International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:6, 2010

OCIRS: An Ontology-based Chinese Idioms Retrieval System

Hu Haibo, Tu Chunmei, Fu Chunlei, Fu Li, Mao Fan and Ma Yuan

***Abstract*—**Chinese Idioms are a type of traditional Chinese idiomatic expressions with specific meanings and stereotypes structure which are widely used in classical Chinese and are still common in vernacular written and spoken Chinese today. Currently, Chinese Idioms are retrieved in glossary with key character or key word in morphology or pronunciation index that can not meet the need of searching semantically. OCIRS is proposed to search the desired idiom in the case of users only knowing its meaning without any key character or key word. The user's request in a sentence or phrase will be grammatically analyzed in advance by word segmentation, key word extraction and semantic similarity computation, thus can be mapped to the idiom domain ontology which is constructed to provide ample semantic relations and to facilitate description logics-based reasoning for idiom retrieval. The experimental evaluation shows that OCIRS realizes the function of searching idioms via semantics, obtaining preliminary achievement as requested by the users.

***Keywords*—**Chinese idiom, idiom retrieval, semantic searching, ontology, semantics similarity.

I. Introduction

CHINESE Idioms are a type of traditional Chinese idiomatic expressions with specific meanings and stereotypes structure which are widely used in classical Chinese and are still common in vernacular written and spoken Chinese today. They are mostly derives from ancient Chinese such as the mythology, fable, historical story, poetry, colloquial language and so on [1]. According to the most stringent definition [2], there are about 5,000 idioms used in the Chinese language, though some dictionaries list over 20,000. They are connotation intensified, brief, easy to be remembered and used, and usually with emotional attachment. If it is not well understood and utilized, the understanding and usage of Chinese will be greatly affected. Moreover, this kind of idiomatic expressions not only exists in Chinese literary expression, but also be widely used in other oriental languages such as the Japanese yojijukugo [3] and the Korean idioms [4].

Chinese idioms can be indexed by means of morphology, phonetics, with the spelling and pronunciation of idioms respectively. The paraphrase and the source of the idiom can be searched through its spelling and pronunciation currently. However, it's very in common that a person would like to express a meaning in her / his literature with an idiom but can not s/he come up with it, or does s/he know any key word or character of the idiom on demand. Existing idiom retrieval methods facilitated with morphology and/or phonetics indexes can not meet the need of such cases due to lacking of methodologies for indexing them semantically. The problem lies in three aspects, that it is difficult for users to faithfully express their search request with one or more key words, the search algorithm does not adopt the semantics matching but the morphology matching so that the search results include lots of ineffective information; and it is difficult to describe the semantic relations among the concepts [5]-[7].

Ontology-based Chinese Idioms Retrieval System (OCIRS) proposed in this paper aims to retrieval Chinese idioms semantically. Ontology is adopted since it is a methodology for concept modeling to present ample and complicated semantic relations among concepts. The semantic relations of the idioms will be analyzed in advance. Therefore, the abundant semantic relations among idioms can be expressed by constructing the idiom domain ontology. Sequentially, the connotation of the search information input by end users will be understood via grammatical and semantics analysis that lead to semantics similarity computation with *Synonym Lexicon*. By mapping the connotation of the user search request to the idiom ontology, the idiom set satisfying the user request can be retrieved with ontology reasoning.

The remainder of this paper is organized as follows. In section II, we give a brief overview of research efforts on Chinese idioms retrieval methods, semantics-based vocabulary, and semantic technologies adopt in this paper as well. Section III describes the framework of OCIRS and introduces the processes of how to retrieval idioms via grammatical analysis, semantics analysis, concept matching and reasoning. In section IV, the construction and development details of OCIRS are provided, and evaluations of OCIRS with comparison to other Chinese idioms retrieval systems are also presented at the end of this section. Conclusions of the paper and our future works are mentioned finally.

Hu Haibo is with the School of Software Engineering, Chongqing University, Chongqing 400030 P.R. China (phone: 8623-65111025; fax: 8623-65111025; e-mail: hbhu@cqu.edu.cn).

Tu Chunmei is with the Chongqing Career College of Information Technology, Chongqing 404000 P.R. China (e-mail: cm.tu@ cqeec.com).

Fu Chunlei is with the Centre for Information and Network of Chongqing University, Chongqing 400030 P.R. China (e-mail: fcl@cqu.edu.cn).

Fu Li is with the School of Software Engineering, Chongqing University, Chongqing 400030 P.R. China (e-mail: fuli@cqu.edu.cn).

