Research on Smart Organization and Retrieval of Government Affairs Archives

Yun-Liang Zhang[1,2,a*], Lin-Na Li[1,2,b] and Zhi-Hui Liu[1,2,c]

[1] Institute of Scientific & Technical Information of China, Beijing 100038;

[2] Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content

[a]zhangyl@istic.ac.cn, [b]liln@istic.ac.cn, [c]Liuzhihui@istic.ac.cn

*Corresponding author

Keywords: Government affairs archives, Knowledge organization, Smart, Indexing, Retrieval.

Abstract. Knowledge organization is important in the services of government affairs archives. To provide better archive service, a new approach of smart organization and retrieval with existing and updated archives related Chinese knowledge organization systems are proposed and implemented. With knowledge organization systems, recommendation, results amount adjustment, ranking and display could be improved.

1. Introduction

Nowadays, with the advent of the information society, information technologies have played an increasingly important role in national economic and social development. With the acceleration of the informationzation process, mining electronic archives and developing public service system are the trends for the archives industry to adapt to the trend. Informatization is also the method to innovate archives service mechanism and to improve archives public service capabilities. There are different types of archives of government affairs, such as legislative archives, administrative archives, military archives, diplomatic archives, economic archives, scientific and technological archives, art archives, etc. With the full implementation of the "Government Online Project" and the "National Archives Information Construction Implement Program", the public service system of government affairs archives based on knowledge organization systems has become the composition of e-government construction.

With the increasing of government affairs archives, the traditional data utilization relying on the searching of file number and title can't satisfy the various kinds of public user's needs for retrieval of enormous file content. But it is the starting point and the foothold of the archival data utilization service for the public, and the public archives services should include more knowledge organization elements, which can utilize the knowledge service mode and ability of the government affairs archives.

2. Related Work

It is important for smart organization to use a specific knowledge organization system, in fact in government affairs archives organization, there are some different knowledge organization systems and applications.

2.1. Related Knowledge Organization Systems

2.1.1. Classification

The "Chinese Archives Classification" has been published to the second edition. The first edition was published in 1987[1] and contained 57 professional tables that contains around 5000 categories. The second edition of the Chinese Archives Classification was published in 1997[2], and it expanded the nineteen categories hierarchy in the first edition. It also adjusted and supplemented the basic

Copyright © 2019, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

categories. The total amount of categories has increased to 100,000, which is 20 times that of the first edition. Relatively speaking, the categories of national economic management, industry and agriculture, and scientific and technological research have increased even more. This was in line with the actual needs of China of the large number of economic and scientific archives. At the same time, the important categories of economic and trade management, taxation, foreign trade and economics were upgraded.

2.1.2. Thesauri

UKAT (UK Archival Thesaurus) [3] is a subject thesaurus and controlled vocabulary created by the British Archives Department for the indexing and retrieval of archive collections and catalogues by subject and provide relevant and consistent subject searches for users. UKAT was created between June 2003 and August 2004 from subject terms contributed by individual archives, projects and users, as part of a project funded by the Heritage Lottery Fund (HLF), The National Archives and University of London Computer Centre (ULCC) etc. UKAT is derived from the UNESCO thesaurus and include different fields such as Education, Science, Culture, Social and human sciences, Information and communication, Politics, law and economics, Countries and country groupings, Events and so on. After 2004, the thesaurus are maintained by Archives in London and the M25 (AIM25) team.

The “Chinese Archives Subject Thesaurus” is a supporting project for the “Archives Recording Rules”. It is a reference book for the indexing and retrieval of Chinese archives. It is the first comprehensive vocabulary of the archives compiled by China. The first edition was published in 1988[4] and contained 27288 terms, in which 22759 are narrative terms. The second edition was published in 1995[5], which contained 21785 narrative terms and 4106 non-narrative terms.

The “Comprehensive E-government Thesaurus” [6] is the first comprehensive e-government thesaurus compiled in accordance with national standards. It was completed in January 2005, and it contains 20252 terms, including 17421 narrative terms and 2831 non-narrative terms.

