Research on Personalized Query Optimization Service in e-Learning System

Xiaojian Li
Computer School
Wuhan University
Wuhan, Hubei, China
e-mail: lixiaojian_k@163.com

Shihong Chen
National Engineering Research Center for Multimedia Software
Wuhan University
Wuhan, Hubei, China
e-mail: chen_lei0605@sina.com

Abstract—**According to the traditional search engine's lack of understanding the semantic query intension and providing personalized service in e-Learning systems, this paper presented a personalized query optimization service to satisfy both semantically and personalized searching demands of the user by semantic mining from user's searching information to build the user model which is exploited to perceive the query intension of the user. To testify the effectiveness of the service, some experiments are carried out and the results show that the search service has been improved.**

Keywords- Query optimization; User model; Query intension; ontology; e-Learning

I. INTRODUCTION

In information systems, search service is the main and important approach for users to acquire information they need. Although the technology for search engine improves fast as the web and communication technology advances. There are still some disadvantages in current search services: 1) the search engine can hardly understand the semantic intension of the user query and always bring thousands of unwanted search results; 2) most search services are focused on one-size-fit-all approach to provide service. They barely take the individual interests into account. In fact, different users have different query intension even use exactly the same query keywords. Unless the personalized search service is provided, the user can hardly be satisfied.

In recent years, e-Learning system becomes the important application of web technology just like e-commerce and e-government. In the project "e-Learning System Platform and Education Resource Base Research and Development" which is supported by the National High Technology Research and Development Program of China (863 Program), we built an e-Learning system with metadata annotation adopted to describe the content of recourses. An ontology represented by RDF(s) [1][2] is used to depict the concepts, instances and relationships of those annotations within a certain domain. To enhance the performance of the search service and satisfy both semantically and personalized resource searching demands of the user, we provide a personalized query optimization service by semantic mining from user's searching information to build the user model and exploiting the user model to perceive the query intension of the user.

II. PERSONALIZED QUERY OPTIMIZAION SERVICE

The system framework of personalized query optimization service is demonstrated in Figure 1. There are 6 main modules which are user query receiver, traditional search engine, semantic query analyzer, personalized result re-arranger, query result deliverer and user model generator. A typical query optimization process is performed as follows: a certain user uses query receiver module as an input interface to collect the query keywords which are then transferred to a keywords based traditional search engine that search through the education resource base and store the original search result in a temporary result set. At the same time, query receiver transfers those keywords to the semantic query analyzer which discovers the user query intension by semantic query expansion using the semantic user model constructed by user model generator module. The semantically expanded query terms are adopted by personalized result re-arranger as keywords to research through the original search result and attain the optimized results that are presented to the user by query result deliverer which also gathers the user's feedback and record them in the search log.

From the above discussion, we can conclude that the core modules are semantic query analyzer and personalized user model generator.

A. Personalized User Model Generator

An efficient approach to providing personalized service in information systems is to introduce user model which had been adopted in various systems to represents identity information and interests of users [3][4][5]. In the e-Learning scenario, user model is exploited to represent the interests and background knowledge of individual learners [6].

During the process of searching, users attend to find the resources in which they are interested. In fact, there are semantic relationships between the user's request and searching target. Research indicates that those relationships reflect the current interest of the user perfectly. The function of user model generator is to construct the personalized user model which represents individual interests of a user and relationships of those interests by semantic mining from user's searching process which is recorded in the search log.

