Modeling User's Cognitive Structure in Contextual Information Retrieval

Xuan Tian
tianxuan@ruc.edu.cn

Xiaoyong Du
duyong@ruc.edu.cn

He Hu
luckh2@163.com

Haihua Li
lihhjs@yahoo.com.cn

School of Information, Renmin University of China
Key Laboratory of Data Engineering and Knowledge Engineering, MOE
Beijing, 100872, China

Abstract

In contextual information retrieval, the retrieval of information depends on the time and place of submitting query, history of interaction, task in hand, and many other factors that are not given explicitly but implicitly lie in the interaction and surroundings of searching, namely the context. User's cognition is one of important contextual factors for understanding his or her personal needs. We propose a model called DOSAM to get user's individual cognitive structure on domain knowledge. DOSAM is developed from the spreading-activation model of psychology and is established on the domain ontology. The cost analysis of algorithm shows that it is feasible to get cognitive structure by DOSAM. Personalized search experimental results on digital library indicate that DOSAM can help improve the search effectiveness and user's satisfaction.

1. Introduction

Contextual Information Retrieval (CIR) has been brought forward and became one of research focuses in IR. The retrieval of information depends on the time and place of submitting query, history of interaction, task in hand, and many other factors that are not given explicitly but lie implicitly in the interaction and ambient environment, namely the context [7]. CIR tries to capture user's needs by augmenting the user's query with contextual information extracted from his or her searching process. For a user, the context within which he or she seeks information consists of cognitive, social and other factors related to his or her tasks, goals and intentions.

As far as cognition is concerned, it is involved in the acquisition and the use of knowledge. It consists of internal Cognitive Structure (CS) and Cognitive Behavior (CB) of knowing in brain [8]. CB is closely related to user's subjective response, such as feedback, experience, browsing response and so on. CS is different from CB. CS depicts a picture of the way in which the contents of cognition are organized in individual brain, namely the individual picture of knowledge.

Cognition has been considerably used for reference in IR. For example, CB was used for intelligent information retrieval interaction and personalized search as in [6, 1]; cognitive framework as a whole, including CS or user's domain knowledge, was taken into consideration in [2]; common knowledge was also widely applied to query expansion on the ground of general ontology such as WordNet[1] and ODP[2]. To the best of our knowledge, how to exploit user's individual CS to improve search has so far not been well addressed in the previous work.

In this paper, we propose a model called Domain-Ontology-based Spreading-Activation Model(DOSAM) to get user's individual CS on domain knowledge. DOSAM is developed from spreading-activation model of psychology, and its goal is to model user's individual CS on domain knowledge. Since spreading-activation model was introduced by Collins and Loftus in 1975 [4], it has been adopted in various fields. In essence, the effectiveness of spreading-activation model is crucially dependent on the availability of a representative node association map, and on the use of activation rules that can distinguish the useful nodes from the extraneous ones. In DOSAM, we bring special semantics to relationships between two concepts, develop semantic distances based on concrete semantics and introduce activated strength on concepts.

The rest of this paper is organized as follows. Section 2 gives the definitions about DOSAM and shows a feasible algorithm to build DOSAM. Section 3 presents the experimental results about personalized search based on DOSAM. Our conclusions and future work are presented in Section 4.

[1]WordNet. http://wordnet.princeton.edu.
[2]Open Directory Project. http://dmoz.org.

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

2. DOSAM

DOSAM is built on the domain ontology. In follows, we first introduce three definitions about domain ontology: *Domain Ontology* describes the organization of domain knowledge, which is the basis of user's individual CS on domain knowledge; *Semantic Relationship* gives a formalization of semantic relationship between two concepts; *Degree of Association* describes the semantic distance between any two concepts.

2.1. Definitions

Definition 1(Domain Ontology) Domain ontology is an explicit specification of a conceptualization about domain knowledge [5]. It can be described as $O = (C, R)$, where C is the set of concepts, and R is the set of semantic relationships between concepts.

Based on the latest standard of ontology description language recommended by W3C[3], we define semantic relationship as the follows.

Definition 2(Semantic Relationship) Let c_i and c_j be two concepts in domain ontology. If c_j is defined as "*equivalentClassof*" for c_i, we say c_i and c_j are semanticlly equivalent, namely $c_i \equiv c_j$(c_i is equivalent to itself, i.e. $c_i \equiv c_i$); if c_j is defined as "*subClassof*" for c_i, we say c_j is semantically contained by c_i, namely $c_j \subseteq c_i$ or $c_i \supseteq c_j$; if c_j is defined as "*propertyof*" class for c_i, we say c_j is semantically associated with c_i, namely $c_i \succ c_j$ or $c_j \prec c_i$.