2.2. Related Applications

2.2.1. The Archival Portal Europe

This application [7] are funded by the Archival Portal Europe Foundation. It provides access to 274,450,640 descriptive units of archives from different European countries as well as information on 26,824 persons and entities, and 7068 archival institutions throughout the continent. The application provide different ways of search, in addition to a simple full text search, the Archives Portal Europe offers an advanced search concentrating on aspects such as dates as well as a navigated search for browsing through the archival material. All approaches can be followed separately or in combination. With the search engine, users can do searching with topics, finding institutions, and exploring featured documents, and it also support multilingual search of English, French, German and Dutch and so on, which is important for an application for all the Europeans.

2.2.2. Shandong Archives Catalogue Center

Shandong Archives Catalogue Center[8] is a typical application of Provincial Archives Service in China. The center gathers archival resources of Shandong Province and its 16 prefecture-level cities such as Jinan, Qingdao and Yantai. It can provide the society with centralized and unified online retrieval of archives which have been opened through the Internet and remote cross-library use of archives which are combined online and offline. Internet users do not need to register and use “Open Archives” to inquire about archives that have been opened through the Internet in Shandong Province. Through the archives, users can check the information related to the utilization of the work of archives in Shandong Province, such as the introduction, contact telephone, working hours and so on. Users registered and authenticated in Shandong Unified User Management and Identity Authentication Platform can use the “Archives Hall” to realize remote utilization and cross-library service of archives in Shandong Province by combining online and offline. They can use “Message Consultation” to submit consultation to archives in Shandong Province on issues related to the use of relevant and

directory center system. Simple full-text retrieval and retrieval function in results are provided. It can provide metadata fields of archival documents such as title, document number, responsible person, time, keywords, notes, file number, belonging archives and operations.

3. Smart knowledge organization

To get smart retrieval, smart knowledge organization are necessary, and it can be implemented by cooperation with suitable knowledge organization systems and proper indexing work. In our opinion, the Classification are stable and can be used now and the thesaurus should be updated.

3.1. Knowledge organization systems adjustment

The “Chinese Archives Subject Thesaurus “and the “Comprehensive E-government Thesaurus” can be used to organize the archival documents. But it should be noticed that they was finished in 1995 and 2005 respectively, so we must update them with some new information resources.

3.1.1. Term from Report on the Work of the Central Government

From the central governmental portal [9], public can get the Report of the Word of the Central Government, public can download the annual report from 1954 to 2019. Refer the publishing time of the two thesaurus, the ones after 1995 are focused, so 25 reports are aquired and put together as a corpus. With a voting term extraction algorithm composed in tf-idf algorithm, c-value algorithm and text-rank algorithm, 2000 high ranking terms are extracted.

3.1.2. Knowledge Organization System Update

Two different thesauri and a list should be merge, of course manpower intensive work can be adopted, but in this paper the terms are combined from literal information into a vocabulary. And then it was processed with a specific procedure, which is shown in Fig. 1.

