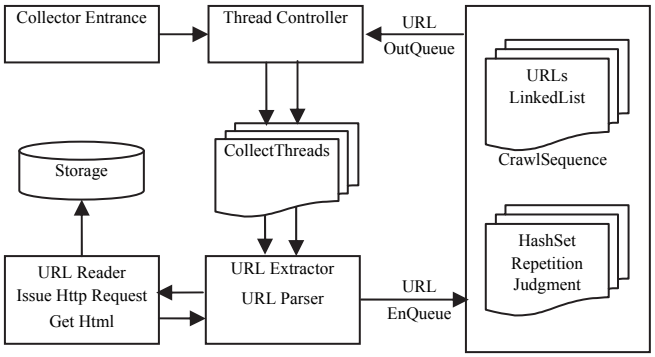


Fig. 2. Crawler collection module

In addition, the resource collection module also needs to implement the resource update, that is, the incremental collection of resources on a regular basis. At the implementation level, update function should provide two ways, namely real-time acquisition and planned updates. Real-time acquisition can be performed by the administrator or automatically after the new address of source database is added to the system. Whereas scheduled updates are set to solve the problem of periodic updates.

### B. Data processing and platform construction

On the basis of the gathered raw data, we start to design the database model, which is directly related to retrieval performance. The stability and efficiency are also considered. In the database design, we follow the appropriate requirements and norms. If the entity can be well modeled, and the relationship between the entities can be well analyzed, then the database design will play a vital role in the entire system.

The relationships in the database must first follow the principle of entity integrity, that is, the key values cannot be empty to ensure that each entity is valid and distinguishable. In addition, the principle of referential integrity is also followed. If there is a link between the two relations, appropriate foreign key should be added. The retrieval system also ensures that data redundancy is as low as possible, while avoiding inserting, deleting, and updating exceptions.

After the data processing, unified retrieval platform can be established. It consists mainly of two functional modules: (1) Integrated search module. The module is the core component of the retrieval system. It needs to provide three different retrieval methods, namely, centralized retrieval based on database index, Lucene-based text retrieval, and real-time retrieval for direct data sources. After removing duplicates, merging, sorting, paging, and clustering, the retrieval results are presented in a unified form. (2) Resource browsing module. The module includes at least two kinds of ways. One is the navigation on e-journal name, and another is on subject name. Among them, e-journal navigation can be directly linked to the home page of the journal, or can be directly localized to

the local resource database with the latest paper resources. Subject navigation can be found through the corresponding subject of the link of source database, or directly located to the relevant subject in resource database.

### *C. Collaborative Filtering Mechanism*

We further introduce the collaborative filtering mechanism into MH-ALSE for analyzing the user-specific behavior. The mechanism provide the services of personalized recommendation, push, and early warning. Fig.3 shows its description.

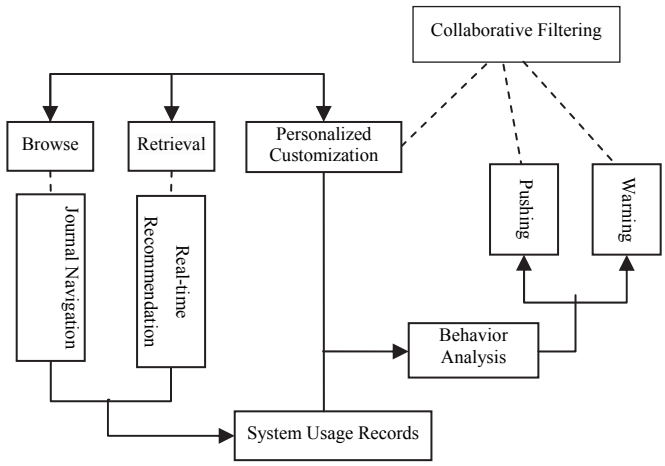


Fig. 3. Collaborative filtering mechanism

Personalized recommendation is usually done by machine learning algorithms. For the push, the MH-ALSE will first combine the user's personalized custom information and the operating behavior records to make a specific analysis, for further establishing the corresponding preference model. Then, the resources that the user may be interested in will be presented or mailed. Early warning can discover new research hotspots and significant academic events, through the analysis and mining of all resources.