Mao Fan is with the School of Software Engineering, Chongqing University, Chongqing 400030 P.R. China (e-mail: f.mao@cqu.edu.cn).

Ma Yuan is with the School of Software Engineering, Chongqing University, Chongqing 400030 P.R. China (e-mail: y.ma@cqu.edu.cn).

II. Related works and Methodologies

The researches on idiom focus on its semantics, source, usage, rhetoric and structures, and search for idiom is the indispensable part of studying, using and researching the idioms. Research on idiom retrieval and searching requests mainly include searching the pronunciation, meaning and source of the idiom through its spelling, or searching the spelling, meaning and source of the idiom through its pronunciation, or searching the spelling, pronunciation and source of the idiom through its meaning. The retrieval technology based on morphology and phonetics has been well-developed, but the current searching technology is still hard to meet the search of idiom itself through its meaning.

In Chinese lexical semantic research, *Synonymy Lexicon* [8], HowNet, CSD, etc. are renowned. As far as the Chinese vocabulary search is concerned, although the synonym of the Chinese word can be searched through *Synonymy Lexicon*, the *Synonymy Lexicon* is far from the user's request of searching the word through its meaning. The Chinese idiom is the significant part of Chinese vocabulary, but it has its own characteristics in a specific structure and stereotype. Research on semantic-based idioms retrieval and searching is so unsubstantial that a lot of researches are pursuing.

Since Tim Berners-Lee put forward the concept of Semantic Web in 2000, the semantic search on the web environment has been the mainstream of network research development, with the research on word semantics obtaining a certain achievement and different types of vocabulary knowledge-bases formed, among which, the WordNet [9] of Princeton University in the Untied States is the most influential one, which is an online English vocabulary searching system. The WordNet realizes the searching request of the matching English word through its meaning, but its limitation lies in that the user can only search with specific word or phrase, and the WordNet itself cannot analyze the connotation and show the search result when the user gives search request with sentences or combination of more than one word.

In 1993, the most popular definition of ontology was the one given by Gruber that "an ontology is the specification of a concept model" [10]. The target of ontology is to define the vocabulary in a certain domain or the relationship between words to describe or indicate the knowledge in the domain and facilitate the automatic process to data. Under the processing of data with the computer, the ontology needs indicating formally in a language so that the clear meaning, nature and reasoning algorithm can be provided. At present, in the semantic web community, many semantic markup languages can be used for the description and indication of the ontology, including RDF, RDFS, OWL, etc.

Description Logics [11], also called Terminology Logic, is the unification of the logic reconstruction and formalization of the knowledge as the tool in its early term. It uses frame system, semantic network and object-oriented method to work as tools and is equipped with the semantics suitable for formalization. DLs are complex of concept formulae to describe and infer the concept knowledge, being capable of providing the decidable reasoning function and serving as the unified logic foundation of such knowledge as the semantic network and frame as tools. DLs are also the logic foundation of the ontology in the semantic web. The logic foundation as well as the core of the ontology language OWL-DL is the DL.

Protégé is a free, well known open source ontology editor and knowledge-base construction tool. The Protégé platform supports two main ways of modeling ontologies via the Protégé-Frames and Protégé-OWL editors. Protégé ontologies can be exported into a variety of formats including RDF(S), OWL, and XML Schema.

The Protégé-OWL API is an open-source Java library for the OWL and RDF(S). The API provides classes and methods to load and save OWL files, to query and manipulate OWL data models, and to perform reasoning based on Description Logic engines.

In this paper, Protégé will be used as the ontology editor to construct the ontology and the Protégé-OWL API to operate the ontology.

III. The Roadmap of OCIRS

A. Framework of OCIRS

The semantics-based idiom searching model of OCIRS is shown in Fig. 1. When users providing their search request in the natural language as a Chinese phrase or sentence, this model will analyze the request grammatically and then analyze the result semantically to get the search target.

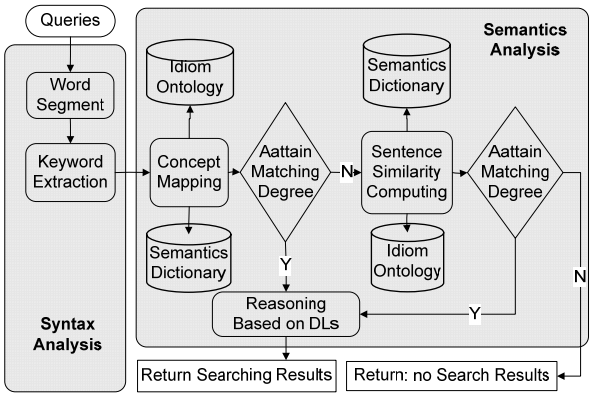


Fig. 1 Semantics-based idiom search model

In grammatical analysis, the word segmentation aims to separate the words input by the user and to mark the parts of speech. The key word extracting module analyzes the syntax of the whole sentence, extracts the key words and works out the matching key words set.

