Domain-Oriented Retrieval Model Research Based on Meta-search

Yiliang Liu and Guishi Deng
Institute of System Engineering
Dalian University Of Technology
DaLian, LiaoNing, China
E-mail: liuyiliang.dlut@gmail.com

***Abstract*-In allusion to the complex information resource and more and more difficulty of retrieval for the user, this paper present a new retrieval model based on the Meta-search. It associates the user retrieval interface with the domain-specific knowledge base which produce the domain-specific formula and the standard formula as the input of general search engine, at last, by means of meta-search dispatching module and document sorting module adopting Extended Boolean Model, it will return the sorted document set according to the correlativity degree. The experiment proves efficient in satisfying the user's query request.**

Keywords-Domain-specific; Meta-search; Domain knowledge base

INTRODUCTION

With the development of information technology, there are great information space containing millions of sites and billions of pages, so how to mine the useful information from the information ocean is a great challenge. In this situation a lot of general search engines come forth such as Google, Yahoo, Baidu etc, but these search engine exist some problems. First, internet contains mass of domain-specific information, but these search engines return much junk information unrelated to the domain as the user expect; second, these search engines may cover a large number of domain information, but not single one can cover all the information resource, so only one search engine may lead to not enough or even no useful information. More query search words or complex query Boolean formula may lead to more accurate query, but it is difficult for the common user. A new Domain-specific search engine as searching the associated literature with domain-specific(such as Cora in Reference [1,2])may settle this problem in some degree, but the time and network bandwidth consumed by crawlers are excessive, so this limit its application prospect.

Although Kokubo and other authors present the KeyWord Spice (KS method) in Reference [3] and it manifest some great effect in the aspect of domain-specific search, there are some problems still exist. First, the weighted value of the domain filter is accessed randomly, so this may eliminate the efficiency of the domain-specific keywords which represent the query request of user, so it decrease the query precision;

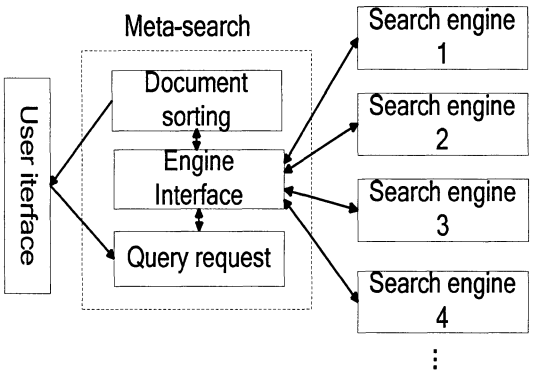

Figure 1. the work flow of the Meta-search

second, KS method does not improve the sorting algorithms of the document set returned by the GSM, so it also decrease the query precision too in some degree. In this situation, a new query model based on Meta-search is presented in this paper, the Domain-Oriented Search Model (DORM).

I. META-SEARCH ENGINE

A. Meta- search Engine

Search engine is an information retrieval tool in internet, it provide a interface of information retrieval, and return the expected information basing on the user's retrieval request. Generally speaking, search engine is the common independent search engine as we know, but Meta-search engine [4] is a kind of search engine based on the general search engine, it consists of several independent search engines, in this paper, we will use member search engine instead of the internal independent search engine of the meta-search engine.

B. The Framework and Run able-mechanism of the Meta-search Engine

Meta-search engine implement retrieval through a unified interface and distributed module helping the user select the suitable search engine from mutil-search engines. It consists of global control mechanism of multiple retrieval tools basing on the distributed networks. Eliminating the disadvantages of the independent search engine, the user can obtain the result of multiple search engines aiming at one query as no need to inquiry one by one. Its workflow as the Figure 1.

