# The Construction of Heterogeneous Database Based on Web Data Grab

Yuhua Chen and Jinghai Yin

Institute of Information Technology, Jiangxi University of Technology, Nanchang 330098, China

**Keywords:** Network Data Capture; Item Resource Library; Regular Expressions

**Abstract.** This article describes the data in search results of crawl a crawl Web search engine to build the item repository. These methods make the efficiency of building Q improve and the costs reduce. Examing the questions based on XML storage format also have a very strong compatibility.

## Introduction

The test resources library adapts in large-scale development of the cause of examination and examination work with more scientific, standardized which needs developed. Strictly follow the theory of educational measurement, based on the accurate mathematical model and the educational measurement tool, which belongs to the field of computer-aided teaching (what??). Its basic unit is a single subject and the test database is a test warehouse. Problem management a good question bank can control the quality of the whole system, to achieve a variety of specialized functions, so as to effectively use questions for education and test. Test resource library construction is a complex systems engineering, first to establish the mathematical model of the system, then determine item attribute index and test structure, to organize a large number of outstanding teachers to write papers, in order to ensure the scientificalness and validity of these questions, which should organize a large number of test samples, sampling and testing, the item parameter label is effective for correction. The traditional method of test resources library to establish the need to consume a large amount of manpower and material resources,After decades of development, many fields have accumulated a lot of questions of resources, if these resources for the collection and processing through the network, as the material for the test resources library, the cost of the item bank construction can be greatly reduced cost, construction time can be shortened, ,it is also conducive to the inheritance of test resources and sharing. The network data capture using the technology basically used vertical search engine technology of web spider (or data acquisition robot), word segmentation system, tasks and index system technology of comprehensive and complete. Vertical search engine for a certain industry professional search engine, search object are usually specific website, and test resource on the network is basically evenly distributed in a large number of web servers, and the question involved in the field is various, content is very wide in the pan. The test resources on the network for data capture the full text search engine more suitable; can use the mainstream search massive search engine results were analyzed to obtain the related questions of resources.

## Basic Function Module

A lot of questions of resources must be through the rational organization and management can be easily used by the majority of teachers and students. Therefore, it is necessary to build the test resources library, resource library cannot only store a lot of questions, but also can store the and related test some important information, such as courses, chapters, types, difficulty, solution analysis, and so on, but also must provide quick query and statistical functions and convenient import and export functions. The main functions modules of the test library are shown in Fig. 1.

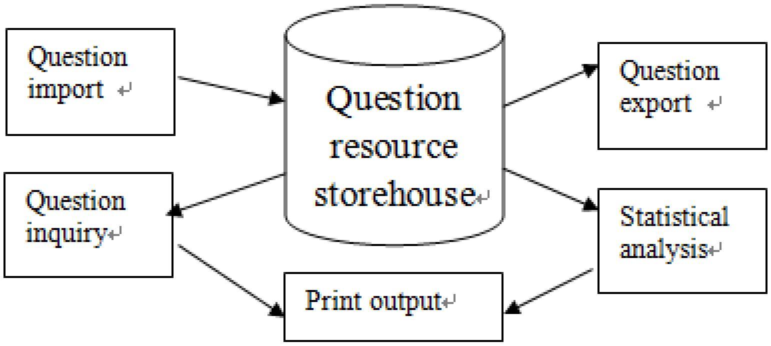

Fig. 1 Function module chart of test resource library

As Fig. 1 shows, test resource library through the test the import function building, also open the question query and statistical analysis function module, the output of the item bank is divided into two parts, the electronic version of the questions through the test module is derived to obtain, the paper version of the test by printing and output module to obtain. At the same time, the cross library transfer of the test resource can be realized through the import and export of the questions based on XML.