During the semantic analysis, the concept mapping will map the key words from the grammatical analysis to the ontology concept. The sentence semantics similarity computation module will practice the concept auxiliary mappings through computing the semantics similarity of the key word set and the individual in the idiom domain ontology when the anticipated matching of the key word set and the concept mapping of the

ontology is not achieved. The semantic dictionary is the dictionary resource for the synonym conversion and semantics similarity computation during the concept mapping. The description logics based reasoning module infers the semantic relations among the ontology concepts with the inference machine so as to obtain the matching search target.

B. Grammatical Analysis

1) Word segmentation

Word is the minimum independent meaningful language element. Firstly, OCIRS will segment the word of the natural language input by the user, which is to segment the search request input in the natural language as a phrase or sentence into several words and mark the part of speech of each word. The ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)[12] which was developed by the Institute of Computing Technology of Chinese Academy of Sciences is used to segment the words and mark the part of speech of each word in this paper. For example, if the user inputs the words like "品格高尚的人" (person with noble characters), a key word set of "品格/n高尚/a的/u人/n" (person/n, with/p, noble/a, characters/n,) with the part of speech of each word can be received via the processing of ICTCLAS.

2) Key Word Extraction

According to Chinese linguistics, any sentence consists of the key elements (such as subject, predicate, object, etc.) and the modifiers (such as attribute, adverbial modifier, complement, etc.). The key element plays a primary role in the sentence and the modifier, the secondary. Therefore, it is unnecessary to keep every segmented word and only the key element of the sentence can be considered for the key word extraction. This system regards nouns, verbs, adjectives and idioms in the sentence as the key words and extracts the key words after the word segmentation, thus a matching key word set can be obtained. The key words extracted in this way boast a certain syntax information expressive ability and can primarily reflect the core idea of the sentence.

C. Construction of Idiom Ontology

Idiom ontology is the important part of the semantics-based idiom search and the key to the semantics analysis of the system. There are lots of ways to construct the ontology, including the Skeletal Methodology [13], TOVE [14], METHONTOLOGY [15], etc. In reference to these ontology idiom construction methods and combination of the reality of the idiom domain, the ontology editor Protégé4.0 by Stanford University will be used to construct the idiom ontology.

The deep analysis and conclusion will be exercised in a certain scope of the idiom domain. The analysis conclusion will be made to the daily idioms which will be classified as the description of people, events and matters [16]-[17]. The analysis will be practiced based on the characteristics described by the idioms to the human beings, events and matters, and the concept in the idiom ontology can be formed after the standard definition and description of these characteristics. For example, human beings have the characteristics of their character, body and feeling, and the character can be classified as the nobility, evil, loyalty, etc. The partial concept of the idiom domain ontology is shown in Table I.

TABLE I
PARTIAL CONCEPTS OF THE IDIOM DOMAIN ONTOLOGY

Class Name	Description
Idiom	Idiom
DescriptionPeople	Idioms described people
DescriptionEvent	Idioms described event
DescriptionThing	Idioms described thing
IdiomTrait	Idioms described the characteristics of the object
PeopleTrait	The characteristics of people
EventTrait	The characteristics of event
ThingTrait	The characteristics of thing
Character	People have the characteristics of character
NobleCharacter	People have the characteristics of noble character
WillingHelp	Person have the morality of willing to help others
CharacteredPeople	Idioms described the character of people
NobleCharPeople	Idioms described the noble character of people
WillingHelpPeope	Idioms described people willing to help others

Only the class itself can not provide enough information to answer the question by the user during the semantics-based idiom search, so the internal structure among the concept must be described besides the class definition. For example, the idiom itself owns the characteristics of phonetics and definition, and some idioms have their homoionym or antonym. These terms to describe the internal structure among concepts will serve as the attribute of every concept. The relevant attribute description in the idiom ontology is shown in Table II.