978-1-4244-2013-1/08/$25.00 ©2008 IEEE

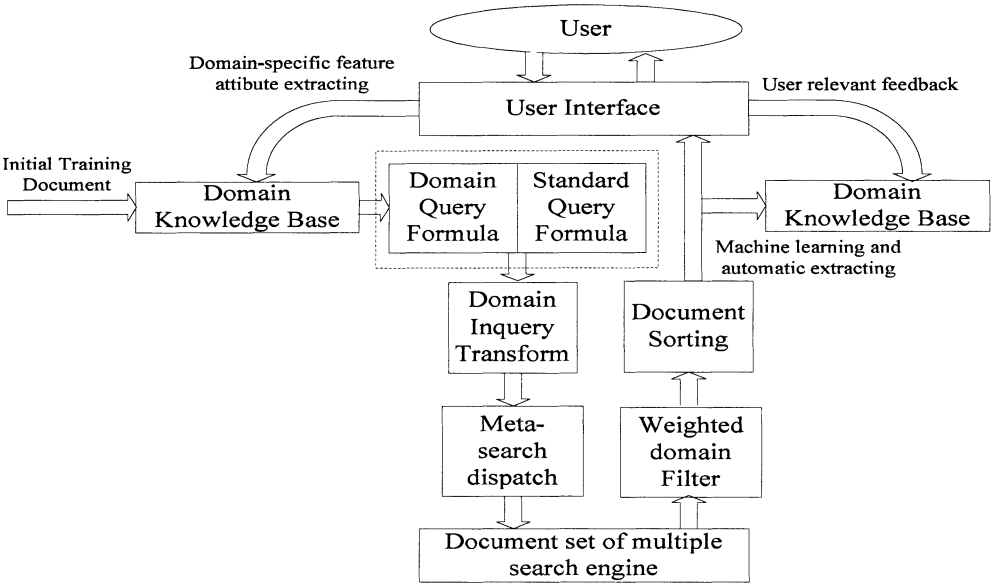

Figure 2. Construction of the DORM based on Meta-search engine

II. THE DESIGN AND REALIZATION OF THE DOMAIN-ORIENTED RETRIEVAL MODEL

In order to reuse the large indices of the general search engine, we present a Domain-Oriented Retrieval Model (DORM) based on the Meta-search engine. It can eliminate the irrelevant documents from returned ones by domain-specific filter module and its construction as the figure 2 as below, it combines the advantage of the meta-search engine with query extension to strengthen the ability of domain-specific retrieval.

A. Construction of the Domain Knowledge Base

1) Representation of the domain-specific knowledge

Knowledge presentation (KP) is also called knowledge description. Among the methods of knowledge representation, we use framework representation method to descript domain-specific knowledge. Its advantage is that framework embedded represents the knowledge details distinctly and the structure can implement the efficient matching and reasoning. In the process of constructing the Domain Knowledge Base, domain-specific knowledge representations need briefness, convenient to extract and match. The knowledge representation framework of Domain Knowledge Base consists of domain class, domain feature name, domain feature value.

<Top frame> ::=< top frame name>
<Slot 1> ::=< slot 1 name>< slot 1 value>
<Slot 2> ::=< slot 2 name>< slot 2 value>
.....
<Bottom frame> ::< bottom frame value>

As to the Domain Knowledge Base(DKB),framework structure of the domain-specific knowledge[5]is: top frame: domain class(D_C) name; slot 1 name: knowledge feature attribute(D_A);slot 1 value: included attribute value(0 or 1);

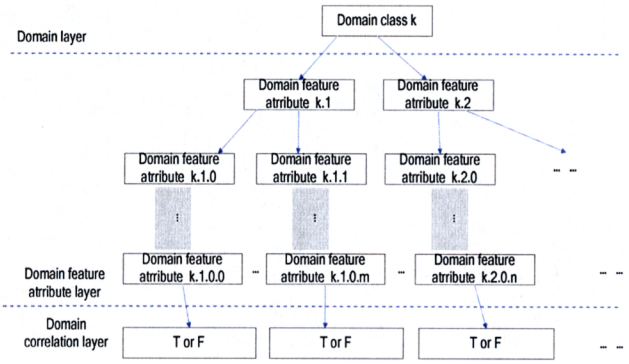

Figure 3. Classifying system of the domain-specific knowledge base according to the hierarchy

bottom frame name: domain-specific correlation; bottom frame value: domain-specific relevant document sets. Among the DKB, every framework represent a domain-specific feature attribute.

