An Effective Semantic Search Technique using Ontology

Jihyun Lee
Korea Advanced Institute of Science and Technology
Yuseong-gu, Guseong-dong
Daejeon, Republic of Korea, 305-701
hyunlee@islab.kaist.ac.kr

Jun-Ki Min
Korea University of Technology and Education
Byeongcheon-myeon
Chungnam, Republic of Korea, 330-708
jkmin@kut.ac.kr

Chin-Wan Chung
Korea Advanced Institute of Science and Technology
Yuseong-gu, Guseong-dong
Daejeon, Republic of Korea, 305-701
chungcw@islab.kaist.ac.kr

ABSTRACT

In this paper, we present a semantic search technique considering the type of desired Web resources and the semantic relationships between the resources and the query keywords in the ontology. In order to effectively retrieve the most relevant *top-k* resources, we propose a novel ranking model. To do this, we devise a measure to determine the weight of the semantic relationship. In addition, we consider the number of meaningful semantic relationships between a resource and keywords, the coverage of keywords, and the distinguishability of keywords. Through experiments using real datasets, we observe that our ranking model provides more accurate semantic search results compared to existing ranking models.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Measurement

Keywords: Semantic Search, Ranking, Semantic Relationship, Ontology, Semantic Web

1. INTRODUCTION

It is common that the current keyword-based search misses highly relevant pages and returns a lot of irrelevant pages for user requests, since the keyword-based search is ignorant of the type of resources a user wants to get and the semantic relationships between the resources and keywords.

In order to effectively retrieve the most relevant *top-k* resources in searching in the Semantic Web, [3, 5] propose ranking models using the ontology which presents the meaning of resources and the relationships among the resources. These works determine the relevance based on the link analysis where the amount of relationships and the specificity of the relationships are considered, but it is insufficient for precise ranking. The ranking model in [5] does not consider the diversity of semantic relationships. The work in [3] depends on domain experts to determine the weights of all relationships in the instance level. Also, the query consisting of multiple keywords with different importance is not effectively handled.

In this paper, we propose an effective semantic search technique considering the diversity of semantic relationships. We propose a measure for weighting a semantic relationship. Based on this, we suggest a novel ranking model.

2. PRELIMINARIES

In this section, we present some definitions for our work.

Definition 1. **(Schema)** Schema S is defined as <*C*, *D*, *P*>. C is the set of classes, D is the set of data types, and P is the set of properties. For $d \in C$, $r \in C \cup D$, there can be a property $p(d, r) \in P$. For each property $p(d, r) \in P$, d is the *domain* and r is the *range* of p. $p^{-1}(r, d)$ is the inverse property of $p(d, r)$.

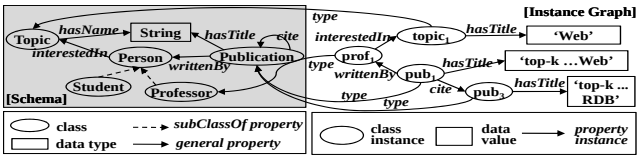

Figure 1: An example of ontology

Definition 2. **(Instance Graph)** An instance confined to a schema *S* = <*C*, *D*, *P*> is defined as a directed graph *G* = <*V*, *E*>. V is the set of resources. $[c]$ indicates the set of instances of $c \in C \cup D$. For each $v \in$ V, $v \in [c]$ if $v.type = c$. E is the set of property instances. $[p(d, r)]$ denotes the set of instances of $p(d, r) \in P$. For each $e(v_i, v_j) \in E$, $e(v_i, v_j) \in [p(d, r)]$ if $e = p$, $v_i \in [d]$ and $v_j \in [r]$. v_i is the subject and v_j is the object of e.

Definition 3. **(Semantic Path)** A semantic path *sp* is a sequence of properties $p_1(d_1,r_1)$ ... $p_m(d_m, r_m)$ in a schema *S* = <*C*, *D*, *P*> where $p_i(d_i, r_i) \in P$ and r_i and d_{i+1} are the same class or have the same ancestor class (excluding the root) in the class hierarchy.