TABLE II
PARTIAL ATTRIBUTES OF THE IDIOM DOMAIN ONTOLOGY

Attribute Name	Domain	Range	Description
hasAntonym	Idiom	Idiom	Idioms have Antonyms
hasSynonym	Idiom	Idiom	Idioms have Synonyms
hasIdiom	IdiomTrait	Idiom	Class IdiomTrait as to the idiom
hasTrait	Idiom	IdiomTrait	Idiom as the IdiomTrait
hasParaphrase	Idiom	XML:Literal	Idioms interpretation
hasName	Idiom	XML:Literal	Idioms name
hasPronunciation	Idiom	XML:Literal	Idioms pronunciation
hasLiterary	Idiom	XML:Literal	Idioms origin
hasContent	IdiomTrait	XML:Literal	IdiomTrait of contents

After the construction of matching concept and attributes of the idiom ontology, the relevant limits should be built for the class attribute to express more semantic relations among concepts. All the idioms with characteristics of describing WillingHelp are the WillingHelpPeope idioms. These attribute limit conditions can be described with DLs as follows:

WillingHelp⊑IdiomTrait⊓∀ hasTrait.WillingHelpPeople

The relation network of the idiom domain ontology is shown in Fig. 2. Because of the relatively large scale of the ontology, only a part of the idiom ontology is shown in Fig. 2. The relevant concept and attribute description can be referred to in Table I and Table II.

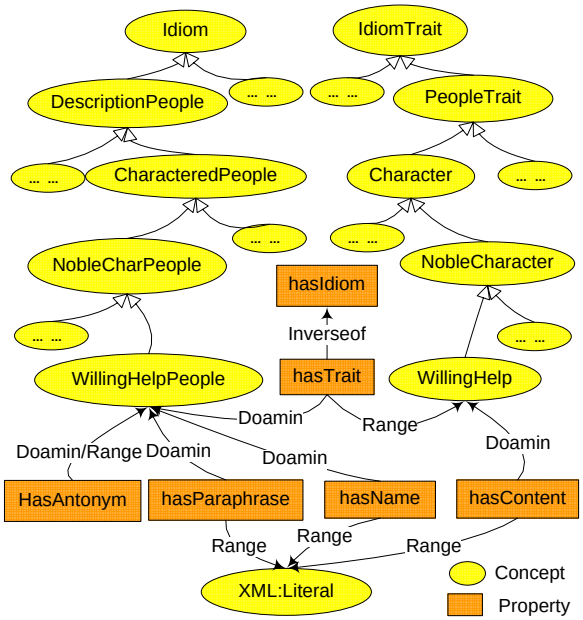


Fig. 2 Partial figure of the idiom domain ontology

D. Concept Mapping

Concept mapping is to map the key words collected after grammar analysis into the concept of idiom domain ontology. The idiom domain ontology describes the semantic relations among the concepts in idiom domain. By mapping the key words into the concept of ontology, the semantic relations will be established between the key words. The phenomena of one word with several meanings and one meaning described by several words usually occur during the concept mapping of key words. Therefore, it is necessary to refer to the semantic dictionary for help. Thus the expanded *Synonym Lexicon* by the Research Center for Information Retrieval of Harbin Institute of Technology (HIT-CIR) is employed in our work as semantic dictionary to facilitate concept modeling of Chinese idioms.

1) Synonym Lexicon

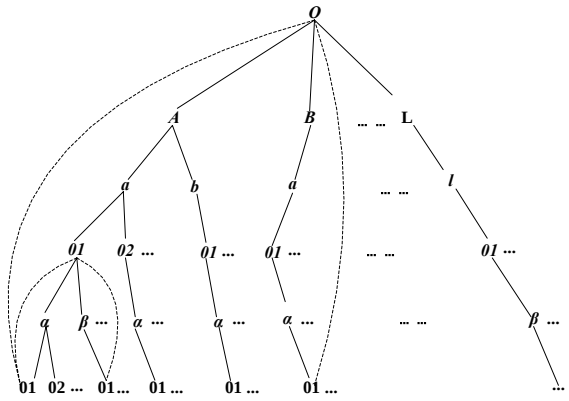
Synonym Lexicon is a common synonymous thesaurus for modern Chinese. With *Synonym Lexicon*, the Chinese words can be classified into a tree structure, including large, medium and small classes. Each small class contains various words, and the words are classified into many word groups according to meaning. In sequence, the word groups are further classified into synonymous groups. The word groups are actually at the fourth level in the tree structure while the synonymous groups are at the fifth level or the end level, which can be called as the atom groups. In short, the words of *Synonym Lexicon* are coded in the five levels based on relation of the words meaning, seen in Table III.