2) Construction of the DKB

Knowledge base organize the knowledge using classification and hierarchy, this method can improve the maintenance of the knowledge base greatly. Domain-specific knowledge is classified according to the category of the domain knowledge, the domain feature attribute, and the organization structure of the DKB and the logical comprehension of the user. Knowledge structure of the DKB based on the ontology can associate the knowledge representation hierarchy into a organic combination system, which implements the share and reuse of the domain knowledge, makes the construction of DKB clear and easy to maintain. In the process of constructing the DKB model, the hierarchical architecture consists of three tiers: domain tier; domain feature attribute tier, domain relation tier. Domain category composes domain

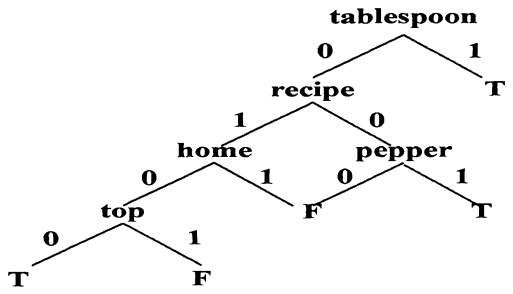

Figure 4. An example of the decision tree that classified the document

tier; domain feature attribute ties consist of initial training document and manual work which represent query end-point and determine the domain query path, the construction is as the Figure 3.

B. Query Request Transformation Module

Query request transformation is a key module in the DORM model, Construction method have a lot ,but Keyword Spice method is a simple but efficient method at producing the domain formula, so we will use this method to reduce the result set unrelated to the user so as to improve the precision ratio. First, we extract the domain-specific key words from the user interface, and then find out the optimal path related to the domain-specific key words by using the decision tree. The relationship of domain-specific key words and the domain tree is as a binary tree and the construction as below. The node represents the domain feature attribute; The value of branch represent correlation degree, the value of the leafy node determine the correlation. Starting from the root node, judge if the query request contain the domain category, then choose the corresponding branch, the process will sustain until reach the leafy node. This three classify the query request (document initially) into two category (related) and F(unrelated).

For example, figure 4, the path from the root node to leafy node which is equal to T is the domain conjunctive formula, then synthesis of the disjunctive formula represent the correlation degree between the document and the query request. Domain request transform to the Boolean formula as below:

$$\begin{aligned} q_d &= tablespoon \\ &\vee(\overline{tablespoon} \wedge \overline{recipe} \wedge pepper) \\ &\vee(\overline{tablespoon} \wedge recipe \wedge \overline{home} \wedge \overline{top}) \end{aligned} \tag{1}$$

C. Query Formula

Formal definition of the query formula is $q = q_s \Lambda q_d$,Among them, q_s represent the standard formula which experience the processing operation of eliminating stop words, extracting the word stem, selecting the index term etc; q_d is the domain query formula produced by the DKB. Over here, q_d is different from the q_d presented in the KS method. The distinction is that the domain keyword is random but here, q_d is determined by the correlation binary tree, so it reflects the user's query request.

D. Meta-search Dispatching Module

Concerning the selection problem of member search engine [6, 7], so how to select the member search engine will directly affect the response time, precision ratio etc. In this paper, we will design the workflow as below:

1) Calculating the weight value ω_d of the domain feature attribute by using the odds ratio;

2) Calculating the correlation degree between domain feature attribute and the member search engine

So as not to affect the recall ratio and the efficiency of the total search process, we will use top feature attribute method: First, discompose the path correlated to the domain from the domain query formula, then, find out the top feature attribute and subordinate top feature attribute. The formula of calculating the correlation degree is:

$$\omega_j = \sum_i \omega_i \times rel(i, j) \tag{2}$$

where ω_j is the correlation degree of domain feature attribute with member search engine j; ω_i is the weight value of the domain feature attribute; rel(i,j) is the correlation degree of top feature attribute or the subordinate feature attribute with the member search engine(in this paper, ω_j have not been normalization processing, so ω_j may be greater than or equal to 1).