978-0-7695-3843-3/09 $26.00 © 2009 IEEE
DOI 10.1109/MINES.2009.245

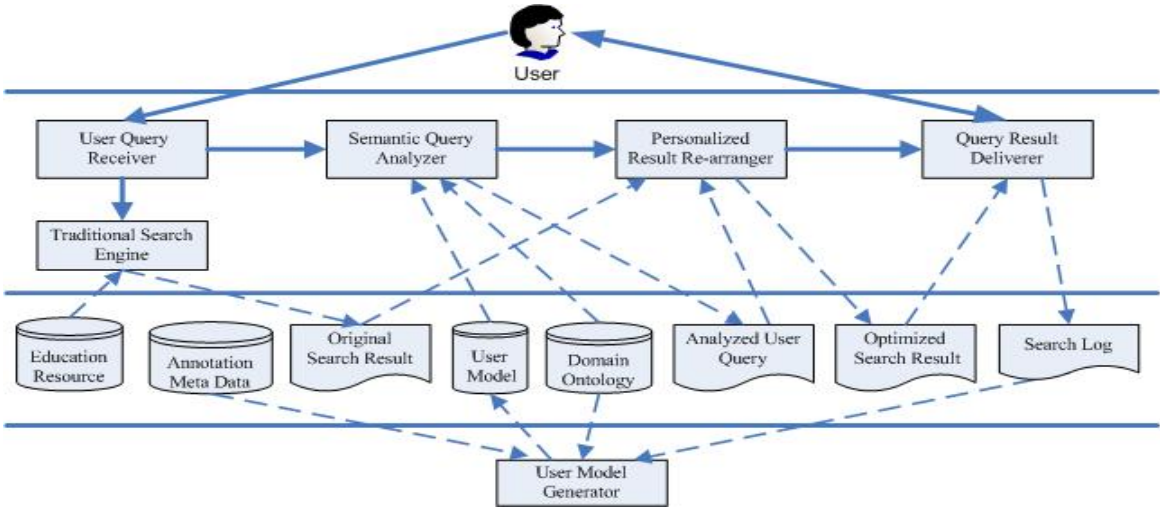

Figure 1. The system framework of personalized query optimizaiton service

The user model is formalized as U_{model}= *(UP, {(C,CS,AS,TW)})* whre *UP* is user's identity information which usually contains registered ID, true name, age, education level, learning orientation and so forth; $C=\{c_1,c_2,\ldots,c_k\}$ is a set of concepts that user is interested in; $CS=\{(c_1,c_2,\ldots,c_m)|(ci \in C,\ 1 \leqslant i \leqslant m \leqslant k)\}$ is a set of combinations of semantically related concepts in the set *C* attained by semantic mining; $AS=\{(c_i, P_1, e_2, P_2, e_3, \cdots, e_{n-1}, P_{n-1}, c_j)|(c_i,c_j \in C,\ 1 \leqslant i,\ j \leqslant k)\}$ is a set of semantic association sequences [7] that connect two related concepts in set *C*, $TW=\{TS_1,TS_2,\ldots,TS_n\}$ is a set of timestamps representing the distribution of time during which user have corresponding interests.

To construct the user model, the search log is utilized for mining. Typical data that recorded in search log are searching keywords used by the user; top *k* resources selected by the user where *k* can be adjusted on demand; the time-point when the search process occurred. The mining process is as follows:

1): Let set $UK=\{uk_1,uk_2,\ldots uk_l\}$ be consisted of query keywords and set $MD=\{md_1,md_2,\ldots md_m\}$ be consisted of annotations to selected resources. Use variable $ts_{current}$ to store the time-point. Let $W=UK \bigcup MD=\{w_1,w_2,\ldots,w_n\}$, where $1 \leqslant n \leqslant l+m$;

2): For each $w_i \in W$, build a corresponding set $SYN_i=\{S_{i1},S_{i2}, \cdots,S_{im}\}$ consisting of synonyms of w_i by analyzing through the domain lexicon, where $1 \leqslant i \leqslant n$;

3): For each SYN_i where $1 \leqslant i \leqslant n$, use the method of keyword matching to map the element in SYN_i to the elements defined in the domain ontology, which will lead to the result of corresponding sets $OC_i=\{OC_{i1},OC_{i2},\cdots\cdots,OC_{ie}\}$ where OC_{ij} is a concept or instance defined in the domain ontology, and $1 \leqslant j \leqslant e,\ 1 \leqslant i \leqslant n$;

Definition 1: the above 3 steps together is defined as concept mapping from keywords to elements in ontology.

4): Let set of combinations of elements in every OC_i be $CS=\{(c_1,c_2,\ldots,c_n)\ |(c_i \in OC_i \wedge OC_i \neq \Phi \wedge 1 \leqslant i \leqslant n) \vee (c_i=\delta \wedge OC_i=\Phi \wedge 1 \leqslant i \leqslant n)\}$ where Φ represents the empty set. Let crv_{ij} be the the semantic relevance degree between c_i, c_j, where c_i, $c_j \in cs_h$, $cs_h \in CS$.

For $\forall\ c_i,\ c_j \in cs_h,\ cs_h \in CS$, let crv_{ij} be the semantic relevance degree between c_i, c_j . If c_i, c_j are not semantically connected or $c_i=\delta$ or $c_j=\delta$, let crv_{ij} =0. Otherwise they are semantically connected and there exists a semantic association as_{ij} connecting c_i, c_j. Let set $AS_{current}$ be the set of these associations, add as_{ij} in the set $AS_{current}$.