TABLE III
CODING OF WORDS IN *SYNONYM LEXICON*

Sequence	Symbol Example	Symbol Properties	Level
1	*A, B,..L*	Large Classes	First Level
2	*a, b,…l*	Medium Classes	Second Level
3	*0*	Small Classes	Third Level
4	*1,2*		
5	*α ,β*	Word Groups	Fourth Level
6	0	Atom Word Groups	Fifth Level
7	1, 2		

2) Synonym Lexicon based Semantic Similarity Computation

The *Synonym Lexicon* based semantic similarity computation method was introduced by B. Wang in 1999[18]. A root node *O* is assumed for the existing five levels of *Synonym Lexicon* and the overall classification based on words meaning which can be described by a tree diagram as shown in Fig.3.



Notes: the route from root/middle node to leaf node is marked by dashed line.
Fig. 3 Tree diagram of semantic classification of *Synonym Lexicon*

As shown in Fig. 3, the symbol at each node represents the corresponding level of words classification as large class, medium class, small class, word group and atom group. The shortest route from the root node *O* to the leaf node represents one meaning, and the meaning code is a character string containing all the codes on the shortest route (excluding *O*). Then, the semantics distance of two words S_1, and S_2, as $SenseDist(S_1, S_2)$, can be defined as the shortest length between node S_1 and S_2 as shown in the tree diagram.

For example:

$$SenseDist(Aa01\alpha01, Aa01\alpha02) = 2$$
$$SenseDist(Aa01\alpha01, Aa01\beta01) = 4$$
$$SenseDist(Aa01\alpha01, Ba01\alpha01) = 10$$

Obviously, the shorter the distance between S_1 and S_2, the closer the meaning between S_1 and S_2, so the semantic similarity between S_1 and S_2 can be defined as:

$$SenseSim(S_1,S_2)=\begin{cases}1/SeneDist(S_1,S_2) & S_1\neq S_2\\ 1 & S_1=S_2\end{cases} \quad (1)$$

And the semantic similarity of two Chinese words (C_i, C_j) can be defined as:

$$ClassSim(C_i,C_j)=SenseSim(S_m,S_n) \quad (2)$$

$$S_m \in Senseof(C_i)\bigcap Senseof(C_j) \quad (3)$$

Function as $SenseOf(S)$ returns to the meaning code collection of word *S*.

3) Concept Mapping Algorithm

The top layer concept in idiom ontology includes Idiom and IdiomTrait. During concept mapping, the search request input by user can be mapped to the concept under IdiomTrait and then be reasoned by the inference machine, so the corresponding idiom can be reasoned out. The specific concept mapping algorithm is as follows.

After the grammatical analysis on the search request input by the user, the key word aggregate containing *m* key words is $S_1=\{k_{11},k_{12},...,k_{1m}\}$.

Since the concept in the ontology is arranged in a tree structure, the key word aggregate will be mapped to the sub-tree, which makes IdiomTrait as the father node. There are three sub-trees under IdiomTrait, including PeopleTrait, ThingTrait and EventTrait. The key word aggregate containing the key words from sub-tree node to leaf node will be formed. For example, {people, quality, high, very high morality}. Since there are four levels from each sub-tree node of people, thing and event to its leaf node, the key word aggregate of the concept in the ontology can be: $S_2=\{k_{21},k_{22},k_{23},k_{24}\}$.

$Sim(k_{1i},k_{2j})$ is the semantic similarity from word k_{1i} to k_{2j}.

$SimW_i=\max\{Sim(k_{1i},k_{21}),Sim(k_{1i},k_{22}),Sim(k_{1i},k_{23}),Sim(k_{1i},k_{24})\}$

The semantic similarity between S_1 and S_2 is:

$$Sim(S_1,S_2)=\frac{SimW_1+SimW_2+...+SimW_m}{m} \quad (4)$$

By making the similarity computation between each key word in S_1 and the concept in the ontology respectively with the same method, the concept reaching the predetermined match degree will be the result of concept mapping. If no word reaches the predetermined match degree, the auxiliary mapping that calculates on the basis of sentence similarity will be adopted. The predetermined match degree is based on experiment.

E. Formal Description of Semantic Similarity Computation

In Chinese, various relations are expressed through words correlated through word meanings, and sentence meanings are expressed centering on the meanings of characters and words, the computation of semantic similarity between sentences can be made through the computation of lexical semantic similarity.

The specific semantic similarity computation method of sentence is as follows:

Step 1. Segmenting the sentence and collecting key words.

Segmenting the search request input by user and marking word property through ICTCLAS. Then the key word aggregate containing the key words of noun (N), verb (V), adjective (A) and idiom (I) will be formed.