An instance graph G confined to schema S can include matches of a semantic path in S. A match of the semantic path is a sequence of instances of properties, which is called *semantic path instance*. For example, in Figure 1, *writtenBy*$^{-1}$*(Professor, Publication) hasTitle(Publication, String)* is a semantic path, and *writtenBy*$^{-1}$ *(prof*$_1$*, pub*$_1$*) hasTitle(pub*$_1$*,'top-k...Web')* is a semantic path instance of the semantic path.

Definition 4. **(Semantic Search)** Given an instance graph G, the semantic search is to find an answer A for query *Q* = <*T*, *K*>. For each resource $a \in A$, there should be at least one semantic path instance from resource a to data value s in G where $a \in [T]$ and value s contains $k \in K$. $IP(a, k)$ denotes the set of such semantic path instances from a to s including keyword k.

3. WEIGHTING FOR SEMANTIC PATH

A semantic path consists of one or more properties. Therefore, in order to measure the weight of a semantic path, we should be able to determine the weight of each property in the semantic path. The weight of a property is determined according to (1) the amount of information contained in the property, and (2) the characterizability between domain and range of the property.

In the information theory, the information content of an event x can be quantified by the occurrence probability of x. As x occurs rarely, x has more information. Based on this, the amount of information contained in a property $p(d, r)$ is computed by

$$I(p(d, r)) = -\log_2 pr(p(d, r)) \quad (1)$$

where $pr(p(d, r))$ is the probability that a resource is a subject of the property $p(d, r)$ in an instance graph.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

In the information theory, the amount of information that one random variable contains about another random variable is measured by Mutual Information [1]. To measure the characterizability between the domain and the range, we adapt the mutual information. The mutual information between the domain d and the range r for a property $p(d, r)$ is computed by

$$MI(p(d,r)) = \sum_{o \in r} \sum_{s \in d} pr(s,o) \cdot \log_2(\frac{pr(s,o)}{pr(s)pr(o)}) \quad (2)$$

where the sample space is $[p(d, r)]$, $pr(s)$ is the probability that $e \in [p(d, r)]$ has s as its subject and $pr(o)$ is the probability that $e \in [p(d, r)]$ has o as its object. Also, $pr(s, o)$ is the probability that $e \in [p(d, r)]$ has s and o as its subject and object at the same time.

By using Equation (1) and (2), we can compute the weight of a property $p(d, r)$ as follows:

$$w(p(d,r)) = \alpha \cdot I(p(d,r)) + \beta \cdot MI(p(d,r)) \quad (3)$$

where $0 < \alpha, \beta < 1$ and *I* and *MI* are normalized to be in the same range [0, 1].

As the length of a semantic path gets longer, the relevance between the source and the destination decreases. Thus, considering the loss of the relevance due to the growing of the length of a semantic path, the weight of the semantic path sp is computed as follows:

$$W(sp) = (\prod_{p(d,r) \in sp} w(p(d,r))) \cdot \delta^{length(sp)-1} \quad (4)$$

where $length(sp)$ indicates the number of properties in sp, and δ is an attenuation parameter which is tunable from 0 to 1.

4. RANKING FOR SEMANTIC SEARCH

In the novel ranking model proposed in this paper, the following three relevance criteria are considered.

- **The number of meaningful semantic path instances:** We regard resources which have many meaningful semantic path instances directed to keywords as more relevant resources. The meaningfulness of the semantic path instance is represented by the weight of the semantic path obtained by Equation (4). The relevance of a resource a for a keyword k_i is computed by $R(a, k_i) = \sum_{ip \in IP(a,k_i)} W(ip)$.
- **The coverage of keywords:** A user prefers results covering all keywords. To reflect this preference, we apply the extended boolean model [4] to our ranking model. A resource is mapped to a point in an n-dimensional space $[0, 1]^n$ where n is the number of keywords. Each dimension represents the relevance of a resource to the corresponding keyword. The relevance of a resource a is in inverse proportion to the distance from the ideal position $[1, ..., 1]$ to the point of a.
- **The distinguishability of keyword:** A resource having semantic paths to distinguishable keywords is more relevant than a resource having semantic paths to undistinguishable keywords. The distinguishability of keyword k_i, $D(k_i)$, is the inverse of the ratio of data values containing k_i to entire data values in the instance graph.