3) Sorting the correlation degree in descending order, then select several membership in front(in this paper ,we select the top 3 member search engine)

E. Weighted Domain Filter Module

1) Calculating the weighted value of the domain feature attribute associating with the initial query.

We know that, in text feature extracting method, the importance of the domain feature attribute in domain query formula is different, bigger is $|OR_i|$,the more importance of domain feature attribute i. so the definition of the weighted value of the domain feature attribute associating with the initial query is:

$$\omega_i = |OR_i| \tag{3}$$

2) Calculating the weighted value of the domain-specific feature attribute returned by the meta-search engine.

Through weighting the domain-specific feature attribute can improve the degree associating with the domain-specific so as to improve the precision ratio, avoiding a large number of document sets unrelated to the user's request. for not losing the generality, we use the classic method TF*IDF.

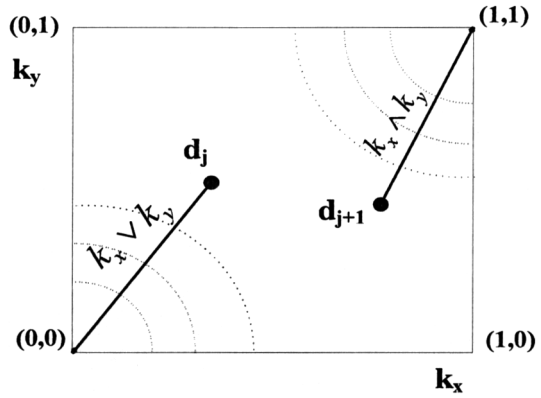

Figure 5. Two-dimensional space consists of k_x and k_y based on Extended Boolean Model

$$\omega_i = TF_i \times IDF_i = \frac{m_i}{\sum m_j} \times \lg(\frac{M}{k_i}) \quad (4)$$

Where ω_i is the weighted value of the domain-specific feature attribute i in document; m_i represent the times of domain-specific feature attribute i in document; $\sum m_j$ is the sum of all the times of domain-specific feature attribute i in document; M represent the number of the document in the document set; k_i represent the document number of domain-specific feature attribute i in document set.

F. Documents Sorting Module

In information retrieval (IR)theory, Boolean model and vector model are two simple but efficient models which are widely used. Due to the limitation of their application,this paper select the extended Boolean model[8].This model have obvious advantage in processing the Boolean formula, moreover, domain-specific formula in this paper is Boolean.

1) Calculating the similarity degree between the document and the standard query

If only considering two query words,we can use 2D drawing to represent the document and the query standard, as the figure5,through binary array $[d_j, k_x]$ and $[d_j, k_x]$,and their weighted value $\omega_{x,j}$, $\omega_{y,j}$ determine the position of document d_j ;assume that the value of $\omega_{x,j}$ and $\omega_{y,j}$ are between (0,1);their values are determined by the standardized factor tf-idf:

$$\omega_i = TF_i \times IDF_i = \frac{m_i}{\sum m_j} \times \lg(\frac{M}{k_i}) \quad (5)$$

where $f_{x,j}$ represent the normalized frequency of query word k_x in document d_j ; idf_i represent inverse document frequency of k_x.

As to the disjunctive query $q_{or} = k_x \vee k_y$,(0,0) is an ineffective point ,so the distance to(0,0) represent the similarity degree between query formula q_{or} and the document:

$$sim(q_{or}, d_j) = \sqrt{\frac{\omega_{x,j}^2 + \omega_{y,j}^2}{2}} \quad (6)$$

As to the conjunction query $q_{and} = k_x \wedge k_y$,(1,1) is ideal point, so the distance to (1,1) represent the similarity degree between the query q_{and} and document:

$$sim(q_{and}, d_j) = 1 - \sqrt{\frac{(1-\omega_{x,j})^2 + (1-\omega_{y,j})^2}{2}} \quad (7)$$

As not to lose the generality, we assume that the number of query words included in the initial query words is t, so the query formula is $q_s = k_1 \wedge k_2 \ldots \wedge k_t \vee k_{t+1} \ldots \vee k_t$,the similarity degree of the standard query formula with the document is

$$sim(q_s, d_j) = \sqrt{\frac{sim(q_{and}, d_j)^2 + sim(q_{or}, d_j)^2}{2}} \quad (8)$$

2) Calculating the similarity between the document and the domain-specific query formula.