The equation to calculate crv_{ij} based on Component Subsumption Weight(S_A), Rarity Weight(R_A) and Length Length(L_A) [7][8] is as follows:

$$crvij = \kappa_1 \times S_A + \kappa_2 \times R_A + \kappa_3 \times L_A \tag{1}$$

where *A* represents the association sequence consisting of a minimum number of entities and properties that semantically connects c_i and c_j, $\kappa_1+\kappa_2+\kappa_3=1$, $\kappa_{1,}\kappa_{2,}\kappa_3>0$ are weight constants that can be adjusted on demand.

Let CR_h be the semantic association rank of $cs_h \in CS$, the equation to calculate it is as follows:

$$CR_h = \sum_{i,j=1}^{i,j=n} crvij \tag{2}$$

5): Arrange the elements in the set *CS* in descending order by corresponding semantic association rank, select top Γ elements in arranged *CS* to form a set $CS_{current}=\{cs_1,cs_2,\ldots cs_{\Gamma}\}$ where Γ can be adjusted on demand. The set $CS_{current}$ is a set of combinations of semantically related concepts representing the current user interests after current semantic mining. Make $CS_{current}$, $AS_{current}$ and $ts_{current}$ be the output of current mining process.

6): Let *ICF=(C, CS, AS, TW)* represents the semantic mining result from a certain searching sample occurred at a

certain point of time, where $C=\bigcup_{i=1}^{\Gamma} cs_i$, $cs_i \in CS_{current}$; $CS=CS_{current}$; $CR=\{(CR_h)|(CR_h$ *is the semantic association rank of* $cs_h) \wedge (cs_h \in CS, 1 \leqslant h \leqslant \Gamma)\}$; $AS=AS_{current}$; $TW=\{ts_{current}\}$.

7): During certain period of time we can get a bunch of searching samples of a user. Correspondingly, we are able to get a set $ICFs=\{ICF_1, ICF_2, ..., ICF_n\}$ by mining from these samples and then we semantically aggregate them to complete the construction of user model;

8): Select a certain period of time ΔT, let t_{start} and t_{end} be the starting time-point and ending time-point respectively, Create a set $ICFs' \subseteq ICFs$, where $ICFs'=\{(ICF_i)|(ICF_i \in ICFs) \wedge (t_{start} \leqslant ICF_i.tw_1 \leqslant t_{end}, ICF_i.tw_1 \in ICF_i.TW)\}$. Let $|ICFs|=k$;

9): For $\forall A,B \in ICFs', A \neq B$, let $sim\_c(A, B)$ represents the semantic aggregation relevance degree between A and B and the equation is defined as:

$$sim\_c(A,B)=\frac{\sum_{i=1}^{t} Aw_i * Bw_i}{\sqrt{\sum_{i=1}^{t} Aw_i^2} * \sqrt{\sum_{i=1}^{t} Bw_i^2}} \tag{3}$$

where $AW=(Aw_1, Aw_2, ..., Aw_t)$ is the aggregation eigenvector of A where $t=|A.C \cap B.C|$, every $Aw_i \in AW$ is the aggregation eigenvector that indicates semantic relevance contribution degree of corresponding concepts $c_i \in A.C \cap B.C$ and be calculated by

$$Aw_i=\sqrt{\sum_{j=1}^{N}\left(\frac{\sum_{k=1}^{n_j} crv_{ik}}{cr_j}\right)^2}\Big/N \tag{4}$$

where $N=|A.CS|$, $n_j=|A.C \cap B.C \cap cs_j|$, $cs_j \in A.CS$, crv_{ik} represents the semantic relevance degree of c_i and c_k, $c_i \in A.C \cap B.C$, $c_k \in cs_j$. The BW is defined likewise.

Thus the equation to calculate the aggregation rank between A and B is:

$$sim\_merge(A,B)=\frac{|A.C \cap B.C|}{|A.C \cap B.C|+\alpha|A.C-B.C|+\beta|B.C-A.C|} * sim\_c(A,B) \tag{5}$$

where $0 <= \alpha, \beta <= 1$, $\alpha+\beta=1$ are constants;

10): Create a aggregation rank matrix $MR=[sm_{ij}]$, where

$$sm_{ij}=\begin{cases} sim\_merge(ict_i, ict_j), & i \neq j, \quad ict_i, ict_j \in ICcuts' \\ 1, & i=j \end{cases} \tag{6}$$

$0 \leqslant i,j \leqslant |ICFs'|$, and apparently $sm_{ij}=sm_{ji}$.

Define an equivalence binary relation

$$\begin{aligned} mm=\{&<a,b>|((sm_{ab}>\theta) \vee \\ &(\exists a_1 ... a_n \in ICcuts' \rightarrow (sm_{aa1}>\theta) \wedge (sm_{a1a2}>\theta) \wedge ... \wedge (sm_{anb}>\theta))) \\ &\wedge a,b \in ICcuts'\} \end{aligned} \tag{7}$$

where $0<\theta<1$ is a threshold.