Based on the above mentioned three factors, the relevance score of resource a for keywords K is computed by

$$Rank(a,K) = 1 - [\frac{\sum_{1 \le i \le |K|} (D(k_i) \cdot (1 - NR(a,k_i)))^p}{\sum_{1 \le i \le |K|} D(k_i)^p}]^{\frac{1}{p}} \quad (5)$$

First, $NR(a, k_i)$ is the normalized $R(a, k_i)$ in the range [0, 1], which reflects the the number of meaningful semantic path instances. Second, the L_p-norm distance form of the above model reflects the coverage of keywords, and p (≥ 1) controls the strength of AND-semantics among keywords. Finally, the distinguishability of keyword, $D(k_i)$, is utilized as the weight of a keyword reflecting the importance of the keyword in the model.

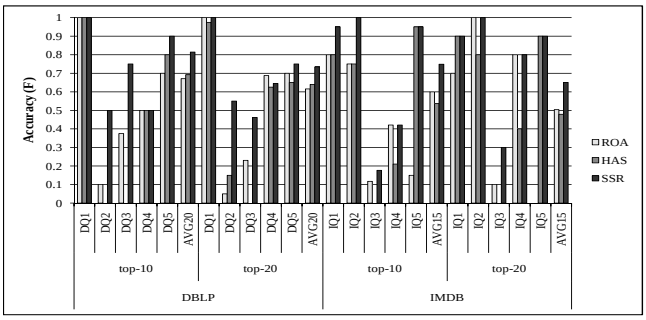

Figure 2: The accuracy for *top-k* results (k = 10 and 20)

5. SEMANTIC SEARCH PROCESSING

Our semantic search for query $< T, K >$ is performed according to the following four procedures: 1) Extract the set of semantic paths SP from T to K in a schema. 2) Find the set of resources R each of which reaches some keywords in K through SP in an instance graph. 3) For each resource $r_i \in R$, compute $Rank(r_i, K)$. 4) Provide r_i in the descending order of $Rank(r_i, K)$.

6. EXPERIMENTS

In the experiments, we evaluate the accuracy of our ranking model, Semantic Search Rank (*SSR*), in comparison with existing methods *RQR* [5] and *HAS* [3]. We use two kinds of ontologies from two real datasets: DBLP (http://dblp.uni-trier.de/xml/) and IMDB (http://www.imdb.com/interface). Figure 2 shows the accuracy of three ranking models for five representative *top-k* queries over each ontology. We use F-measure (= $\frac{2*precision*recall}{precision+recall}$) [2] to measure the accuracy. Also, it includes the average of the accuracy for 20 queries over DBLP ontology (i.e,. AVG20) and 15 queries over IMDB ontology (i.e., AVG15). The tunable parameters of *SSR* in our experiments are set as follows: α = 0.2 and β = 0.8 in Equation (3), δ = 0.6 in Equation (4), and p = 3 in Equation (5). As we can observe from the experimental results, *SSR* outperforms the existing ranking models in general.

7. CONCLUSION

In this paper, for an effective semantic search, we proposed a new ranking model considering the diversity of semantic relationships and the coverage of keywords with various importance to determine the relevance. The experimental results showed that the accuracy of our new ranking model was better than those of other raking models using the ontology.

Acknowledgment

This research was supported by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the Institute of Information Technology Advancement. (grant number IITA-2008-C1090-0801-0031)

8. REFERENCES

[1] T. M. Cover. *Elements of Information Theory*. John Wiley and Sons, INC., New York, NY, USA, 1991.

[2] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[3] C. Rocha, D. Schwabe, and M. P. de Aragao. A Hybrid Approach for Searching in the Semantic Web. In *Proc. of WWW*, pages 374–383, May 2004.

[4] G. Salton, E. A. Fox, and H. Wu. Extended Boolean Information Retrieval. *Communications of ACM*, 26(12):1022–1036, December 1983.

[5] N. Stojanovic, R. Studer, and K. Stojanovic. An Approach for the Ranking of Query Results in the Semantic Web. In *Proc. of ISWC*, pages 500–516, October 2003.