After the segmentation, the sentence *S* contains *k* words, $S=\{w_1,w_2,...,w_k\}$

According to the property of the words, the words in the sentence *S* can be aggregated into four sets, $S=\{N,V,S,I\}$

Then, add the words into their own group as per property, $S=\{N(w_i,...),V(w_j,...),A(w_k,...),I(w_l,...)\}$

Step 2. Calculate the semantic similarity between the words.

The semantic similarity computation method for words based on *Synonym Lexicon* put forward by Wang Bin will be adopted in this paper. Assume $\{N_1,V_1,A_1,I_1\}$ and $\{N_2,V_2,A_2,I_2\}$ as the key word aggregates of s_1 and s_1, and take the similarity computation between N_1 and N_2 as an example, $N_1=(w_{11},w_{12},...,w_{1m})$, $N_2=(w_{21},w_{22},...,w_{2n})$.

$SimW(w_{1i},w_{2j})$ is the semantic similarity between the word w_{1i} and w_{1i}, and refers to previous specific computation method in subsection III.D.

Step 3. Calculate the semantic similarity between the sentences [19]-[20].

In the same way, assume $\{N_1,V_1,A_1,I_1\}$ and $\{N_2,V_2,A_2,I_2\}$ as the key word aggregate of s_1 and s_1, and take the noun aggregate *N* as an example to discuss the semantic similarity computation between the sentences, $N_1=(w_{11},w_{12},...,w_{1m})$, $N_2=(w_{21},w_{22},...,w_{2n})$.

Assume N_{12} as the characteristic matrix of the semantic similarity between s_1 and s_2.

$$N_{12}=N_1\times N_2^T=\begin{pmatrix} w_{11}w_{21} & ... & w_{1m}w_{21}\\ ... & ... & ...\\ w_{11}w_{2n} & ... & w_{1m}w_{2n}\end{pmatrix} \quad (5)$$

where $w_{1i}w_{2i}=simW(w_{1i},w_{2j})$

Traverse the characteristic matrix of similarity, and then take out the maximum value of each line and row in the characteristic matrix. For example, make a_i as line *i* and b_i as row *i*, then:

$$a_i=\max(w_{11}w_{2i},w_{12}w_{2i},w_{13}w_{2i},...,w_{1m}w_{2i}) \quad (6)$$

$$b_i=\max(w_{1i}w_{21},w_{1i}w_{22},w_{1i}w_{23},...,w_{1i},w_{2n}) \quad (7)$$

And the semantic similarity between the noun aggregates of sentence s_1 and s_2 is:

$$simS_1(s_1,s_2)=\left(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{i=1}^{n}b_i}{n}\right)/2 \quad (8)$$

In the same way, the semantic similarity between the verb aggregates, adjective aggregates, etc. can also be calculated. As a result, the semantic similarity of sentences s_1 and s_2 as $simS(s_1, s_2)$ based on the weighted average of aggregate similarity is shown as follows:

$$simS(s_1, s_2) = \sum_{i=1}^{4} \beta_i simS_i(s_1, s_2) \quad (9)$$

Where β_i is the weight coefficient based on linguistics and experiment.

IV. System Design and Development

A. System Design

The OCIRS in this paper adopts B/S structure and the user can search through browser directly. As shown in Fig. 4, the system contains 3 levels, including user interface level, business logic level and data level.

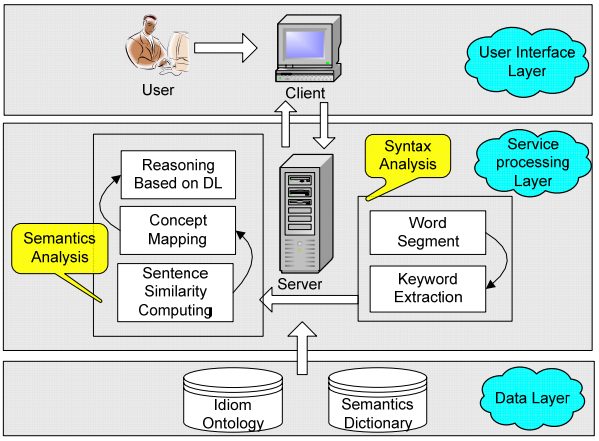


Fig. 4 Architecture of OCIRS

User interface level provides an interface for the end-users to access OCIRS and the user can also access the system through Web browser. *Services processing level* consists of the JSP and JavaBean components on Web server. It can respond to the HTTP request from the user and transmit the request data to the applied logic level and then transmit the processing result back to the user. *Services processing level* mainly consists of the programs of Servlet and Java, responsible for the semantics-based idiom search, and operates the idiom ontology according to the request of user. *Data level includes* semantic dictionary and idiom domain ontology.

