*I.J. Education and Management Engineering*, **2012, 11, 8-13**
Published Online November 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2012.11.02

***Available online at http://www.mecs-press.net/ijeme***

# The Exclusive Search Engine designed for personalized second language Chinese learner

Chen Lin, Xu Yan

*College of Information Sciences Beijing Language and Culture University Beijing 100083, China*

**Abstract**

With the Chinese learning craze on the increase, the huge lack of Chinese language teachers and the complication of the background of second language Chinese learner, it is becoming increasingly urgent to provide specific search of Chinese learning which can meet individual needs. The paper is proposed on the foundation of analysis of the characteristics of the special profession-second language learning. It focuses on three detail parts: extraction of individual information of learners, extraction of Web pages, acquisition of the data of Web database.

**Index Terms:** Second language acquisition; exclusive search; extraction of information; data mining



## 1. Introduction

With the Chinese learning craze on the increase and the huge lack of Chinese language teachers, it is really difficult to find some Chinese learning websites or materials which can faithfully meet the individual needs. With the internet resources multiplying and the diversified characteristics of various areas, even the general search engines such as Baidu or Google or other professional search engines at home and abroad do not satisfy the second language learners. The reality of the growing number of learners from different countries, ethics, regions, social classes or vocational backgrounds make the service of special search engine more urgent as learners want definite objectives and individual information services.

Taking advantage of the unique environmental strength and achievements of international education of Chinese, combining with natural language processing and data mining technology, we propose the individual search service platform based on personalized background. The background-based resolution can extract formatted background information to meet searching demands by using the information users provide such as a variety of documents, files, records on the Internet ect. In the searching process, both background information and what the users input will be considered to reflect the user's search intentions truthfully.

Sponsor:
supported by the National Natural Science Foundation of China (No. 60973148, 60873166), the Project of the Chinese Ministry of Education (No. 109028) , the Project of the Education science Foundation of Beijing ((No. AHA0911) and the Project of the Excellent Course of Beijing language and culture university (No. B201012)

* Corresponding author.
E-mail address: chenlin@blcu.edu.cn, xuy@blcu.edu.cn

In this paper, the characteristics and demand of the current second language acquisition will be analyzed, and then the details of the three parts of precision and efficiency provided to the second language learners will be talked about.

## 2. The characteristics and demand of the current second language learners.

As China is growing, the world's expectation of China's future development is becoming better and better. At the same time, the enthusiasm of foreigners learning Chinese is also warming up. According to the statistics of Ministry of Education [1], More than 2,500 universities in more than 100 countries around the world offer Chinese courses to students. The number of people learning Chinese around the world is over 40 million and will reach 100 million by 2010. Chinese government has been committed to the promotion and publicity of Chinese. By November 2009, 282 Confucius Institutes and 272 Confucius Classrooms have been launched in the whole world [2]. Chinese Proficiency Test (HSK) has been set more than 100 test centers in dozens of countries and regions overseas, known as the "Chinese TOEFL".

According to Chinese Institution that Chinese teachers and volunteers are far from meeting the global needs. A variety of Chinese learning and teaching sites related to IT technology are flourishing. In addition to government and schools, various organizations, companies and individuals actively jump into the field, some of which even use smart phones, PDA and other mobile device to download resources in order to learn on the move through the web site. [3] In addition to comprehensive website, as well as numerous distinctive web site for Chinese learning, such as http://you.video.sina.com.cn/ezychinese learning [4], but it is still difficult to find some that can meet individual needs. Because of the huge group of various individuals, no web site is good enough to meet the needs of everyone. Even in the teaching of more formal Confucius Institute classes, meeting individual need is still a problem, Even if in the same school, the same grade lesson plans are different from class to class. Requirement of Non-student learner is different as some need business Chinese, while others need travel Chinese. Therefore Chinese teaching is expected to offer more diversified and personalized services.[5],[6]and[7]

## 3. Resolution of individual specific search based on personalized background

With such a huge number of learners from different country and their different ethnic backgrounds， different religious and different cultural backgrounds, in addition to the second language learners' different professional backgrounds, how to meet the individual needs?

### *A. Internet Resources programs and search services*

The use of Mass network resources is a sound program, as the best and most efficient way to find the resources is to use search service. Search service refers to collecting information on the Internet based on certain strategies and the use of specific computer programs, providing search services for users after organization and treatment of information. [8].