## System Data Flow

The source of the data source is the Internet resource, and the output is the format of the paper, and the map is the complete data flow chart of the resource library:

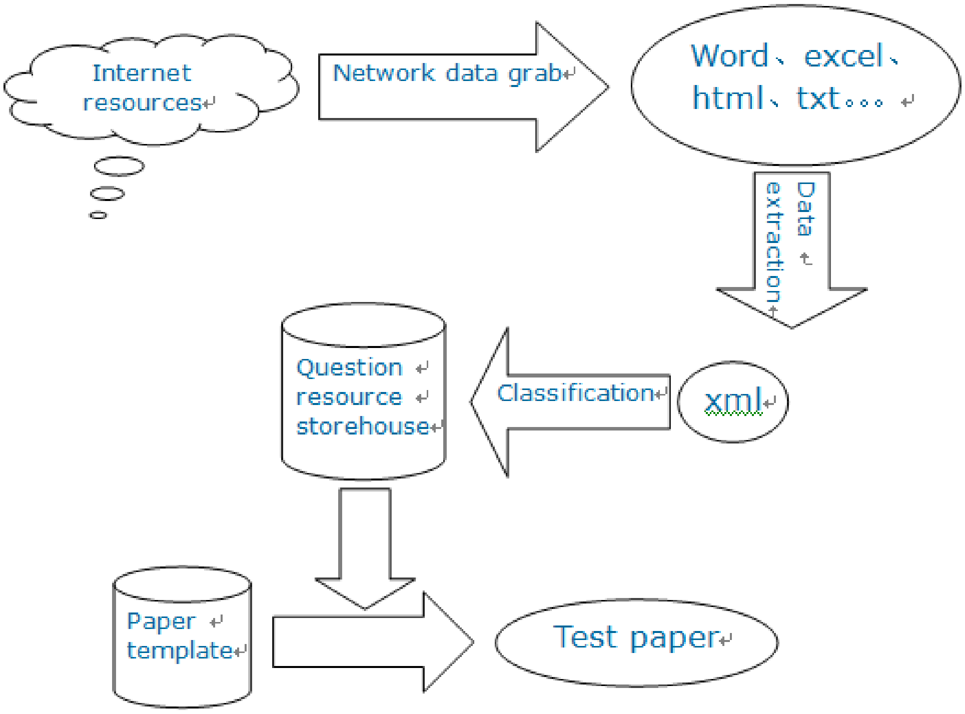

Fig. 2 Data flow chart of test resource database

As Fig. 2 shows, network data capture tools can get scattered in the test resource on the Internet, the full text search engine based on, through the analysis of processing the HTML text segments, the questions, such as word, Excel, HTML, TXT, PDF and other formats, can be got. These documents examination information by pattern matching and recognition will be effective information is extracted and transformed into XML format standard test resources. Through the import module of the question resource library, the questions in the XML format will be converted to the database mode to be stored and managed in uniform. If papers are needed to be output, load the corresponding paper format template to generate all kinds of papers with the requirements of the papers.

The data exchange of the system includes the data exchange of the heterogeneous database, the data exchange between the database and office, the data exchange between the database and the XML. The main exchange is shown as below, the core of data exchange is in ADO. Net dataset and through ADO. Net in the relevant functional classes will be with the database, word documents, EXCEL documents and XML documents for data exchange and XML document, and become a bridge for the exchange of data between heterogeneous databases.