A domain ontology can be viewed as a directed graph, where a concept is viewed as a node and the relationship between two concepts is viewed as a directed edge. The domain ontology can thus be represented as $OntoGraph = (C, R)$, where $< c_i, c_j >\in R(c_i, c_j \in C)$ if there is a kind of semantic relationship between c_i and c_j. According to different types of semantics, we associate a weight $\omega(c_i, c_j)$ with each edge in the graph, where $\omega(c_i, c_j)$ is given by equation(1).

$$\omega(c_i, c_j) = \begin{cases} 1, & \text{if } c_i \equiv c_j \ ; \\ \alpha, & \text{if } c_i \supseteq c_j \text{ or } c_i \subseteq c_j \ ; \\ \beta, & \text{if } c_i \prec c_j \text{ or } c_i \succ c_j \ ; \\ 0, & \text{otherwise} \ . \end{cases} \qquad (1)$$
$$\text{where } 0 \leq \beta < \alpha < 1 \ .$$

In equation(1), three different semantic relationship, equivalence relationship, super-sub relationship, and association relationship, are taken into consideration. $\omega(c_i, c_j)$ represents the associative degree implicit in each type of semantic relationship. Intuitively, the associative degree implicit in equivalence can be defined as 1, while the associative degree in super-sub is less than that in equivalence, and the associative degree in association relationship is less than that in super-sub relationship. Thus $0 \leq \beta < \alpha < 1$ is required. In practice, α and β may be different in various domain anthologies, and the advices of domain experts can be referred to when they are given. Castano gave his suggestions of different semantic relationships on tourism otology in [3].

[3] Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/.

According to Definition 2, there is $\omega(c_i, c_j) = \omega(c_j, c_i)$, therefore we get an undirected weighted graph from the original directed graph of domain ontology. This weighted domain ontology can be described as:
$WeightedOnto = (C, E, \omega)$, for $c_i, c_j \in C$, if $(c_i, c_j) \in E$, then $0 \leq \omega(c_i, c_j) \leq 1$.

Definition 3(Degree of Association(DOA)) Given a weighted domain ontology $WeightedOnto = (C, E, \omega)$, $c_i, c_j \in C$, the DOA between c_i and c_j is denoted by $DOA(c_i, c_j)$, which can be computed as in equation(2).

$$DOA(c_i, c_j) = \begin{cases} \omega(c_i, c_j), & \text{if } (c_i, c_j) \in E \ ; \\ \max_{(c_i, c_k) \in E} \{\omega(c_i, c_k) * DOA(c_k, c_j)\} & \text{else} \ . \end{cases} \qquad (2)$$

DOA describes the association between any two concepts in domain ontology. It reflects the intuition of the semantic distance between any two concepts, and provides a quantitative description for the association between two concepts.

By equation (2) we can know that the DOA of a concept with itself is 1, and the DOA of two different concepts is the maximum product of DOAs of two concepts along the paths between the two concepts.

According to the definition of DOA, we have to compute a complete graph from the given domain ontology. It's very expensive to construct this complete graph. In Section 2.2, we will analyze the cost and give a feasible algorithm for the computation.

To build user's CS, we define *Cognitive Center Concepts* to depict the center of his cognitive structure, and assign every concept in the domain ontology with the *Degree of Cognition*. Thus, given a threshold value, starting from Cognitive Center Concepts, activation can decide whether to spread to other related concepts according to their Degree of Cognition.

Definition 4(Cognitive Center Concept and Cognitive Center) The concept that user u gives to describe his attention on domain knowledge is called a cognitive center concept. The collection of cognitive center concepts is called the cognitive center V_u .

The cognitive center can be used to depict individual knowledge center in specific domain. For example, the V_u of a user u in economic domain is the set of concepts {macroeconomic, world economic, financial crisis}.

Definition 5(Degree of Cognition(DOC)) For a concept c_i in domain ontology, $DOC_u(c_i)$ is a real numbered weight given by a user to describe the extent of his knowing on it,$0 < DOC_u(c_i) \leq 1$.