Vertical search such as Industry search, shopping search and community search, emerge as new hot spots of development in recent years [9].Due to a deep understanding of specific areas, particular group or specific needs, vertical search, cannot simply be regarded as a general search of an industry or some area. Vertical search pays more attention to usability, and provides differentiated services. Now at home and abroad no specific search for the global Chinese network service is provided for this particular industry in the Chinese language learning, and no one is provided for the more characteristic area of second language learning, though is demands for exclusive search service by the second language learners is huge an urgent.

### *B. Resolutions for individual-specific search base on personalized background*

Using the unique environmental advantages and research achievements of international education of Chinese, combining with natural language processing and data mining technology, and special network environment in BLCU we propose the research and design of search service platform based on personalized special background

of. It provides specific search services for a huge number of globally distributed students from different countries, different ethnic backgrounds, different religion, different cultural backgrounds and different social levels, as well as for second language learners with different professional backgrounds.

*1) Extraction solution based on background*

Now it is difficult for the search engine to deal with the information from the user because the information is far from useful or accurate enough to express the objectives. Search designers always feel confused as they have to grasp what the users want from the limited and unclear information. This resolution based on background extraction is based on the information that user can provide such as a variety of documents, files, records on the Internet to extract formatted background information to meet the needs . Both background information and what the users input will be to reflect the user's search intentions truthfully.

*2) Construction design for client service program*

It puts the processing relevantly sensitive to users on the client, which not only balances the server to handle stress, but also solves the private problem most concerned to customers, as users' background is dynamically updated which includes the user's interest orientation, access to the Internet, input information, online Chinese language learning content, time on a certain page, and time spent on the Internet. Client service procedures as an independent program or pendant gives full control to clients, including open, close, uninstall or delete, which can help build confidence of customers, while avoiding the complexity and server resource consumption, reducing legal disputes.

*3) Consideration of different cultural backgrounds*

Extraction based on background search can also reduce or avoid the disputes aroused by different responds to sensitive information of clients of different nations, ethnics or religious beliefs because the extraction of information contained the background of states, nations, religions, therefore better shielding can be achieved.

## 4. Three key design for individual search service for second language learners of Chinese

### *A. Information Extraction of second language learners of Chinese*

Accurate, and complete perception of the needs of second language learners, is the foundation to achieve the service by the search management platform of teaching resources. Learners are both the network of information resources user and provider, so we can get the needs of learners by analyzing the learners' browse contents, behavior, and release of information. After we get their needs, we need to know how to filter relevant information among the vast database of information and how to show the learner in more special ways. That is the key issue to success. In addition, the learners usually have high demand for the timeliness of the information. As a result, how to protect the information retrieval system performance is also one of the important researches of acquiring teaching resource from the database information.

*1) Information gathering of Chinese language learners*.

Chinese language learners hope to search for their own learning strategies and learning content. However, the current major search engines try to grasp the intention of user from the limited information input which is not clear, as a result, such search results often fails to meet the requirements of Chinese language learners. Because the Chinese learner: "cannot describe what he is looking for, unless he finding what he wants." Based on this, we add the Chinese language information extraction program for Chinese language learners which can provide background information to individual learners when they search the information with inputting both background information and key words to truly reflect the user's intentions. We mainly achieve the extraction of personal information of Chinese language learners from the following aspects:

*a) Language learners learning initiative provided personal background material:* For example, language learners take the initiative to subscribe to certain types of learning content, leave their age, nationality, education background, religious background, learning objectives when the register. Also answer questions of search management platform of teaching resources and so on.

*b) Providing stand-alone client tracking software to learn the Chinese language:* It can collect the document and internet history of language learners have browsed and interacting with the server side. In order to reduce leakage of personal information of learners, we can calculate the amount of information of key words through the rate of appearance of the collected key words--information entropy-and set the threshold to control the exposure range of entropy.

*c) Interest documents update:* After sending language learners the information gathered from the Internet according to their interests, the server real-timely monitors the process of filtering, browsing, recording the last access time, and updates language learners' interest documents. In the update process, due to restrictions on the dimension key words, for every language learner profile document, the server has established a corresponding backup documentation, and set the size and number of key words and the upper limit to the number of occurrences. It will delete the key words with long absence and put most frequently updated keywords into the profile document, which adopts update algorithm of the LRU of cache to update profiles.