In this paper, domain-specific query may exist the "not contain" forms, so the corresponding expression is $k_x \wedge \overline{k_y}$: contain k_x but not contain k_y ;so (1,0) is the ideal point, so the distance to (1,0) represent the similarity degree of the domain-specific query formula with the document:

$$sim(q_{and}, d_j) = 1 - \sqrt{\frac{(1-\omega_{x,j})^2 + \omega_{y,j}^2}{2}} \quad (9)$$

As not to lose generality, we assume that number of domain-specific words included in the domain-specific query is n, so the domain-specific query formula is $q_d = k_1 \wedge k_2 \wedge \ldots k_m \wedge \overline{k_{m+1}} \ldots \overline{k_n}$ and the similarity degree between the domain-specific and the document is:

$$sim(q_d, d_j) = 1 - \sqrt{\frac{\sum_{i=1}^{m}(1-\omega_{x,j})^2 + \sum_{y=m+1}^{n}\omega_{y,j}^2}{n}} \quad (10)$$

3) Calculating the similarity degree of sorting document

The final query formula is the conjunction of the standard query formula and the domain-specific query formula: $q = q_s \wedge q_d$ .So the similarity degree of the document which will be finally returned is:

$$sim(q, d_j) = \sqrt{\frac{sim(q_s, d_j)^2 + sim(q_d, d_j)^2}{2}} \quad (11)$$

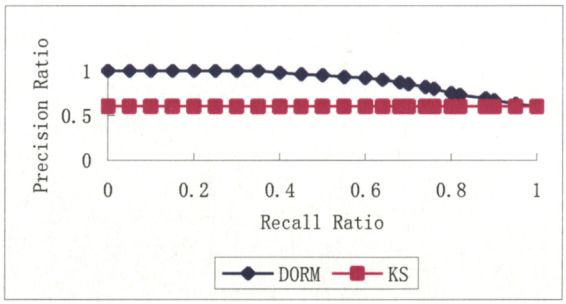

Figure. 6. Query effect comparison of DORM and KS

TABLE I. ETRIEVAL EFFECT OF DIFFERENT QUERY FORMULA

Query formula	Effect		
	Precision Ratio	Recall Ratio	Harmonic mean value
match	0.317	1	0.241
match∧ football	0.843	0.821	0.416
match∧q_d	0.794	0.915	0.425

TABLE II. COMPARISON OF PARAMETER INDEX BETWEEN DORM AND KS

Parameter index of search engine	Method		
	DORM	KS	Target value
Average number of documents returned	291	347	227
Average number of related documents returned	191	211	227
Average precision ratio	0.656	0.608	1
Average recall ratio	0.841	0.930	1
Harmonic mean value	0.457	0.368	0.5

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Evaluation Method of Retrieval Effect*

This paper proves the retrieval effect of the DORM model by example. Though precision and recall are widely applied to evaluate algorithm performance, yet there are some problems in the two measure methods when they are applied to this paper: firstly, searching reasonable evaluation of the maximal precision should know all documents in set, while the DORM model would get many unknown documental sets based on Meta-search, therefore we can not accurate estimate of the recall; Secondly, recall and precision are the capabilities of searching set based on batch processing pattern, while the DORM model focuses on the interaction with customers to improve the satifaction. Hence fore, we use the harmonic mean method to measure:

$$F(j)=\frac{2}{\frac{1}{r(j)}+\frac{1}{p(j)}} \tag{12}$$

Where r(j) and p(j) are recall and precision of the j document, respectively; F(j) is the harmonic mean of r(j) and p(j).