Suppose DOC value λ_i has been given for every cognitive center concept c_i by the user, $0 \leq \lambda_i \leq 1$, then the DOC values of all concepts in domain ontology can be figured out by equation(3).

$$DOC_u(c_i) = \begin{cases} \lambda_i \ \text{if} \ c_i \in V_u \ ; \\ \max_{c_j \in V_u} \{DOC_u(c_j) * DOA(c_i, c_j)\} \ \text{if} \ c_i \in (C - V_u) \ . \end{cases} \quad (3)$$

For each concept c_i in cognitive center, its DOC value is λ_i. For the concept falling out of V_u, its DOC values can be computed according to the DOAs on the path to cognitive center concepts. Therefore, the DOC value of the concept excluded in V_u changes with the association with the cognitive center concepts, which can be seen as a spreading-activation process.

For a threshold value θ given by a user, $0 \leq \theta \leq \min(\lambda_i) \leq 1$, we can get the user's cognitive extension by pruning concepts within the domain ontology, and then obtain his CS by DOSAM.

Definition 6(Domain-Ontology-based Spreading-Activation Model(DOSAM)) $O = (C, R)$ is a domain ontology, its corresponding weighted representation is $WeightedOnto = (C, E, \omega)$; θ is a threshold value given by user u; V_u is the user's cognitive center. The user u's CS on O, i.e. DOSAM, is $O_u = (C', E')$, which is defined in equation(4).

$$\begin{array}{l} C' = \{c_j | DOC_u(c_j) \geq \theta\} \ ; \\ E' = \{(c_i, c_j) | (c_i, c_j) \in E, c_i \in C', c_j \in C'\} \ . \end{array} \quad (4)$$

For instance, as for the user in the example following definition 4, his cognitive center is defined as {macroeconomic, world economic, financial crisis}, and the corresponding degree of cognition is given as {1,1,0.9}. An economic ontology as EO, which is described in Section 3.1, is given, and θ is set as 0.67. Thus, according to the definition of DOSAM, there are 35 concepts, including 3 cognitive center concepts, in his CS on economics domain, and the minimal DOC value is 0.675(when $\alpha = 0.85$ and $\beta = 0.75$). Fig. 1 describes his CS for the concept 'Financial Crisis'.

Definition 7(Cognitive Extended Concept and Cognitive Extension) Given a user u, his DOSAM is $O_u = (C', E')$, and his cognitive center is V_u. If $c' \in (C' - V_u)$, c' is called a cognitive extended concept. All of cognitive extended concepts of user u are defined as his cognitive extension V'_u.

2.2. Constructing DOSAM

According to Definition 6(in the follows, we called the algorithm of definition 6 as Naive Construction Algorithm(NCA)), to build DOSAM, all of DOA values of any two concepts in domain ontology have to be computed. This would be very inefficient. If there are N concepts in domain ontology, and $\omega(c_i, c_j)$, $DOA(c_i, c_j)$ represent the elements of two matrices respectively (according to equation(3), they are both $N \times N$ matrices), the time complexity is $O(N^3)$ for the product of two matrices. In addition, because the length of any path between two concepts in domain ontology is $N - 1$ at most, the maximum time cost is $O(N^4)$. This cost leads NCA to be hardly realized in practice.

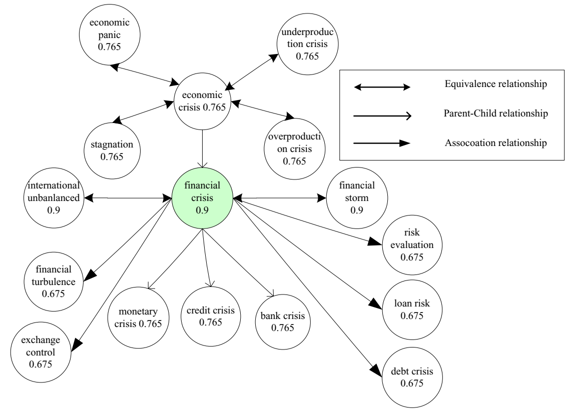

Figure 1. An example of CS on the concept 'Financial Crisis'.

In the following, we give a feasible Improved Construction Algorithm (ICA) of DOSAM. In ICA, not all DOA values need to be computed. The idea of ICA is: start from the cognitive center concepts, traverse every concept in domain ontology by breadth-first search, and put the concept whose DOC value is larger than the threshold value into DOSAM and as a new cognitive center concept. This goes on recursively until there are no new concepts added into DOSAM. Algorithm. 1 illustrates the ICA of DOSAM.