*2) Short-term interest model based on long associated content delivery.*

The disseminator of information on network resources and information visitors of network platform push technology using model-based personalized information: for information disseminator with the background of the current language learners' personalized information (age, nationality, religion, work, etc.) design for learners interest model of long or short term, and divide learners into different groups, recommend learning strategies to learners of similar interest (content); for language learners, it takes the visited pages as the representation of the learners. In this way, it not only pushes the theme of learning content and learner-related, but also with the web content potentially to maximize the extent consistent with the background of the learner which is more targeted. With the achievement of both short and long term interest of the modeling, by combining machine learning-based approach and decision-making of the Fuzzy Reasoning, BP neural network algorithm and Rough Sets attribute reduction, attract the interest of language learners through analysis of language learners access to records of web sites [10]. And then, integrate with the websites visited by the learners and find the change of interests to automatically update the interest-database.

## *B. Web page information extraction for Chinese learner*

Information extraction (IE) is the structural information processing from the text into an organized form as a table. Information point extracted from a variety of documents, and then integrated into a unified form. A large number of Chinese second language acquisition's information exists in the form of natural web document so that web-based form of information extraction has become a hot topic. There are a variety of different principle-based information extraction technologies, and they have different properties. Existing Web information extraction methods can be divided into four categories:

*1) Data extraction based on natural language processing.*

Early text of the document information extraction systems often rely on people manually creating the extraction rules, which is difficult to ensure a systematic and logical feature, and also with poor portability. Thus machine learning in the information extraction system to automatically extract information rules to deal with the massive web documents has become an urgent demand. Current machine learning methods widely used: Maximum Entropy method, hidden Markov model, Maximum Entropy method of hidden Markov model (MEMMs), Conditional Random method, kernel-based machine learning methods.

*2) Wrapper-based data extraction.*

Data through the wrapper can be extracted from the HTML page, and transformed into a structured software program. Generally speaking, a wrapper just can handle a specific source of information. Different type of web pages for data extraction need different wrapper. Wrapper-based information extraction has two research areas: Wrapper generation and Wrapper balance.

*3) Data extraction based on ontology.*

Construction of ontology is essentially a sense of organization or group decision-making behavior. Expert knowledge is contextual and constructed independently, relatively powerful but inevitably one-sided. It is difficult to build one that can meet the needs of all members of the noumenon. Even in some good framework, it is difficult to reach consensus. However, those more successful applied ontology are largely from the vast majority of experts which can reach consensus in terminology and concepts. In ontology evaluation, due to inability to obtain knowledge of all areas, many high-level of ontology applications could not be verified.

*4) HTML-Structure-based data extraction.*

Thanks to the structure characteristics of HTML page itself, HTML-structure-based information extraction method has a very strong automatic capability and high performance automatic data extraction ways. It parses a syntax tree before the document information extracted, with the automatic or semiautomatic extraction rules transform information extracted into the syntax tree operated, which is in order to achieve the information extraction.

*5) Improving on the ontology and natural language processing.*

One page often contains more than one kind of theme, the information distributed throughout this page. Therefore, This wants a more accurate access to Web information, dividing Web document into a certain number of data block is the primary task. Granular Computing Based on Maximum Entropy and Rough neural net algorithm in data classification and attribute reduction is efficient, with this method calibrating property of each block, classifying according to the approximation, reducing redundant attributes, training Reduction Model, optimizing identified data, completing the Chinese learning characteristic rules for the automatic extraction of the information. Second, with ontology approach, we intend to introduce Fuzzy Inference System into the ontology framework. Based on Fuzzy Inference method, CRI constructs expert knowledge into an expert knowledge base, providing a quantitative framework for ontology-based data extraction, and then constructs a multi-ontology layer of decision-making system, which provides decision support feedback for Chinese learners' search. Rough Set-based fuzzy cluster the search results to establish the type of users, enabling different users get different search results. The versatile method can cluster large data with numerical value attribute and symbol attribute, improving the efficiency and quality of data mining.

### *C. Data mining of Chinese Learning Web-Database*

With the collection of Second language learners' personal preferred information, Modeling and establishing a Chinese learning Web-database. Information extracting technology on the Web page builds a Chinese acquisition structured database. It provides differentiated learning support for learners mainly depend on mining Web information.

