The Research on Domain-Oriented Information Resource Management and Retrieval

Gao Shaochen

***Abstract*—All kinds of information repositories now abound. In large-scale information repositories, it's a very important thing to get the required resource quickly and intelligently. This paper presents a framework of intelligent resource management and retrieval which is based on domain-oriented and natural language processing. A framework is given to indicate how the system is constructed. And many key technologies, which support the system implementation and implementation process, are also described. Finally, the paper introduces operation of real prototype system and discusses future work of the system.**

***Index Terms*—Information resource, domain-oriented, search.**

I. Introduction

With the database and network technology and other computer-related rapid development and application of information technology, the Internet has been to provide the valuable information about the main ways the Internet can be retrieved increasingly large number of information resources. This information can be retrieved for the relevant personnel, increasingly complex social information needs and the value of information is difficult to determine, for each one trying to find a specific resource problems have been proposed, especially in certain professional and technical personnel, provides a traditional retrieve the does not have the fast and efficient features at the same time, information resources, search results will bring cell problems such as duplicate resources, similar resources, non-professional resources, these resources do not want to get in a lot of invalid information and spam, and useful information but was lost in the flood of information resources, cannot be reasonably and effectively be queried. In particular, some professional and technical person will need for specific areas of information resources retrieval.

During domain-oriented information resource construction, are often faced with the following considerations: each of these areas and what kind of resources; each of these areas where access to resources; built area resources how convenient and practical. These three aspects of the problem is usually the repository construction field problems, therefore, in the field of information resources is necessary to form the effective management of resources, effective organization as a key consideration. Advanced manufacturing such as the present, the concept of free trade, it is necessary to form these features under the field of information resources management platform.

Information retrieval refers to the information represented, stored, organized and accessed on the basis of the user to deal with the problem solving find, identify, access to relevant facts, data, events and processes the literature. Retrieval of information resources is the core of user information needs and comparative literature collection and selection process are match each other. On the one hand, is the user information needs, on the other hand is well-organized collection of information that is retrieved from the user-specific information needs, a collection of information on the specific use of certain methods, techniques, and rules based on certain clues to find out relevant information. The Fig. 1 gives the design procedure.

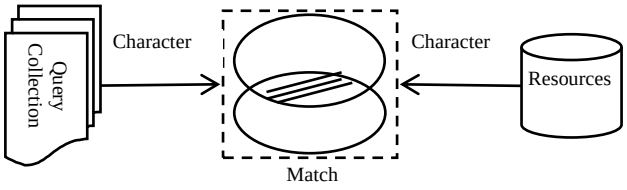

Fig. 1. Design procedure.

Retrieval of information resources is a common problem faced, especially in the field of information resources for the retrieval, the retrieval keyword is relatively simple, but the results for the search results in a professional and authoritative demanding. Currently, there is a clear need for keyword searches, be possible to retrieve the need to locate resources according to the formulation can be quickly found the necessary resources.

The organization of this paper is, the second part describes the information resources framework, the third section describes the information retrieval, the next illustrated a detailed experiment and a case study. Finally, there are conclusions and future works.

II. Information Resource Framework

Currently, there are several ways of information resources, such as information resources: books, electronic books, audio, video, policies and laws, case. Now, a variety of resources are available on the Internet access.

The information resources also have a certain particularity. In the traditional information resources, there are the main sources, such as journals, newspapers, magazines. According to the existence of the field, and its information resources will also involve a lot of "gray literature", and these documents only from specialized research resources institutions or organizations in order to obtain, so in the field of specialty

Manuscript received September 20, 2013; revised November 25, 2013.

Gao Shaochen is with the Shanghai University of International Business and Economic, Shanghai, 201112 China (e-mail: gaosc@ suibe.edu.cn).

oriented information resources sharing of these resources has important significance.

In domain-oriented information resource, whether it is serving the external economic development, or for professional research institutions to enhance the level of research, with the explicit demand-driven. And now there are information resources, all the related resources are put together without distinction. This will not only conducive to the management of information resources, and but also conducive to the use of the user's query.

A. Framework

Domain-oriented resource retrieval framework resource management and resource retrieval will open to independent, specifically shown in Fig. 1, the top of the figure to provide various system resources on the domain-oriented management interface and the user interface.

