Association Search in Semantic Web: Search + Inference

Liang Bangyong
Deparment of Computer Science,
Tsinghua University
Beijing, 100084
China
861062789831
liangby97@mails.tsinghua.edu.cn

Tang Jie
Department of Computer Science,
Tsinghua University
Beijing, 100084
China
861062789831
j-tang02@mails.tsinghua.edu.cn

Li Juanzi
Department of Computer Science,
Tsinghua University
Beijing, 100084
China
861062781461
ljz@keg.cs.tsinghua.edu.cn

ABSTRACT
Association search is to search for certain instances in semantic web and then make inferences from and about the instances we have found. In this paper, we propose the problem of association search and our preliminary solution for it using Bayesian network. We first minutely define the association search and its categorization. We then define tasks in association search. In terms of Bayesian network, we take ontology taxonomy as network structure in Bayesian network. We use the query log of instances to estimate the network parameters. After the Bayesian network is constructed, we give the solution for association search in the network.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods -- Relation systems, Semantic networks

General Terms
Algorithms, Measurement, Performance, Languages.

Keywords
Ontology, Bayesian Network, Knowledge Management, Inference

1. INTRODUCTION
Currently, there are a few semantic search systems in semantic web which search entities in ontology. Semantic Search[1] searches the objects instead of texts. The search results are instances in the knowledge base. Swoogle[2] searches the knowledge base using keywords, the results are URIs of entities in knowledge base including concepts, properties and instances. This kind of search can be viewed as variation of the traditional search in the semantic web.

The associations among instances may be direct or indirect. The direct association means that two instances are directed connected with a property. The indirect association means that two instances are connected by a set of instances and properties. Such associations are very helpful for users to understand the search results, especially on semantic web.

Consider the domain of computer science department which contains the concept "Professor", "lab", "project" and "Student". A search with two keywords may return the instance "Jack" which belongs to the concept "Professor" and the instance "Jerry" which belongs to the concept "Student". Actually it's just what current semantic web search systems do. After these two instances are retrieved, users may want to know their associations. If fortunately, they have direct associations like the following triple:

Jack hasStudent Jerry

Most of existent semantic search systems success in finding the associations between them. But the association is often not directed. Current semantic search systems will give no associations between the two instances in this situation. Looking into the ontology, we may find actually there are two indirect associations:

1) Jack isInProject "search engine"; Jerry isInProject "search engine". Jack and Jerry can be associated via "search engine".
2) Jack inSameLabWith Mike; Mike hasStudent Eric; Eric isMemberOf "football team"; Jerry isMemberOf "football team". Jack and Jerry can be associated via the instances Mike, Eric and "football team".

2. Problem Statement
Now we formally define the association search problem that we are solving.

We first give the definition of knowledge base in our scenario. A knowledge base can be viewed as a three tupel:

$$KB = (I, C, P)$$

where C denotes the set of concepts; P denotes the set of property; I denotes the instance set of all concepts. Specially, let $c_i \in C$ denote a concept, $p \in P$ denote a property and $i_i \in I$ denote an instance of concept c_i.

We now define the **Association Search** on the semantic web.

Definition 1. Given a knowledge base KB and a user's query q is represented by a set of keyword $q=\{k_i\}$, the task of association search is to:

(1) Search for instances I_q related to q. Let $I_{ki}=\{i_{ki}\}$ denote the set of instances that whose URIs contain the keyword k_i as one of their sub strings. We have $I_q = I_{k1} \cup ... \cup I_{ki} \cup ...$

(2) Infer the association A between i_{ki} and i_{kj}, where $i_{ki} \in I_{ki}$ and $i_{kj} \in I_{kj}$.

- If instance i contains both k_i and k_j, then we define the association between k_i and k_j as **Null Association** (also denoted as $A_{\emptyset}=(i, null, i)$) for the given instance i;

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

- If instances i_{ki} and i_{kj} are related by a property p, then we define the association between k_i and k_j as **Direct Association** (also denoted as A_D= (i_{ki}, p, i_{kj})) for the given instances i_{ki} and i_{kj};
- If instances i_{ki} and i_{kj} do not have any direct relation, then our target is to find the **Indirect Association** between i_{ki} and i_{kj} (denoted as a set A_Φ={(i_{ki}, p_a, i_j), {(i_i, p_m, i_j)}, (i_j, p_b, i_{kj})}), which constructs a relation path from i_{ki} to i_{kj} via **Intermediate Association** A_I={(i_i, pi, i_j)}.

(3) Rank all possible associations for the query q. Given all possible associations (including Null Association, Direct Association, Indirect Association), represent them to the user by a ranked list according to their "relevance" to the query.