B. Concrete Realization of Each Key Module

1) Grammatical Analysis

The grammatical analysis mainly includes sentence segmentation and key word collection. The sentence segmentation is available for us since ICTCLAS provides source code in C/C++, C#, and Java as well. The sentence segmentation and word property marking can be realized with method provided in the source code.

2) Idiom Domain Ontology

Protégé3.4 is employed as the editing tool for ontology in this paper. The conceptual level structure for ontology is shown in Fig. 5 as follow.

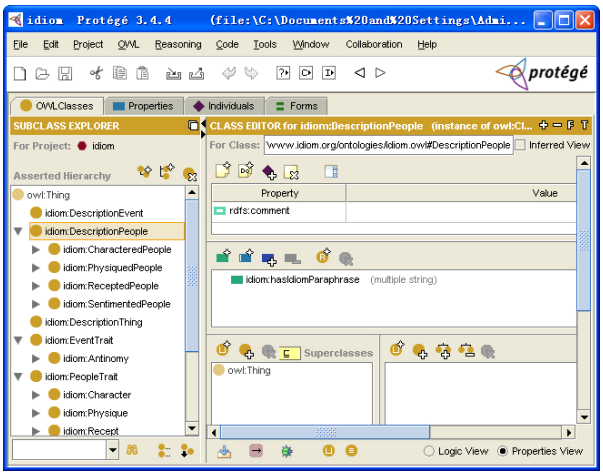


Fig. 5 Conceptual level system of idiom domain ontology in Protégé

3) Conceptual Mapping

Conceptual mapping module is the semantic mapping between the concepts of key word aggregate and that of the idiom ontology, during which ontology operation is required and Protégé-OWL API is used for this request in this paper.

4) Reasoning based-on Description Logics

The advantage of ontology reasoning function distinguishes Protégé-OWL API from other knowledge organization systems. In this paper, Protégé-OWL API is used to operate the ontology and dig is used to call the exterior inference machine (Pellet) to reason. If "a person is willing to help others" is input, the final searching result of OCIRS is shown in Fig. 6, as follows.

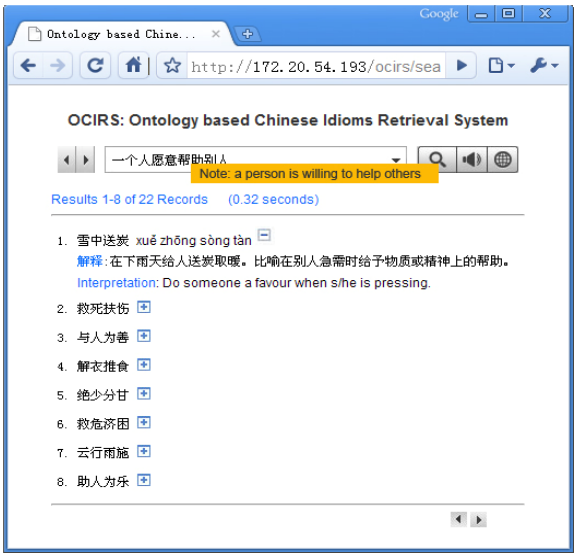


Fig. 6 OCIRS interface with searching result

C. Comparison and Evaluation

Three typical idiom search systems on the Internet are adopted in this paper for comparison. The three idiom search systems include: *iCIBA* Beta online idiom dictionary, *Shuifeng* online idiom dictionary and *Wuyou* online idiom dictionary.

Input "People enjoy high quality in morality", "The morality of people is high" and "A person with high morality" into OCIRS as well as the above-mentioned three idiom search systems, and the search result is shown in Table IV as follows:

TABLE IV
SEARCH RESULT COMPARISON (A)

Searching Request	*OCIRS*	*Wuyou*	*Shuifeng*	*iCIBA*
A person enjoy high morality	y	y	y	n
The morality of a person is high	y	y	n	n
A person with high morality	y	y	n	n

As the meanings of the three short sentences are similar, the meanings of their search results should also be similar. According to the result of Table V, the search results of OCIRS are all the same, which can better meet user's request with respect to the meanings. However, the search results of the other three idiom search systems are different, which fails to meet the user's request with respect to meaning. After 60 times of similar tests, the ratio of successful search to total tests is shown in Table V, as follows:

TABLE V
SEARCH RESULT COMPARISON (B)

Comparison Items	*OCIRS*	*Wuyou*	*Shuifeng*	*iCIBA*
Searching Results	52	59	20	0
Semantically Matching	49	29	18	0
the ratio of Matching	81.67%	48.33%	30%	0

As shown in Table V, the successful search ratio of OCIRS in terms of meaning is 81.67%, which is much higher than the other idiom search systems. Although there is search result for each search in Wuyou online idiom dictionary, the result contains a lot of irrelevant information. The above comparison and analysis show that OCIRS can better meet user's request of searching idiom through meaning.

