Enriching Concept Search across Semantic Web Ontologies

Chetana Gavankar[1,2,3], Vishwajeet Kumar[2],
Yuan-Fang Li[3], and Ganesh Ramakrishnan[2]

(1) IITB-Monash Research Academy, Mumbai, India
(2) IIT Bombay, Mumbai, India
(3) Monash University, Melbourne, Australia

Abstract. Semantic Web ontologies are fast-growing knowledge sources on the Web. Searching relevant concepts from this large repository is a challenging problem. The current Semantic Web search engines provide either (1) coarse-grained search over ontologies or (2) very fine-grained search over individuals. We believe searching and ranking concepts across ontologies provides an ideal granularity for certain tasks such as ontology population and web page annotation. Towards this objective, we propose a novel approach of indexing concepts using ontology axioms in an inverted file structure and ranking them using a dynamic ranking algorithm. Our proposed method is generic and domain-independent. A preliminary evaluation indicates that our proposed method is effective, outperforming the search function of BioPortal, a large and widely-used ontology repository.

Keywords: Semantic Web, Ontologies, Concept Search, Indexing

1 Motivation

The current breed of semantic web search engines can be broadly grouped into 2 categories: (1) those that search over ontologies, and (2) those that search over individual resources. The former may be too coarse-grained as a large ontology may contain hundreds of thousands or even millions of concepts. On the other hand, the latter approach may be too fine-grained – many resources may be relevant and returning them individually may not be the best approach. We describe an approach of retrieving relevant concepts from semantic web ontologies. We propose a novel technique of indexing concepts using axioms in ontologies. Our system supports semi-structured queries where names of concepts and relevant properties can be specified.

2 Related Work

Semantic search engines such as Sindice [1], Swoogle [2, 3], Falcon [4, 5], SWSE [6] provide semantic web search engine interface. They provide search over coarse-grained ontology level and fine-grained resources [7] on the semantic web. We provide search at concept level with middle level granularity. SchemEX [8] is

stream based approach and tool for real time indexing and schema extraction of LOD data. Hu, Bo et. al [9] use information retieval tfidf for indexing the ontology documents. Semplore [10] use standard IR style indexing for semantic web content and textual information. In comparison we build index using context information around concept that makes it easy to search for relevant concept along with all its context information. The current work semantic web resources ranking is by adapting and modifying pagerank algorithm used in classical search engines. ReConRank [11], TripleRank[12] adapt Pagerank/HITS [13] algorithm for semantic web data. Our ranking function is parameterized using context features.

3 System Architecture

Our interface provides keyword query input as well as allows to select contextual information around concepts in an ontology corpus. Given a concept in an ontology, all its contextual features are indexed using an inverted file structure. Such features include the concept's label, ID, URI, synonyms, data and object properties used in axioms about the concept, sub classes, super classes, equivalent classes. This approach enriches concept search by disambiguating a concept from those with similar names. For example, if *heart* concept is searched in context of *diseases* using our approach, results related to diseases of heart will be ranked higher, while results in other contexts such as *functionality* will be ranked lower. We now explain the ranking algorithm based on contextual features.

Let α, β, γ represent weights of concept label, data properties ($i = 1$ to m) and object properties ($j = 1$ to n) of the concept respectively. Let δ represent weights of context features ($k = 1$ to t) like synonyms, provenance of the concept. The weights α, β, γ and δ are currently are assigned values based on heuristics. In future we plan to learn these weights using machine learning algorithms. The weight of concept c in the ontology corpus, denoted W_c, is calculated as follows:

$$W_c = \lambda.[\alpha + \beta.\sum_{i=1}^{m} i + \gamma.\sum_{j=1}^{n} j + \delta.\sum_{k=1}^{t} k] \tag{1}$$

$$\lambda = \begin{cases} 1 & \text{if exact match} \\ similarity(x, y) & \text{where } x \text{ and } y \text{ represent 2 strings} \end{cases} \tag{2}$$

4 Evaluation

For evaluation purposes we compare our system[1] with the search function on BioPortal,[2] a large and widely-used biomedical ontology repository. In our experiment a large portion of ontologies, 252 out of 348 in total, were downloaded from BioPortal and indexed. Together these ontologies contain more than 660,000 classes.

[1] Available at `http://qassist.cse.iitb.ac.in/LOD/`

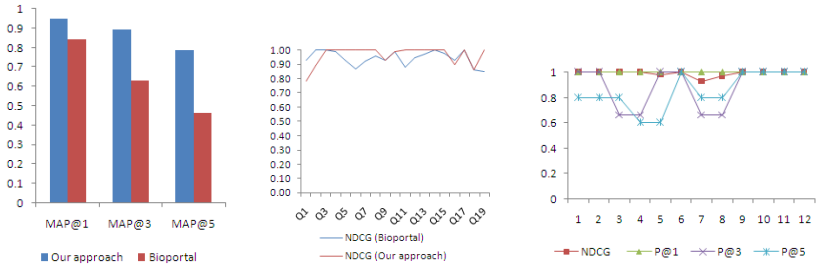
[2] `http://bioportal.bioontology.org/`

Algorithm 1: Ranking Algorithm

Data: Query Tokens $Q = Q_c, Q_{d_1}, ..Q_{d_m}, Q_{o_1}, ...Q_{o_n}, Q_{f_1}, ..Q_{f_t}$, Concepts $C = C_1, C_2...C_n$
Result: Weight of Concept W_c