11): Let set $MSX=\{msx_1, msx_2, ..., msx_h\}$ where $msx_i \subseteq ICFs'$, $1 \leqslant i \leqslant h$ be a equivalence class set of $ICFs'$ based on binary relation mm. Thus MSX is a set partition of $ICFs'$. Every $msx_i \in MSX$ is prepared for aggregating:

$$ics_i=(\bigcup_{i=1}^{l} a.C, \ \bigcup_{i=1}^{l} a.C, \ \bigcup_{i=1}^{l} a.AS, \ \bigcup_{i=1}^{l} a.TW) \tag{8}$$

where $a \in msx_i$, $l=|msx_i|$.

Thus the final user model we get is $U_{model}=(UP, \{ics_i\})$.

B. Semantic Query Aanalyzer

The user's search intension lies in the query keywords which usually are hardly captured by search services nor correctly described by the user. To perceive the search intension, we need to find out the semantic information in those keywords and make the detailed information more available. The user model we presented can convey the latent semantic requirement and personal interests of the user. The semantic query analyzer exploit the user model to perform personalized semantic query expansion to perceive the query intension of the user. The process is as follows:

1): Let set $SK=\{sk_1, sk_2, ..., sk_n\}$ be the set of original user's query keywords, use the concept mapping steps in definition 1 to construct corresponding sets $SOC_i=\{SOC_{i1}, SOC_{i2}, \cdots\cdots, SOC_{il}\}$ where OC_{ij} is a concept or instance defined in the domain ontology, $1 \leqslant j \leqslant l$, $1 \leqslant i \leqslant n$;

2): Let set of elements' combinations in every SOC_i be $SS=\{(s_1, s_2, ..., s_n) \ | (s_i \in SOC_i \wedge SOC_i \neq \Phi \wedge 1 \leqslant i \leqslant n) \vee (s_i=\delta \wedge SOC_i=\Phi \wedge 1 \leqslant i \leqslant n)\}$ where Φ is the empty set. Let CR_h be the semantic association rank of $s_h \in SS$ which is calculated by (7) where crv_{ij} in (7) is the semantic relevance degree between $s_i, s_j \in cs_h$;

3): Arrange the elements in the set SS in descending order of corresponding semantic association rank, pick up top k elements in arranged SS to form a set $SCS=\{scs_1, scs_2, ...scs_k\}$ where k is a constant that can be adjusted on demand. The set SCS is a set of combinations of semantically related concepts representing the semantic mapping results of current user keywords;

4): Let UM be the user model of current user. Let $SIC \subseteq UM.(C,CS,AS,TW)$ be the interest space of current searching which is created by calculating the interest relevance degree(sim_ci) between SCS and $UICi \in UM.(C,CS,AS,TW)$, then

$$sim\_ic_i=\frac{|SCSunion \cap UICi|}{|SCSunion \cap UICi|+\alpha|SCSunion-UICi|+\beta|SCSunion-UICi|} * \frac{|UICi.TW|}{UICi.TW_{max}-UICi.TW_{min}} \tag{9}$$

where $UIC_i \in UM.IC$, $1 \leqslant i \leqslant |UM.IC|$, $SCSunion = \bigcup_{i=1}^{k} scs_i$, $UIC_i.TW_{min} = min\ (t_j - t_0)$, $t_j \in UIC_i.TW$, t_0 is a selected starting time-point; $0 <= \alpha, \beta <= 1$, $\alpha + \beta = 1$ are constants that can be adjusted on demand.

Arrange the elements in the set $UM.(C,CS,AS,TW)$ in descending order of corresponding interest relevance degree, select top π elements in arranged $UM.(C,CS,AS,TW)$ to form the interest space set $SIC=\{ sic_1, sic_2, ...sic_\pi \}$ where π is a constant that can be adjusted on demand.

5): The expansion relevance degree between elements of concepts combinations in SCS and elements of concepts combinations in $sic_j \in SIC$ is defined as:

$$sim\_expand(scs, cs) = \frac{|scs \cap cs|}{|scs \cap cs| + \alpha |scs - cs| + \beta |scs - cs|} * cr_{cs} \quad (10)$$

where $scs \in SCS=\{scs_1, scs_2, ...scs_k\}$, $cs \in sic_j.CS$, $sic_j \in SIC=\{ sic_1, sic_2, ...sic_\pi \}$, $1 \leqslant j \leqslant \pi$, $cr_{cs} \in sic_j.CR$ is the semantic association rank of cs, $0 <= \alpha, \beta <= 1$, $\alpha + \beta = 1$ are constants that can be adjusted.