V. CONCLUSION

This paper aims at the research of the idiom search method. The domain ontology for common idioms is built to express the rich semantic relations in idiom domain. Meanwhile, the grammatical and semantic analyses of the search request input by users are carried out on the basis of idiom domain ontology and semantic dictionary. The prototype system is generally applicable to search idioms through meaning. The semantic search of OCIRS establishes a new way for idiom search.

Moreover, many aspects of OCIRS should be further researched and improved, including the research on the self-learning capability of ontology and the improvement of idiom domain ontology, the deep analysis on the various semantic relations between the concepts in idiom domain, the establishment of an ontology to fully express the various semantic relations between the concepts in idiom domain, and the effective reasoning of inference machine on ontology.

ACKNOWLEDGMENT

This work is supported by NSF Foundation of China under Grant No.60803027, and the Natural Science Foundation in Chongqing City of China under Grant No. CSTC2008BB2312.

REFERENCES

[1] J. Xu, "A study of classification of Chinese idioms," *Journal of Yichun University (social science)*, 2003, vol. 25, no. 5, pp.86-88, in Chinese.
[2] M. Zuo, " Development of idioms in meaning," *Journal of Wuhan University of Science & Technology (Social Science Edition)*, 2004, vol. 6, no. 3, pp.78-81, in Chinese.
[3] *Yojijukugo*, http://en.wikipedia.org/wiki/yojijukugo, last viewed on April 12, 2010.
[4] *Korean idioms*, http://wiki.galbijim.com/Category:Korean_idioms, last viewed on April 12, 2010.
[5] R. Baeza-Yates, *Modern Information Retrieval*. Addison Wesley, 1999.
[6] B. O. Szuprowicz. *Search Engine Technologies for the World Wide Web and Intranets*, Computer Technology Research Corp., 1997.
[7] *Search engines*, http://www.lib.berkeley.edu/teachinglib/Guides/Internet/SearchEngines.html, last viewed on April 12, 2010.
[8] J. Mei. *Synonymous with the Word Forest*. Shanghai Lexicographical Publishing House, 1983, in Chinese.
[9] C.Fellbaum, *WordNet-An Electronic Lexical Database*, MIT Press, 1998.
[10] R. Grubera A translation approach to portable Ontology specifications. *Knowledge Acquisition*, 1993, vol.5, no.2, pp.199-220.
[11] *Description Logics home page*, http://dl.kr.org/, last viewed on April 12, 2010.
[12] Q. Liu and H. Zhang, "Chinese lexical analysis using cascaded hidden Markov model," *Journal of Computer Research and Development*, 2004, vol.41, no.8, pp.1421-1429, in Chinese.
[13] M. Uschold, "Ontologies principles, methods and application," *Knowledge Engineering Review*, 1996, vo.11, no.2, pp.93-155.
[14] M. Gruninger and S. Fox, "Methodology for the design and evaluation of ontologies," *In proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, held in conjunction with IJCAI-95, Montreal, Canada.
[15] M. Fernandez, A. Gomez, N. Juristo, "METHONTOLOGY: From ontological art towards ontological engineering," In *Proceedings of AAAI-97 Spring Symposium on Ontological Engineering Stanford*: AAAI Press, 1997, pp.33-40.
[16] L. Wang and X. Hou, *Chinese Classify Idioms Dictionary*. Guangdong People's Publishing House, 1985, in Chinese.
[17] *Chinese Idioms Dictionary*. http://cy.kdd.cc/sy/, last viewed on April 12, 2010.
[18] B. Wang, *Research on Automatic Chinese-English Bilingual Corpus Alignment*, PhD Thesis, Beijing: Institute of Computing Technology Chinese Academy of Sciences, 1999, in Chinese.
[19] B. Qin and T. Liu, "Question answering system based on frequently asked questions," *Journal of Harbin Institute of Technology*, 2003, vol. 35, no. 10, pp.1179-1182, in Chinese.
[20] B. Jin, Y. Shi, "Similarity algorithm of text based on semantic understanding," *Journal of Dalian University of Technology*, 2005, vol. 45, no. 2, pp.291-297, in Chinese.