Resource management including resource organizations to provide various storage resources in a way. According to the different ways of resource organization can be divided into centralized resource storage, distributed storage resources, resource description information stored centrally. In domain repository in the field of library storage including single and multi-field information resource database maintenance.

Resources to support local search and retrieval across the database search. Retrieval pretreatment and retrieval strategies for field re-search word generation and optimization, domain support, keywords differences (body support), multi-lingual support.

Information Retrieval architecture includes a description of information resources and processing, indexing information resources management, information resource retrieval and cross-database search. The Fig. 2 gives the information retrieval architecture.

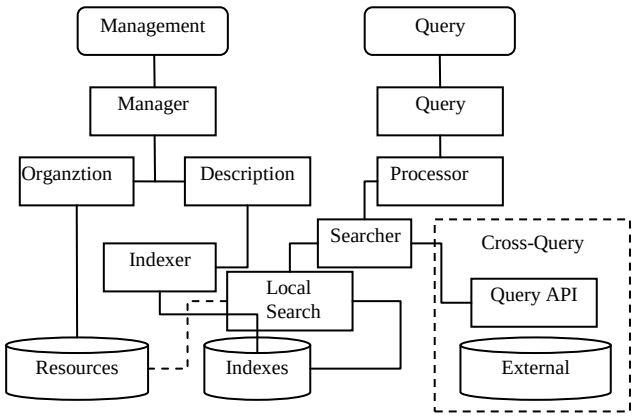

Fig. 2. Framework.

B. Meta Description

Meta information is used to describe resource properties. In order to correct and effective organization, massive resources management system must have the ability to access resources and structural meta-information of the description.

Information resource meta-information including title, subject, abstract, author, type, URL, keywords and other content, the system uses an XML file that describes the resource meta-information.

C. Meta Information Process

Resource library resources to be indexed well-defined xml document, defined by the corresponding xmlschema xml constraint specification. Resources are not to be indexed separate xml document set, but in the xml document defined resource nodes, each node represents a resource in a resource repository, which is a resource retrieval system, various elements of resources information through the resource node contains multiple child nodes (such as Title, Subject, etc.) describe. For xml document parsing using apache's Digester [5], according to the label xml schema defined using digester corresponding analytical rules defined, the information resource definition document for each resource node converted to the corresponding meta-information object definitions , a resource node extract the entire contents of a meta-information defined in the object, then the resource index.

D. Incremental Index

Indexing function uses Lucene the API, the system was built in the resource index, the index is created using org.apache.lucene.document.Document class and make retrieval, therefore, the information resource of data sources, it is not right All tagged information for indexing, segmentation and query returned. In the Document class object, which is a group of org.apache.lucene.document.Field class objects, such a domain name and domain of values, the data source name is the label information. The domain name value is the label values, indexing and retrieval of information resources value in the label for the retrieval and indexing. The data source information resource indexing and retrieval rules in Table I:

TABLE I: INDEX RULE

Label	Inder	Store
title	Yes	Yes
subject	Yes	Yes
abstract	Yes	Yes
identifier	No	No
medium	Yes	Yes
created	No	Yes
learner	No	Yes
Resource type	Yes	Yes
uid	No	Yes

In order to improve the efficiency of the index, the system provides a new index is created and expanded incrementally indexed two ways to meet the needs of different users. When a resource library a small part of the resources of the content has been changed, re-create the index for all the resources relatively time-consuming, so incremental indexing is necessary. Incremental indexing, first open the existing index, through the resource identifier that uniquely identifies each tag Delete Index same identifier tag library that resource, and then add the new access to the document indexing library.

III. INFORMATION RETRIEVAL

Introduction of synonyms and related words to support the treatment of the concept -based information retrieval methods to improve retrieval system for domain-oriented semantic understanding of user request. The core issue is how to extend the search criteria in order to help users get the information.

This system consists of database from the 2000 edition Text extracted synonym sets. Retrieved from the actual needs, in order to reduce as a result of expansion entries noise data increases, the system only for queries containing the largest semantic information for nouns and verbs synonymous expansion , and on other parts of speech of words without processing .

Users enter keywords such as " WTO " repository resources are not and the " WTO " This keyword matching information , while on "WTO" information in the thesaurus "WTO" and " World Trade Center " are synonymous , if not Extended use of semantic words , the result will be empty , the use of synonyms expanded into a "WTO" to retrieve the keyword , will return the "WTO" relevant information to improve the search capabilities .