$\alpha \leftarrow 0$, $\beta \leftarrow 0$, $\gamma \leftarrow 0$, $\delta \leftarrow 0$, $W_c \leftarrow 0$;
foreach *element* $C_i \in C$ **do**
  if $sim(Q_c, label(C_i)) > 0$ **then**
    $\alpha \leftarrow \alpha + \lambda$
    foreach *data property of* C **do**
      for *i=1 to m* **do**
        if $sim(Q_{d_i}, dp(C_i)) > 0$ **then**
          $\beta \leftarrow \beta + \lambda$
    foreach *object property of* C **do**
      for *j=1 to n* **do**
        if $sim(Q_{o_j}, op(C_m) > 0$ **then**
          $\gamma \leftarrow \gamma + \lambda$
    foreach *context feature of* C **do**
      for *i=k to t* **do**
        if $sim(Q_{f_k}, feature(C_m) > 0$ **then**
          $\delta \leftarrow \delta + \lambda$
    $W_c = [\alpha + \beta + \gamma + \delta]$

Two metrics widely-used in information retrieval, normalized discounted cumulative gain (NDCG) and mean average precision (MAP) [14], were used to measure the effectiveness of our approach viz-a-viz BioPortal search across 20 queries. Figure 1 (a) and (b) depict MAP and DNCG results for queries that do not contain property information as contextual features. It can be seen from Figure 1 (a) that our system outperforms BioPortal for MAP. Figure 1 (b) shows that the NDCG values are comparable for the two systems. For queries that contain property information, BioPortal fails to return search results. The results for queries with property information in Figure 1 (c) depict high precision and NDCG values.



(a) without property infomation

(b) without property infomation

(c) with property infomation

Fig. 1. Preliminary evaluation results.

5 Conclusion and Future work

Semantic Web search is primarily divided into two types - one which allows keyword query capability and other which needs SPARQL query input. The latter gives exact results due to precise input queries. This requires user to have

technical knowledge about writing a SPARQL query. We present an approach of searching for concepts using semistructured keyword queries that incorporates *contextual features* to improve precision. A preliminary evaluation and a comparison with BioPortal's search function shows the effectiveness of our system. In future we will investigate the incorporation of ontology reasoning to include implicit contextual features. Currently the ranking algorithm derives feature weights heuristically. Going ahead we will learn the weights using machine learning methods. In addition to enriched concept search, our further work will also include property search across ontologies.

References

1. Tummarello, G., Delbru, R., Oren, E.: Sindice.com: Weaving the open linked data. In: ISWC/ASWC. (2007) 552–565
2. Finin, T., Peng, Y., Scott, R., Joel, C., Joshi, S.A., Reddivari, P., Pan, R., Doshi, V., Ding, L.: Swoogle: A search and metadata engine for the semantic web. In: In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, ACM Press (2004) 652–659
3. Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P.: Finding and ranking knowledge on the semantic web. In: Proceedings of the 4th International Semantic Web Conference. (2005) 156–170
4. Qu, Y., Cheng, G.: Falcons concept search: A practical search engine for web ontologies. IEEE Transactions on Systems, Man, and Cybernetics, Part A **41**(4) (2011) 810–816
5. Cheng, G., Ge, W., Qu, Y.: Falcons: searching and browsing entities on the semantic web. In: World Wide Web Conference Series. (2008) 1101–1102
6. Hogan, A., Harth, A., Umrich, J., Kinsella, S., Polleres, A., Decker, S.: Searching and browsing linked data with swse: the semantic web search engine. Web Semantics: Science, Services and Agents on the World Wide Web **9**(4) (2011)
7. Blanco, R., Mika, P., Vigna, S.: Effective and efficient entity search in rdf data. In: International Semantic Web Conference (1). (2011) 83–97
8. Konrath, M., Gottron, T., Staab, S., Scherp, A.: Schemex - efficient construction of a data catalogue by stream-based indexing of linked data. J. Web Sem. **16** (2012)
9. Hu, B., Croitoru, M., Dasmahapatra, S., Lewis, P., Shadbolt, N.: Indexing ontologies with semantics-enhanced keywords. In: Proceedings of the 4th international conference on Knowledge capture. K-CAP '07, ACM (2007) 119–126
10. Wang, H., Liu, Q., Penin, T., Fu, L., Zhang, L., Tran, T., Yu, Y., Pan, Y.: Semplore: A scalable IR approach to search the Web of Data. Journal of Web Semantics **7** (2009) 177–188
11. Hogan, A., Harth, A., Decker, S.: Reconrank: A scalable ranking method for semantic web data with context. In: 2nd Workshop on Scalable Semantic Web Knowledge Base Systems. (2006)
12. Franz, T., Schultz, A., Sizov, S., Staab, S.: Triplerank: Ranking semantic web data by tensor decomposition. In: International Semantic Web Conference. (2009) 213–228
13. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of The ACM **46** (1999) 604–632
14. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. An Introduction to Information Retrieval. Cambridge University Press (2008)