In ICA, a queue is used to retrieve the concepts in domain ontology. Whenever a concept is visited during the search, all of its unvisited neighbors will be visited recursively. But if the concept's DOC value is smaller than the threshold value, this concept and all its directed neighbors will be skipped. In the algorithm, every concept is visited at most once in the queue, and it has $N - 1$ neighbors at most, so the time complexity is $O(N^2)$. It is clear that ICA is more efficient than NCA defined in Definition 6.

3. Experimental Evaluation

3.1. Experimental Settings

We built DOSAM based on the Chinese Economic Domain Ontology of Renmin University of China (EO), and made use of the document resources of Personalized Digital Library System Version 2.0 of Renmin University of China (DLPers V2.0) as the test data sets, because a set of document resources in DLPers V2.0 have been preprocessed with economic labeling on the basis of EO. The current version of EO contains 9760 classes (concepts) and 15222 relations. It covers almost all of key concepts and relations of economics.

Fifty-two users of DLPers V2.0, consisting of both the students and the teachers of School of Economics of Renmin University of China, took part in our experiments. Every user provided the cognitive center concepts and the corresponding DOC values to generate his or her CS on economic domain, submitted 1-3 queries to DLPers V2.0, and then evaluated the search results.

Three experiments were designed to evaluate the effects of DOSAM. Firstly, 52 users' personalized search results based on DOSAM were evaluated as a whole. Secondly, 7 users' individual satisfactory rates for search results were investigated. At last, the search results of 10 queries based on DOSAM were compared with that of two traditional representation models, Keyword Vector Model(KVM) and Hierarchical Model(HM), since both KVM and HM could be used to represent individual cognitive structure too.

3.2. Experimental Measurements

In both the first and the third experiments, two measures were used. One was Precision@n(Precision at n retrieved documents), and the other was MAP@k(Mean Average Precision at k retrieved documents).We set n at 20 for Precision@n and k at 20 for MAP@k to exam user's satisfaction for top-20 retrieved documents.

In the second experiment, we wanted to investigate detailed individual satisfactory extent for search results based on DOSAM. Every user was required to rate each of the top-10 relevant results from 1 to 5 for his/her queries, 1 defining a very poor result with respect to their expectations and 5 a very good one. The statistical rates of users was selected as the measurement in this experiment.

3.3. Experimental Evaluation

In all three experiments, we set $\alpha = 0.85$ and $\beta = 0.75$ on the ground of both the intuition and the test experiences on EO, such as [3] recommended. In addition, considering that too many semantic equivalent relationships in EO would lead to overmany extended concepts, we set the DOA of semantic equivalent relationship to 0.99 as the approximation of 1 to limit the size of user's cognitive extension.

Experiment 1: Overall Evaluation on Personalized Search Effects. In experiment 1, search effects were evaluated as a whole. There were two runs here, one for personalized search based on DOSAM, and the other for naive search of DLPers V2.0. In the naive search, no personalized techniques were used. In each run, total 112 queries provided by 52 users (1 query at least and 3 queries at most for a user) were submitted. Search results were evaluated manually as relevant or non-relevant.

Table 1. The comparison of search effectiveness.

	Average #	Precision@20	MAP@20
Naive Search	8	21.58%	26.72%
DOSAM Search	18	62.37%	66.65%

Table 1 gives the comparison of search effectiveness of personalized search based on DOSAM with that of naive search of DLPers V2.0. We can see that search effectiveness based on DOSAM is greatly better than that of naive search of DLPers V2.0. From the improvement of MAP@20 of personalized search based on DOSAM, we can infer that the added relevant docs in the top-20 retrieved are almost all ranking ahead, and it is easy to understand since the query expansion based on DOSAM improves the relevance of retrieved documents with the query.

Experiment 2: Investigation of Individual Satisfactory Rates For Search results. In the second experiment, we wanted to investigate detailed individual satisfactory extent for search results based on DOSAM, because experiment 1 provided just an overview on search effectiveness. 7 users took part in this experiment. The measurement has been introduced in Section 3.2. Detailed statistical data of the 7 users was counted and collected in Table 2. According to the statistical data, users had 88% satisfactory degree on the top-10 relevant results on average.

Experiment 3: the Comparison of DOSAM with Other Two Representation Models. In the third experiment, we compared the effectiveness of DOSAM with that of KVM and HM. KVM and HM are always used to describe user profile. All three models were evaluated based on EO to avoid impartiality. We used "equivalentClassof" concepts to simulate synonymic concepts in KVM, and used "subClassof" and "superClassof" concepts to simulate hierarchical relationship in HM. 10 queries were selected randomly from experiment 1 to take part in this experiment. They were submitted to personalized search for three runs, one run on DOSAM, another run on KVM and the other on HM. From Table 3, we can see that for the same query, the improvement on KVM is more than double that on HM such as q1 and q7. As we observed, besides the influence of threshold value, the fact that more related concepts were involved in DOSAM than in HM led DOSAM to produce more exact and richer interpretation of user's personal needs.