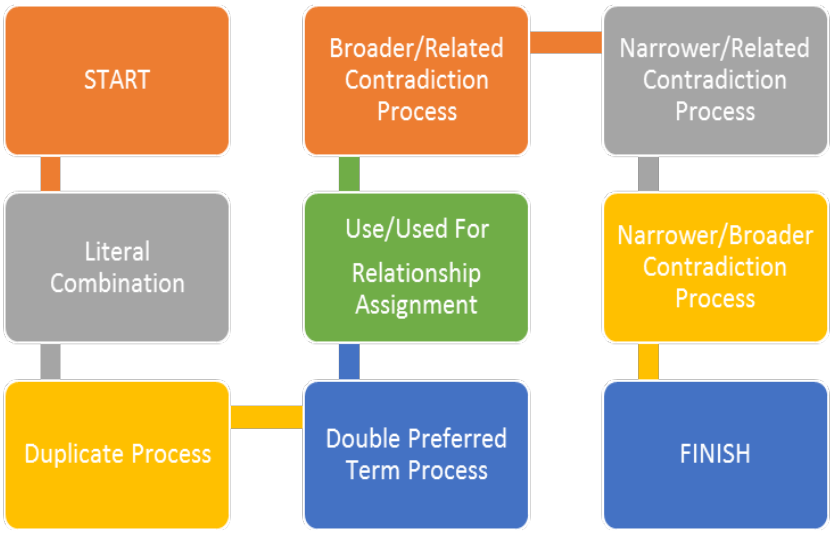

Figure 1. The Process of Knowledge Organization System Update

At first, all the terms from three sources are combined together, and then the duplicate triples of two entries and a relationship are deleted. If there are two preferred terms in it, a preferred term should be identified. At present, the most relational terms from the synonym set are chosen as the preferred words, and the rest are the alternative terms. The relationship be assigned with Use and Used For. Then all the other relations appended to the non-preferred terms are transferred to the new preferred terms. The remaining types of relations include related, narrower and broader. Next, relationship with same terms but different types of relationships should be chosen and deal with according to the following principles: 1) if only one of the Narrower/Broader relationship appears, this relationship will be it, otherwise, the related Related will be used; 2) if the relationship of Narrower and Broader

appears simultaneously, the Related relationship will be selected. In this process, we hope to minimize manual judgment, mainly because the original thesaurus is built and validated by experts. If we rely on one or two people for manual judgment, it is difficult to ensure accuracy. According to the above principles, the consistency can be relatively guaranteed, and the speed is high.

3.2. Government Affairs Archive Indexing

Use the “Chinese Archives Classification” and archival vocabulary to index the archival documents will increase the convenience and satisfaction of the public use the archive services.

3.2.1. Class Indexing

To give the archival documents proper classes, machine learning methods are used. Several algorithms such as SGD, Bayes and SVM are used, but the results are not good enough. So some neural network algorithm are involved in our work, so we can get a classifier with a higher average micro mean precision and recall.

3.2.2. Subject Indexing

Here an open sourcing tool named Maui indexer [10] are used to provide the subject indexing. Some indexed archival documents are used as training set, and the vocabulary should be changed into rdf triples. Here D2R[11] are used as the engine to change the vocabulary contents from MySQL database to rdf file. part of the mapping file are shown as follows:

```
@prefix : <#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2009/08/skos-reference/skos#> .
@prefix jdbc: <http://d2rq.org/terms/jdbc/> .

:database a d2rq:Database;
    d2rq:jdbcDriver "com.mysql.jdbc.Driver";
    d2rq:jdbcDSN "jdbc:mysql://127.0.0.1/vocabulary?autoReconnect=true";
    d2rq:username "root";
    d2rq:password "istic";
    jdbc:keepAlive "3600";                    # sends noop-query every 3600 seconds
#   jdbc:keepAliveQuery "SELECT 1";           # optional custom noop-query
    .

# Table concept
:classmap_concept a d2rq:ClassMap;
    d2rq:dataStorage :database;
    d2rq:uriPattern "http://www.istic.ac.cn/archive#c_@@concept.CID@@";
    .

:concept_Czh-CN a d2rq:PropertyBridge;
    d2rq:belongsToClassMap :classmap_concept;
    d2rq:property skos:prefLabel;
    d2rq:column "concept.Czh-CN";
    d2rq:lang "zh-zh-CN";
    .
```

```
# Table relation
:classmap_relation1 a d2rq:PropertyBridge;
      d2rq:belongsToClassMap :classmap_concept;
        d2rq:property skos:broader;
        d2rq:refersToClassMap :classmap_concept;
      d2rq:condition "relation.REL = 'broader'";
      d2rq:join "relation.CID1 = concept.CID";
      d2rq:join "relation.CID2 = conceptcopy.CID";
      d2rq:alias "concept AS conceptcopy";
      .
:classmap_relation2 a d2rq:PropertyBridge;
      d2rq:belongsToClassMap :classmap_concept;
        d2rq:property skos:narrower;
        d2rq:refersToClassMap :classmap_concept;
      d2rq:condition "relation.REL = 'narrower'";
      d2rq:join "relation.CID1 = concept.CID";
      d2rq:join "relation.CID2 = conceptcopy.CID";
      d2rq:alias "concept AS conceptcopy";
      .

:classmap_relation3 a d2rq:PropertyBridge;
      d2rq:belongsToClassMap :classmap_concept;
        d2rq:property skos:related;
        d2rq:refersToClassMap :classmap_concept;
      d2rq:condition "relation.REL = 'related'";
      d2rq:join "relation.CID1 = concept.CID";
      d2rq:join "relation.CID2 = conceptcopy.CID";
      d2rq:alias "concept AS conceptcopy";
      .
:classmap_relation4 a d2rq:PropertyBridge;
      d2rq:belongsToClassMap :classmap_concept;
        d2rq:property skos:alterLabel;
        d2rq:column "conceptcopy.Czh-CN";
        d2rq:lang "zh-CN";
      d2rq:condition "relation.REL = 'alterLabel'";
      d2rq:join "relation.CID1 = concept.CID";
      d2rq:join "relation.CID2 = conceptcopy.CID";
      d2rq:alias "concept AS conceptcopy";
      d2rq:lang "zh-CN";
      .
```