It has been carefully studied that the average number of keywords used in web search is 2.35 [3]. Thus, to facilitate the illustration, we focus on the scenario of two keywords in association search in the rest of the paper. Generalizing our approach to multiple keywords is straightforward.

3. Our Approach

Here we present our approach for association search. Concerning the query with two keywords, i.e. q={k_1, k_2}, we concentrate on how to do an association search. Association search needs to aware the graph structure of the domain knowledge. Bayesian network is a compact graphical representation of joint probability distributions. It permits the explicit encoding of conditional independencies in a natural manner. Thus, Intuition shows that Bayesian network can be useful for association search. A Bayesian network consists of two components – the graph structure and the parameters [4].

In our case, ontology on the semantic web naturally provides the graph structure for Bayesian network. Hence, our focus is the parameter estimation and inference on it. We first present how to model our problem by Bayesian network, and then simply describe the parameter estimation, finally illustrate how the association search is realized in Bayesian network.

3.1 Bayesian Network based Association Search Model

In terms of Bayesian networks, given two instances i_1 and i_2, the association from i_1 to i_2 can be defined as p(i_1|i_2).

If the probability is zero, it means that there is no association from i_1 to i_2. If the value of the probability is not zero, it means that there are one or more associations from i_1 to i_2.

3.2 Association Search in Bayesian Network

After the Bayesian network of the domain is constructed, the association search can be performed. The association search includes the following aspects: searching the instances by keywords, association finding between two instances and calculating association scores.

The traditional web search technology is used to perform the instance search. An instance is processed as a document. The search takes the keywords {k_1, k_2} and searches for I_1={i_{k1}} (containing k_1) and I_2={i_{k2}} (containing k_2) in the knowledge base. Instead of returning a document list, it returns the instance list.

The association finding includes two steps. Given two instances i_{k1} and i_{k2}, if they are the same instance, which means there is a Null Association between them, we quantify the relevance by $P(i_{k1})$. Otherwise, we quantify its relevance by $P(i_{k2}|i_{k1})$. $P(i_{k2}|i_{k1})$ can be rewritten as:

$$P(i_{k2} \mid i_{k1}) = \frac{p(i_{k2}, i_{k1})}{p(i_{k1})}$$

For direct association, computing its score is straightforward.

For indirect association with the intermediate association A_I={(i_i, pi, i_j)}, we can compute $P(i_{k2},i_{k1})$ by:

$$P(i_{k2}, i_{k1}) = \sum_{i_j, \Lambda, i_i} P(i_{k2}, i_j, \Lambda, i_i, i_{k1})$$
$$= \sum_{i_j, \Lambda, i_i} \Lambda \sum P(i_{k2} \mid i_j) \Lambda\ P(i_i \mid i_{k1}) P(i_{k1})$$

And compute $P(i_{k1})$ by

$$P(i_{k2}) = \sum_{i_j, \Lambda, i_i, I_{k1}} P(i_{k2}, i_j, \Lambda, i_i, I_{k1})$$
$$= \sum_{i_j, \Lambda, i_i, I_{k1}} \Lambda \sum P(i_{k2} \mid i_j) \Lambda\ P(i_i \mid I_{k1}) P(I_{k1})$$

where I_{k1} is the set of instance of c_i, $P(I_{k1})$ denotes $\forall i_{kx} \in I_{k1}, P(i_{kx})$.

We are currently under the development of the presented approach for software search. We plan to represent the search results by visualized path (corresponding to association) using JUNG (http://jung.sourceforge.net), a graph render engine. One of versions of our work is available at http://keg.cs.tsinghua.edu.cn/project/pswmp.htm.

4. Conclusion

Semantic search is becoming one of the most crucial challenges on semantic web. The inherence of semantic web determines that to directly adapt the traditional search to semantic web is not so reasonable. In this paper, we propose the problem of association search on semantic web for the first time. We think that association search should be one important kind of semantic search. We give a complete definition of association search from three aspects. We also present our preliminary solution for the problem by using Bayesian network. Specially, we describe how we construct the Bayesian network structure by ontology, how we learn the parameters in the Bayesian model and how we use the network to infer the association.

REFERENCES

[1] Guha, R.V., McCool, R. and Miller, E. Semantic Search: Proceedings of the twelfth international conference on World Wide Web (WWW2003), ACM Press, 2003

[2] Ding, L., Finin, T.W., Joshi, A., Pan, R., Scott Cost, R., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. CIKM 2004: 652-659

[3] Oyama, S. Query Refinement for Domain-Specific Web Search. PHD thesis, 2002.

[4] Tong, S. Active Learning: Theory and Applications. PHD thesis, Augest, 2001.

[5] Kowalski, R. How to be Artificially Intelligent – the Logical Way. http://www-lp.doc.ic.ac.uk/UserPages/ staff/rak/, 2005.