B. *Retrieval Example*

We just research six fields of network newsgroup documents[9]: football, hockey, hardware, software, economy and politics. Choosing 600 documents per field, and the total quantity of documents is 3600, and the searching goal is the documents in football field. At first, we choose 30 words, which include ten words with maximal positive values and ten words with minimal negative values, as the field words by frequency ratio method. Then, construct the field query formula with these field words based on decision tree method. By simplifying, the field query formula about football field changes to:

$$q_d=(foot)\vee(attacker)\vee(football\wedge\overline{basketball}\wedge\overline{baseball})$$
$$\vee(court)\vee(team\wedge\overline{NBA}\wedge doorkeeper)$$

At present, we want to search the news about match in this field, and given that there are 817 documents including the news of match and 411 documents including the news of football.

1) Retrieval effect of different query formula

Searching "match","match∧footbal"and "match∧q_d" (q_d is the domain-specific query formula according the method introduced in front chapters) respectively, and the results are as the TABLE I.

2) Comparison between DORM model and KS method

The comparison results about precision and recall are shown in Figure 6, and the retrieval result of KS method will not change with precision for sequencing because this method directly returns the results to customers.

To further prove the efficiency of DORM, select ten times of query which the keywords belong to different aspects of the domain and descript the information result as the TABLE Ⅱ.From the comparison of the parameter index of search engine, we know that KS method may be more efficient in recall ratio, but junk information irrelevant to the user's request too much, So on the aspect of balancing the satisfaction of the user, DORM presented in this paper is more excellent in some degree.

IV. CONCLUSIONS

In this paper, we propose the searching strategy model based on Meta-search engine, which may collect information of appointed field more quickly and exactly, meanwhile delete unrelated information and improve efficiency and precision of searching. Customers would just simply reconstruct domain-specific knowledge base and the initial training document sets, if they want to construct a domain-specific search engine about another field. Therefore, this searching model has universal applicability.

ACKNOWLEDGMENT

The authors gratefully acknowledge National Natural Science Foundation of China supporting this project, The Grant No is 70671016;meanwhile,this paper is also supported by the project of Digital Case Library Construction of DUT.

REFERENCES

[1] A McCallum, K Nigam, J Rennie, K Seymore. "A Machine Learning Approach to Building Domain-Specific Search Engines "[C]//Proceedings of 16th Int'l Joint Conf. Artificial Intelligence. SanFrancisco: Morgan Kaufmann Publishers Inc, 1999: 662-667.

[2] Andrew McCallum, Kamal Nigam ,Jason Rennie,Kristie Seymore. Buiding Domain-Specific Search Engine with Machine Learning Technology[J]. In AAAI Spring Symposium on Intelligent Agents in Cyberspace(2003)

[3] KOKUBO TAKASHI, OYAMA SATOSHI, YAMADA TERUHIRO. KITAMURA YASUHIKO, ISHIDA TOORU. Keyword Spice Method for Building Domain-specific Web Search Engines[J]. Transactions of Information Processing Society of Japan ,1804-1813(2002).

[4] Lieming Huang , Matthias Hemmje 1, Erich J. Neuhold. ADMIRE: an adaptive data model for meta search engines[C]. Computer Networks 33 (2000) 431–448.

[5] LV Chuan-yu,LI Hua,GEN Hu." Analytic Method for the Free Vibration of Plane Frame Structur".Journal of Chongqin University. Vol.27 Jul 2007.

[6] Adele E.Howe , Daniel Dreilinger . A Meta Search Engine that Learns which Search Engines to Query[J]. Computer Networks (1997) .

[7] XU Ke, HUANG Guojing,CUI Zhimin. "Personalized scheduling algorithm based on user profile for meta search engine".Journal of Tsinghua University(Sci&Tech).Vol.45,NO.S1(2005).

[8] YUE Wen, CHEN Zhi-ping,LIN Ya-ping. Information-retrieval Algorithm Based on Query Expansion and Classification [J]. Journal of System Simulation, 2006,18(7): 1926-1929.

[9] T M Mitchell. Machine Learning [M]. USA: McGraw-Hill, 1997: 187-189.