4. Smart Retrieval

The framework of smart retrieval is shown in Fig. 2, the archival vocabulary are used in three main function: smart recommendation, Smart results amount adjustment, and the classification can be used to display the archival documents in different categories.

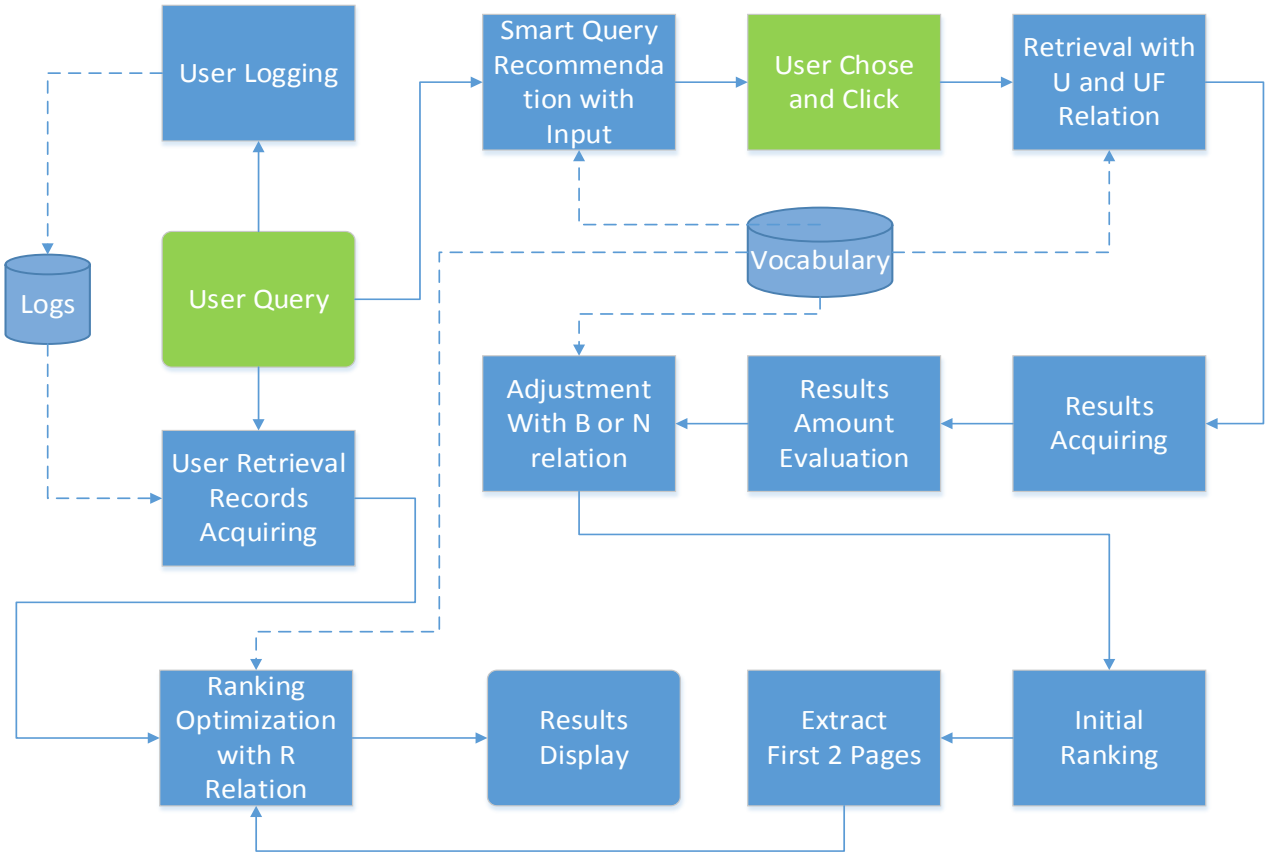

Figure 2. The framework of Smart Retrieval

4.1. Smart recommendation

To provide smart recommendation, all the terms of the archival vocabulary are used, and it is a dynamic process, with the input change by the user, the recommended terms are also changed. So the users would feel the application is smart and he would input less than expected.

4.2. Smart results amount adjustment

To provide smart results amount adjustment, we use Use/Used For relations to expend the retrieval results, since the synonyms are considered. Then the results amount need to be evaluated and if it is not enough, the system will use Broader term to increase the amount of results, otherwise the system will use the Narrower term to decrease the amount of results.

4.3. Smart ranking

Up to now, only Related terms are not used, and they could be used to provide smart ranking, since it is time-consuming, so we only process the first two pages, and re-rank the results with the query and its related terms, if there are more related terms in a document, it will be more closer to the query and should be ranked higher.

5. Summary