As a result, we construct the MH-ALSE framework is shown in Fig.4.

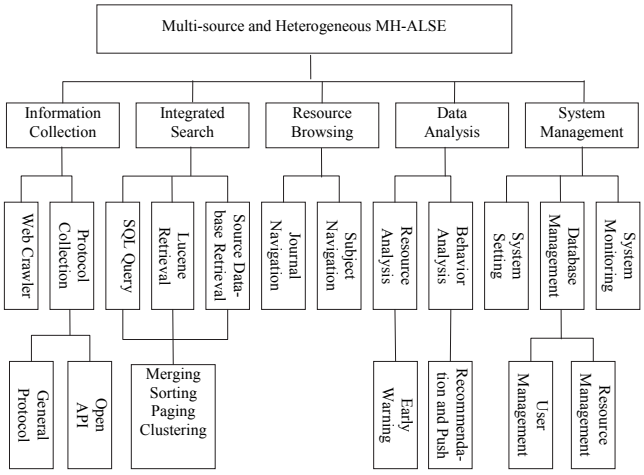


Fig. 4. MH-ALSE Framework

## III. CONCLUSION

In this paper, we presented a unified retrieval system, i.e., MH-ALSE. It collects academic resource in protocol-based way or web crawler, and standardize and structure these resource. Then, the database model is built by following the appropriate requirements and norms, for reducing redundancy and avoiding exceptions. After the data processing, unified retrieval platform is finally established. It provides different retrieval way, such centralized retrieval based on database index, Lucene-based text retrieval, and real-time retrieval for direct data sources. Moreover, the collaborative filtering mechanism is introduced, for achieving personalized recommendation and early warning service. It is expected that MH-ALSE can provide easy-to-use retrieval functions to help researchers develop novel ideas and improve research efficiency.

## ACKNOWLEDGMENT

This work was supported in part by the Scientific Research Project of Liaoning Provincial Committee of Education under grant LZ2016005, the National Natural Science Foundation of China under grant 61602056, the Doctoral Scientific Research Foundation of Liaoning Province under grant 201601348.

## REFERENCES

[1] China National Knowledge Infrastructure. http://www.cnki.net/
[2] IEEE Xplore Digital Library. http://ieeexplore.ieee.org/Xplore/home.jsp
[3] Elsevier. http://www.sciencedirect.com/
[4] Springer. https://link.springer.com/
[5] Sheth A P, Larson J A, Cornelio A, et al. A Tool for Integrating Conceptual Schemas and User Views[C]// International Conference on Data Engineering. IEEE Computer Society, 1988:176-183.
[6] Silberschatz A, Stonebraker M, Ullman J. Database systems: Achievements and opportunities[J]. Communications of the ACM, 1991, 34(10): 110-120.
[7] Lenzerini M. Data integration: A theoretical perspective[C]//Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002: 233-246.
[8] Hollingsworth D, Hampshire U K. Workflow management coalition: The workflow reference model[J]. Document Number TC00-1003, 1995, 19.
[9] Calì A, Calvanese D, De Giacomo G, et al. Data integration under integrity constraints[M]//Seminal Contributions to Information Systems Engineering. Springer Berlin Heidelberg, 2013: 335-352.
[10] Fagin R, Kolaitis P G, Miller R J, et al. Data exchange: Semantics and query answering[C]//International conference on database theory. Springer Berlin Heidelberg, 2003: 207-224.
[11] Vaughan J. Ex Libris Primo Central[J]. Library Technology Reports, 2011, 47.
[12] Breeding M. Summon: A New Search Service from Serials Solutions[J]. Smart Libraries Newsletter, 2009, 29(3):1-3.
[13] Explorit overview. http://www.deepwebtech.com/products/explorit-overview/
[14] León J C C, Par C P, Planas C R, et al. MetaLib[J]. MetaLib - ResearchGate, 2008.
[15] Fifarek A. The Birth of Catalog 2.0: Innovative Interfaces' Encore Discovery Platform[J]. Library Hi Tech News, 2007, 24(5):13-15.
[16] Breeding M. SwetsWise Searcher improves its federated search performance with Deep Web Technologies[J].librarytechnology.org, 2010.