For each $scs_i \in SCS$, arrange the elements in the set $sic_j.CS$ in descending order of value calcutlated by $sim\_expand(scs_i, cs_l)$ where $cs_l \in sic_j.CS$, select top μ elements in arranged $sic_j.CS$ to form a set $CSexpnd_{ij}=\{cse_1, cse_2, ...cse_\mu\}$ where μ is a constant that can be adjusted.

Let set $SCSexpnd_{ij}=\{scse_1, scse_2, ...scse_\mu\}$ where $scse_g = scs_i \cup (scs_i - cse_g), 1 \leqslant g \leqslant \mu$ be a query expansion result of scs_i on sic_j which represents the query expansion of the i^{th} element in SCS according to the j^{th} element in interest space SIC. Thus we have set $s\_expand_i = \bigcup_{j=1}^{\pi} SCSexpnd_{ij}$ be the total query expansion of the i^{th} element in SCS and set $SE=\{s\_expand_1, s\_expand_2, ..., s\_expand_k\}$ be the final query expansion output. The constants k, π, μ used in above process can be adjusted on demand.

III. Experiment and Analysis

The personalized query optimization service we provide intends to improve the search performance and user's satisfaction in our e-Learning system. To testify the effectiveness of the approach, some experiments are carried out. We asked a user to perform 20 searching processes using the query optimization service, and repeat searching processes using traditional search engine. Finally, we compare the searching precision ratio between those searching processes to do analysis and results are showed in Figure 2. The results indicate that the average precision ratio had risen by 13.25%.

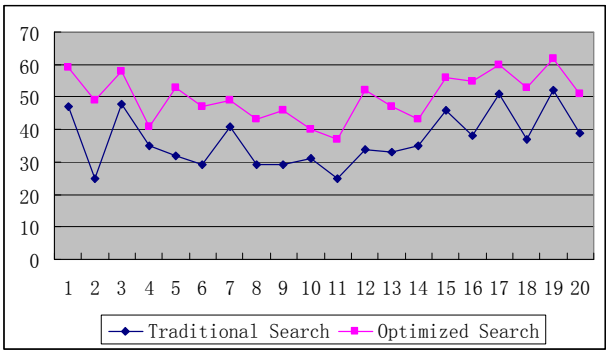

Figure 2. The comparison results of precision ratio

IV. Conclusion

To enhance the performance of search service in e-Learning system, this paper presented personalized query optimization service to satisfy both semantically and personalized searching demands of the user by semantic mining from user's searching information to build the user model which is exploited to perceive the query intension of the user. We illustrate the system framework for the service; discuss the function and algorithm of core modules in the system. To testify the effectiveness in providing personalized service and conveying the latent semantic requirement of the user, we did some experiments. The results of the experiments show that the performance has been improved. Our future work includes the adaptive update and optimization of the user model.

References

[1] Gomez-Perez, Asuncion and Corcho Oscar, "Ontology languages for the semantic web", *IEEE Intelligent Systems and Their Applications*, Institute of Electrical and Electronics Engineers Inc., 2002, pp. 54-60.

[2] O. Lassila and R. R. Swick, "Resource Description Framework(RDF) Model and Syntax Specification". *W3C Recommendation,* http://www.w3.org/TR/REC-rdf-syntax/, 1999.

[3] Balabanovid, "Learning to surf: multiagent systems for adaptive Web page recommendation", Ph.D. dissertation, Dept. of Computer Science, Stanford University, 1998.

[4] L. Chen and K. Sycara, "WebMate: Personal agent for browsing and searching", *In Proceedings of the 2nd International Conference on Autonomous Agents*, ACM Press, New York, USA, 1998, pp. 132-139.

[5] K. Lang, "NewsWeeder: Learning to filter net news", *In Proceedings of 12th International Conference on Machine Learning*, Denver, USA, 1995, pp. 331-339.

[6] W. R. Murray, "Practical approach to Bayesian student modeling", *Lecture Notes in Computer Science*, *Intelligent Tutoring Systems,* Springer Berlin, Heidelberg, 1998, pp.424-433.

[7] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar, and Amit Sheth, "Context-Aware Semantic Association Ranking", *In Proceedings of SWDB03,The first International Workshop on Semantic Web and Databases*, Berlin, Germany, 2003, pp. 33-50.

[8] Halaschek C, Aleman-Meza B, and Arpinar I B. "Discovering and ranking semantic associations over a large RDF metabase", *In Proceedings of the 30th VLDB Conference,* Toronto, Canada, 2004, pp. 1317-1320.