The archival documents should be used better and provide sound service to the public. With existing knowledge organization systems and corpus, we update the archival vocabulary and then with this vocabulary and the “Chinese Archives Classification” we indexed all the archival documents and then provide smart retrieval with the indexed documents, and it is a more friendly and effective retrieval system.

6. Acknowledgement

This research was financially supported by Key R&D Program Projects in Shandong Province (Grant No. 2017G006002), CKCEST Project Program (Grant No. CKCEST-2019-2-2), ISTIC Key Project Program (Grant No. ZD2019-10).The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

7. References

[1] China Archives Classification Editorial Committee, China Archives Classification(In Chinese),China Archives Publishing House, Beijing, 1987.

[2] China Archives Classification Editorial Committee, China Archives Classification(In Chinese), second ed., China Archives Publishing House, Beijing, 1997.

[3] Information on http://www.weld.labs.gov.cn https://ukat.aim25.com/

[4] Chinese Archives Subject Thesaurus Editorial Committee, Chinese Archives Subject Thesaurus(In Chinese),China Archives Publishing House, Beijing, 1988.

[5] Chinese Archives Subject Thesaurus Editorial Committee, Chinese Archives Subject Thesaurus (In Chinese), second ed.,China Archives Publishing House, Beijing, 1995.

[6] Zhao Xinli, Comprehensive E-government Thesaurus (In Chinese), Science and Technology Literature Publishing House, Beijing, 2005.

[7] Information on https://www.archivesportaleurope.net/

[8] Information on http://mlzx.sdab.gov.cn/

[9] Information on http://www.gov.cn/guowuyuan/baogao.htm?baike

[10] Information on https://sourceforge.net/projects/maui-indexer/

[11] Information on http://d2rq.org/