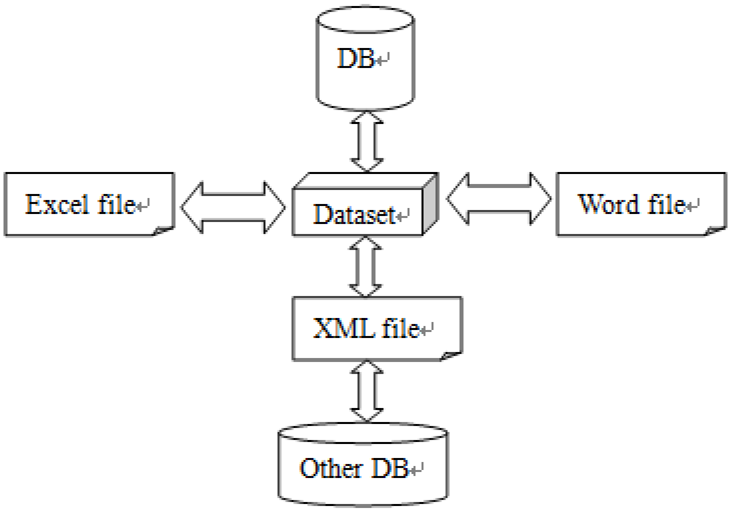

Fig. 3 Data type conversion chart

## Grab Test Resources

The traditional vertical search engine is on a site search depth, by recursively from the home page for your site to enter, by analyzing and matching each hyperlink to judge whether to visit the links page, if access, continue into the next layer of recursive, generally vertical search must specify the depth of the search, otherwise easily lead to the unlimited expansion of the network link and spread. Compared with this method, the search engine based full-text search method has obvious advantages in the following aspects. Firstly, the powerful search feature of the search engine can maximize the excavation of the relevant resources on the Internet above. Secondly, due to the optimized processing, the matching rate of the resources in the search results is much higher than that of the vertical engine. Finally, the resources to obtain high stability are wide, through the combination of the different key search elements, different types of material types can be got, instead of being confined to a particular class of test resources.

The crawl of network resources mainly has the following steps:

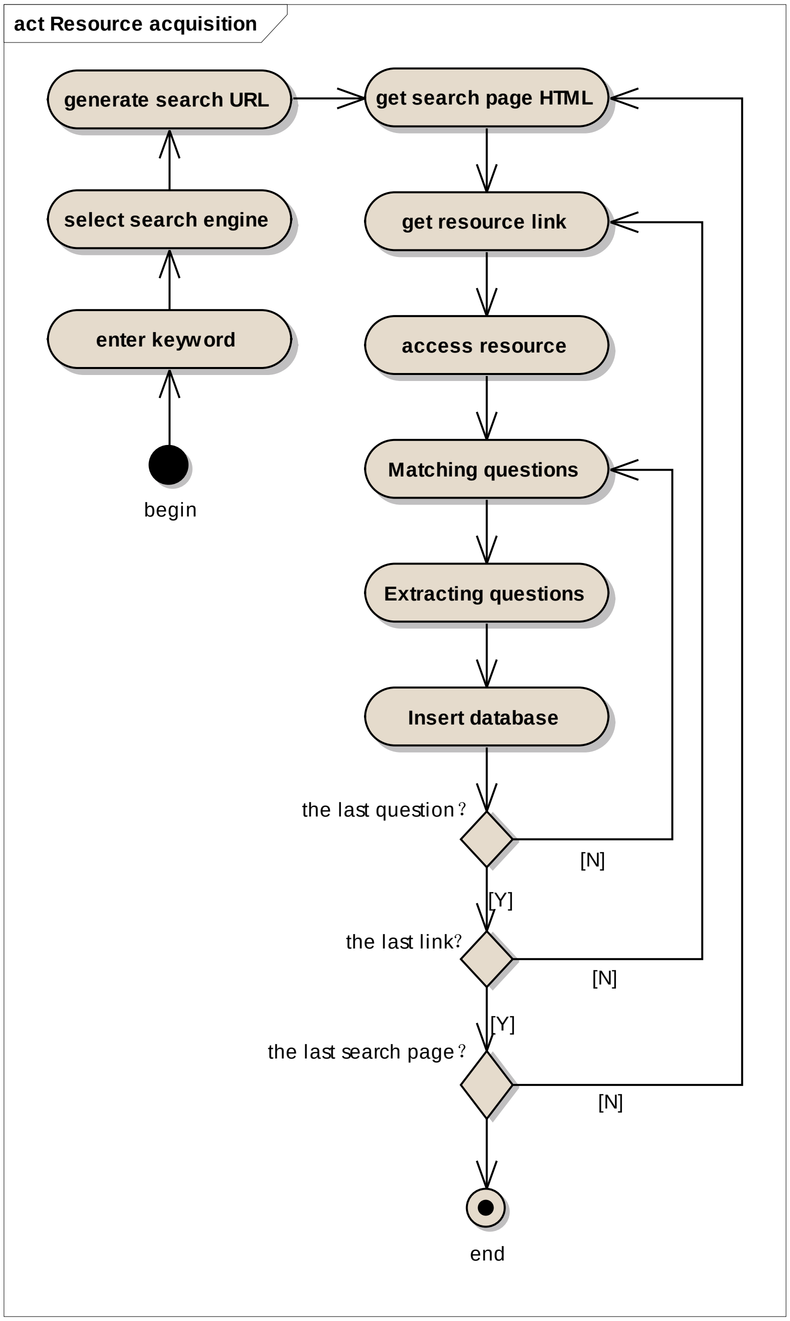

Fig. 4 System flow diagram

1. The user enter the keyword and select search engine to generate the search URL, access to the URL and get back to the search page HTML coding.

2. By matching the HTML coding, access to the resource link, and then get to the question resource information.

3. Access to search engine page turn button of the HTML code, automatic page turning is realized by software, and then continue to step operation until it reaches the preset range of pages or on the last page.

## Discussion

Through the network data digging getting in the way, to construct the database of test questions, to greatly accelerate the item bank construction speed, to reduce the item bank construction costs, based on XML standard item storage format also makes the test questions library cross bank transfer and merge operation becomes simple. However, this method has not automatically identified the function of multimedia examination questions or material, and will be added to the grab function of the multimedia resources in the following research.

## Acknowledgment

This work was supported by Jiangxi province department of science and technology support project [20142bbe50030] and Natural Sciences Project of Jiangxi Science and Technology Department [20122BAB201049]. The authors are grateful for the anonymous reviewers who made constructive comments.

## References

[1] JoSePh Fong，FraneisPang. Converting relational database into XML document.IEEE，2001.61-65

[2] Fawsy Bendeek.Automation of XML Document Translators Generation. IEEE，2001.37-38

[3]Rogers Gary , Moscato Pablo, Langston Michael. Graph algorithms for machine learning: a case-control study based on prostate cancer populations and high throughput transcriptomic data[J]. BMC Bioinformatics, 2010, 11.

[4] Sotiris Batsakis,Euripides G.M. Petrakis,Evangelos Milios. Improving the performance of focused web crawlers[J]. Data & Knowledge Engineering . 2009 (10)

[5] Peng Tao,He Fengling,Zuo Wanli. A new framework for focused Web crawling[J]. Wuhan University Journal of Natural Sciences . 2006 (5)

[6] G. Almpanidis,C. Kotropoulos,I. Pitas. Combining text and link analysis for focused crawling—An application for vertical search engines[J]. Information Systems . 2006 (6)

[7] Alberto H. F. Laender,Berthier A. Ribeiro-Neto,Altigran S. da Silva,Juliana S. Teixeira. A brief survey of web data extraction tools[J]. ACM SIGMOD Record . 2002 (2)

[8] Vladimir M. Krasnopolsky,Michael S. Fox-Rabinovitz. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction[J]. Neural Networks . 2006 (2)

[9] William E. Winkler. Methods for evaluating and creating data quality[J]. Information Systems . 2004 (7) [3] Monika R. Henzinger,Rajeev Motwani,Craig Silverstein. Challenges in web search engines[J]. ACM SIGIR Forum . 2002 (2)

[10] Peter W. Foltz,Sara Gilliam,Scott Kendall. Supporting Content-Based Feedback in On-Line Writing Evaluation with LSA[J]. Interactive Learning Environments . 2000 (2)

[11]Gorchetchnikov Anatoli, Versace Massimiliano, Ames Heather et al.. General form of learning algorithms for neuromorphic hardware implementation[J]. BMC Neuroscience, 2010, 11.