*1) Repeat identification of the website reprint and the similar resources.*

Chinese teaching resources copying result in a lot of duplicated content on the Internet, to take a great waste of time for data analyzing, and increase the burden of the server. So we add the identification when we deal with these resources. We will clean page content noise and segmentation before identifying, meanwhile leaving the article's signal. During the segmentation remove the script code and stop words. After segmentation, compute the frequency of each word and chose three or five highest ones to be key words. Supposing "you/he/I "are the three keywords, respectively 10/8/5 times, then forming a string "you 10 he 8 I 5".Using MD5 or other code forming a string numbers called information fingerprint for this article, and search this fingerprint in the database, if exists, this page maybe duplicate. It must exist many holes if only duplication identifying, such as some deceive behaviors with the core key words slightly changed, or artificially added or removed. Therefore, the search engine need modify this algorithm, like search the key words in frequency from 6 to 9, or even take sub-multiple identification.

*2) Clustering the Chinese learning resources for different background and learning purpose.*

Using Fuzzy-Cluster Analysis analyze the search results. Clustering is able to scale the data according to the similarity of the data without any prior knowledge of the situation, and divided it into clusters for users' quickly

understand of the elements contained in the document. Cluster Technology is an indispensable part of the Search Engines research field. Thanks to the magnitude of variable quality of network data, we use search engine filters and cleans out the Web-data that is in low-quality, then clusters the search results. And the construct of a smart search engine through fuzzy clustering will cluster the search results in different user-type.

*3) Semantics-based represented spatial model of Fuzzy factor.*

In allusion to Web-resources for Chinese teaching proposing a semantics-based represented spatial model of Fuzzy factor, the basic idea is to use ontology and semantic-linked-network to achieve a unified and standardized framework from the disordered heterogeneous resources, which can clearly reflect the semantic structure between the learning resources. Define a two-dimensional factor space, as the object and its attributes. Object space is the abstraction of the heterogeneous resources, location and format of the resources and some other details are transparent. This space can be regarded as an object of a single semantic reflecting the description logic-based formal model including knowledge category and property database. Knowledge category model is used to describe areas' concept, attributes gallery is the abstract of areas' resources. This model is characterized by a semantic-linked-network representing the same types of semantic relations between individuals, and gives the semantic-linked-network's design criteria and constraints, rules of inference and evolution of the operation.

## 5. Summery

With the Chinese learning craze on the increase, the huge lack of Chinese language teachers, and the complication of the background of Chinese learner, internet and distant education has become the trend of development of Chinese teaching and learning. The difference in nationalities, ethics, regions, cultures, classes and vocations of Chinese learners lead to the difference of learning objectives and individual-suitable information. However the general or specific search engines cannot meet the demands. The individual background-based specific search can perfectly solve the problems. Compared with other search services, this service has its own unique strength and will surly support the teaching and learning of Chinese.

## References

[1] Global Chinese learning craze net of Chinese education http://News.wenzhouglasses.com 2010.9.15
[2] Everson,M.E.& Ke Chuanren. An inquiry into the reading strategies of intermediate and advanced learners of Chinese as a foreign language. Journal of the Chinese Language Teachers Association, 1997.32(1), 1-20.
[3] Gu,Yongqi. & R.K.Johnson. Vocabulary Learning Strategies and Language Learning Outcomes. Language Learning 1996.46(4):643-79..
[4] Chen, G. M. & Starosta, W. J. Foundations of intercultural communication. Boston. Boston Press, 2007.
[5] Berry, Kim, Unichol. & Boski, P. Psychologycal acculturation of immigrants. In Kim,Y. Y., & Gudykunst, W. B. (Eds). Cross-cultural adaptation: Current approaches. Newbury Park: Sage, 1987.
[6] Ward, C. Thinking outside the Berry boxes:New perspectives on identity, acculturation and in-tercultural relations. International Journal of In-tercultural Relations, 2008.
[7] Dodd, C. H. Dynamics of international communication. (5th ed). Shanghai: Shanghai Foreign Language Education Press, 2006, 96-99.
[8] Xu Yabo, Wang Ke, Zhang Benyu, et al. Privacy-enhancing personalized web search. Proceedings of the 16th International Conference on World Wide Web, 2007.
[9] Singhal A. Modern information retrieval: A brief overview.Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001, 24(4):35-43.
[10] Dou Zhicheng, Song Ruihua & Wen Ji-Rong.A large-scale evalua-tion and analysis of personalized search strategies. Proceedings of the 16th International Conference on World Wide Web, 2007.