Table 2. Detailed survey of individual satisfactory rates on search results based on DOSAM.

	u1	u2	u3	u4	u5	u6	u7	Avg-Value
Min rate	3	2	4	3	4	3	3	3.14
Max rate	5	5	5	5	5	5	5	5
Average rate	3.96	4.53	4.60	4.33	4.76	3.80	4.83	4.40

Table 3. The comparison on Precision@20 of three models.

	q0	q1	q2	q3	q4	q5	q6	q7	q8	q9	Avg-Value
KVM	35%	30%	35%	30%	40%	35%	20%	30%	35%	35%	32.5%
HM	50%	55%	45%	50%	65%	45%	50%	60%	55%	45%	52.0%
DOSAM	75%	70%	85%	65%	75%	65%	75%	75%	80%	75%	74.0%

4. Conclusion and Future Work

In contextual information retrieval, cognition is one of important contextual factors to understand user's personal needs. Cognition consists of internal structure and cognitive behavior of knowing in brain. In this paper, we proposed a model called DOSAM to get user's individual cognitive structure on domain knowledge. DOSAM is developed from the spreading-activation model of psychology, and its goal is to get the individual cognitive structure. Algorithm analysis indicates that it is highly efficient to get cognitive structure by DOSAM, and experimental results show that it is effective to apply cognitive structure produced by DOSAM in personalized search.

References

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR2006*, pages 19–26, New York, NY, USA, 2006. ACM Press.

[2] A. S. Belhassen, N. B. Abdallah, and H. H. B. Ghezala. Cognitive approach for building user model in an information retrieval context (poster session). In *SIGIR2000*, pages 337–338, Athens,Greece, 2000. ACM Press.

[3] S. Castano, A. Ferrara, S. Montanelli, and G. Racca. Matching techniques for resource discovery in distributed systems using heterogeneous ontology descriptions. In *ITCC'04 Volume 2*, page 360, Washington, DC, USA, 2004. IEEE Computer Society.

[4] A. M. Collins and E. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

[5] T. R. Gruber. A translation approach to portable ontology specifications. *knowledge Acquisition*, 5(2):199 – 220, 1993.

[6] P. Ingwersen. *Information Retrieval Interaction*. 1992.

[7] P. Ingwersen and N. Belkin. Information retrieval in context - IRiX. *SIGIR Forum*, 38(2), 2004.

[8] D. L. Medin and B. H. Ross. *Cognitive Psychology*. Harcourt Brace College Publishers, 1997.

Algorithm 1 the Improved Construction Algorithm (ICA) of DOSAM.

Input: User u's cognitive center V_u and $DOC_u(v_i)$ for every $v_i \in V_u$;
The threshold θ;
The Weighted domain ontology $WeightedOnto = (C, E)$.
Output: DOSAM $O_u = (C', E')$.
Begin

$InitQueue(Q)$;//initialize a queue as NULL
$V_u \rightarrow Q; V_u \rightarrow C'; null \rightarrow E'$;//center concepts enter the queue
while not eof(Q) **do**
  $DeleteQueue(Q) \rightarrow v_i$;//remove the header element
  for all adjacent edges of vertices v_i in E **do**
    $\text{the corresponding adjacent vertices} \rightarrow v'_j$;
    $DOC_u(v_i) \times \omega(v_i, v'_j) \rightarrow newDOC_u(v'_j)$;//find the adjacent vertices
    if $newDOC_u(v'_j) >= \theta$ **then**
      if $v'_j \in C'$ **then**
        $DOC_u(v'_j) \rightarrow oldDOC_u(v'_j)$;
        if $newDOC_u(v'_j) > oldDOC_u(v'_j)$ **then**
          $newDOC_u(v'_j) \rightarrow DOC_u(v'_j)$;
        end if
      else
        $EnterQueue(Q, v'_j)$;
        $AddVertices(v'_j, C')$;
        $newDOC_u(v'_j) \rightarrow DOC_u(v'_j)$;
      end if
      if not $(v_i, v_j) \in E'$ **then**
        $AddEdges((v_i, v'_j), E')$;
      end if
    end if
  end for
end while
Return $O_u = (C', E')$;

End.