For related lexicon , the system provides the tools by the system maintenance personnel for maintenance. In this particular area facing massive information system, known in the art due to maintenance personnel relationships between various information, relevant content thesaurus will be efficient and rational organization.

When the system is constructed indexes constructed three types of indexes, retrieval query words entered by the user or after treatment, the use of these three indexes query words do retrieval. According to information resources describe the characteristics of meta-information that title, abstract, subject labels can best embody a combination of resource content, so the system default for each resource title, abstract, subject index label content for retrieval.

According to the Lucene Boolean query model, the query words to make an exact match, with these words segmentation associated Boolean operators AND , constitute a Boolean expression (" computing " AND " computer " AND " machine network " AND " network " AND " network teach " AND " Guide") , these words are included in the meta information will there return results containing " computer ", "network " , "Tutorial ", " computer network " and other resource information of these words will not be returned. Then the user at query time recall rate is very low.

Retrieval system will query input into the vector space model, based on user input keywords entered by the user determines the order of the weight of the query words , the first word is a query has the highest weight , then the weight of query terms in descending order [8]. Based on user input query words to construct a query tree as a query object. Query term for the leaf nodes of the tree, the query tree structure shown in Fig. 3:

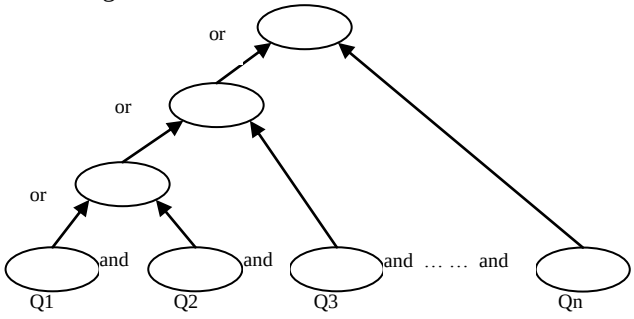

Fig. 3. Query tree.

Word in the query sequence query tree, the first tree of the first leaf node and all non-leaf nodes of the tree of the query sequence of words query , the query words and symbols as follows logically associated to four query words Q1 , Q2, Q3, Q4 for example :

Q1 or (Q1 and Q2) or (Q1 and Q2 and Q3) or (Q1 and Q2 and Q3 and Q4) or (Q 2 and Q3) or (Q2 and Q3 and Q4) or (Q3 and Q4)

In keyword query, the user enters the query words separated by spaces , the first word entered by the user as the standard for the separation boundary queries, that is to say between words are separated by spaces , if hit document is returned to the user, otherwise , the user enters a query word segmentation lyricist table , do the query. The reason for this lies in the premise to ensure precise queries to improve query recall ratio.

For example, the user enters "Computer Network Course" , the system first uses vocabulary segmentation , "Computer Network Course" in accordance with the vocabulary words cut into a " computer ", "network ", " course" , according to the system configuration and the use of query tree vector space model , the return of the resources must contain the first keyword " computer " and return the results may not include the "Network" or "Tutorial ." Sort results returned by these three words appear many times, the word weights via IF-IDF formula to sort the results of the score . Therefore, the ratio provided by third parties on the Chinese Lucene search method is more effective.

The system set a different query strategy to simplify the user's query input to improve the query accuracy and efficiency. Retrieval method has two modes: Exactly Query and expanded query.

A. Exactly Query

Based on user input keywords entered by the user determines the order of the weight of the query words, the first word is a query has the highest weight, followed by the query words in descending order of weight. According to the general user's query habits, usually, the input query words in 3 to 4 or less, therefore, this system only the first three of the user input query word given a higher weight, the weight of the subsequent query term average the default value of 1. Words and word are the logical relationship between or. Therefore, the original Boolean query: Q1 or Q2 or Q3 convert: Q1 ^ 4 or Q2 ^ 2.5 or Q ^ 1.5.

Keyword query, the user enters the query words separated by spaces, the first word entered by the user as the standard for the separation boundary queries, ie between words are separated by spaces, if hit document is returned to the user, otherwise , the user enters a query word segmentation lyricist table, do the query. The reason for this lies in the premise to ensure precise queries to improve query recall.

B. Expanded Query

Query expansion is a concrete realization of the concept retrieval technology to realize the concept of retrieval, through the introduction of domain ontology, the user enters the keyword semantic conversion, based on semantic information retrieval. The establishment of the initial domain ontology and ontology supplement and perfect the way through the man-machine combination to achieve. The initial

domain ontology vocabulary through analysis of a representative sample of the document obtained in the information retrieval process replenishes the emerging field of vocabulary, dynamic sound domain ontology.

The expanded search retrieves achieve query words entered by the user as the target word in the thesaurus to find synonyms for the word set. And synonym expansion set to replace the original query words entered by the user in the index for the query. Words and logical relationship between words or, but the original query term expansion will have a word synonymous higher than its weight, the default value is 3.0. Therefore, the original user enters a query: Q1 ^ 4 or Q2 ^ 2.5 or Q3 ^ 1.5, converted to : (Q1 ^ 3or Q11 or Q12 or Q13) ^ 4 or (Q2 ^ 3 or Q21 or Q22 or Q23) ^ 2.5 or (Q3 ^ 3 or Q31 or Q32) ^ 1.5. From the expansion of the scale can be seen, although the concept of expansion can find out which documents related to the query with the input, but at the same time also increases the noise data returned results.

IV. Application

Resource retrieval subsystem is responsible for analyzing user submits a request to index database, thesaurus and related thesaurus as the basis, to complete intelligent resource retrieval. The application of the Chinese information processing system of segmentation techniques, allows users to submit natural language queries, without having to be entered by keyword query. This friendly interactive mode enables the user to obtain a new, natural user experience. Additionally, the system is designed in accordance with the results of the query algorithms are sorted, making it easier for users to get the most needed resources. Resource retrieval subsystem features include:

- User described by natural language queries as word processing, to obtain keywords. This process, the system discards the statement auxiliary words, such as conjunctions, auxiliary verbs, pronouns, etc. For example, the user enters "I want on plant resources", the system obtained after the treatment in the word "plant", "resource" these two keywords.
- Relying thesaurus and related thesaurus, get the keyword synonyms and related words for subsequent data provide the basis for intelligent retrieval.
- Combining words and vocabulary extension, construct fuzzy query conditions.
- According query, retrieves from the index database qualified resources.
- Index calculated according to the weight value of the matching degree of the search resource to be sorted. The best resource information to meet user needs are arranged on top.

The system will be expanded synsets all packaged in a BooleanQuery object, where the concept of expansion during this time, allowing maintenance personnel to develop applications based on the actual expansion of the number of query words, by default, we are only the first query words for expansion.

If the user enters a query term is not defined in the thesaurus, the system also provides for the system maintenance personnel associated user-defined word thesaurus maintenance functions. When the query words also appear in the user-defined dictionary and thesaurus in relevant when associated with user-defined priority thesaurus expansion.

V. Conclusion

In this paper, domain-oriented information resource retrieval for clues, put forward for the field of information resource organization, resource management and service related technologies. The paper gives the system architecture, key technology and main features.

To further strengthen the scientific information resources organization and management, will be messy fragmented, not easy to readers to search queries, low utilization of information resources online through scientific organization and management of information resources structure consists of "information resources" to "knowledge system "transformation, which is the establishment of scientific classification, the level of knowledge and significant information resources to improve the efficiency and effective way.

References

[1] Apache Lucene. [Online]. Available: http://lucene.apache.org/

[2] C. Dong. The full-text retrieval function join in application—an introduction of java full-text searchengine. [Online]. Available: http://www.chedong.com/tech/lucene.html

[3] Apache Jakarta Commons Digester. [Online]. Available: http://jakarta.apache.org/commons/digester/

[4] L. M. de Campos, J. M. Fernandez, and J. F. Huete, "Building bayesian network-based informationretrieval systems," in *Proc. the 11th International Workshop on Database and Expert Systems Applications*, pp. 543-553.

[5] L. M. D. Campos, J. M. Fernandez, and J. F. Huete, "Improving the efficiency of the Bayesian network retrieval model by reducing relationships between terms," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 26, no. 3, pp. 101-116, 2002.

[6] W. R. Caid, S. T. Dimais, and S. I. Galiant, "Learned vector space models for document retrieval," *Information Processing and Management*, vol. 31, no. 3, pp. 419-429, 1995.

Shaochen Gao was born in March 1980. She has received her master degree and is currently a technical staff at Library of Shanghai University of International Business and Economic, with a focus on information process, library and information.