Generating Summaries for Ontology Search

Gong Cheng
petercheng456@gmail.com
Weiyi Ge
geweiyi@gmail.com
Yuzhong Qu
yzqu@nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China

ABSTRACT

This poster proposes a novel approach for generating summaries for ontology search. Following previous work, we define ontology summarization as the problem of ranking and selecting RDF sentences, for which we examine three aspects. Firstly, to assess the salience of RDF sentences in an ontology, we devise a bipartite graph model for representing the ontology and analyze random walks on this graph. Secondly, to reflect how an ontology is matched with user needs expressed via keyword queries, we incorporate query relevance into the selection of RDF sentences. Finally, to improve the unity of a summary, we optimize its cohesion in terms of the connections between constituent RDF sentences. We have implemented an online prototype system called Falcons Ontology Search.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Markov processes*; H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors, human information processing*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*abstracting methods*

General Terms

Algorithms, Human Factors

Keywords

Cohesion, ontology summarization, query relevance, random walk, ranking

1. INTRODUCTION

When building a Web application, reusing an existing ontology could not only facilitate domain modeling but also place the application on a "semantic bus" where different applications easily interchange the meaning of their semantically homogeneous data. For systems that support ontology reuse such as an ontology search engine, a key feature is how to assist users in finding relevant ontologies efficiently, given that a returned ontology may define a great many term (i.e. class and property) specifications. To this end, ontology summarization [2] has been proposed to extract a salient part of an ontology to enable fast investigation. In this work, we propose a new method for computing salience. Further, we combine it with two other metrics, namely query relevance and cohesion, to form a solution to generating summaries for ontology search. The solution has been applied to an ontology search engine called Falcons Ontology Search (`http://ws.nju.edu.cn/falcons/ontologysearch/`).

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

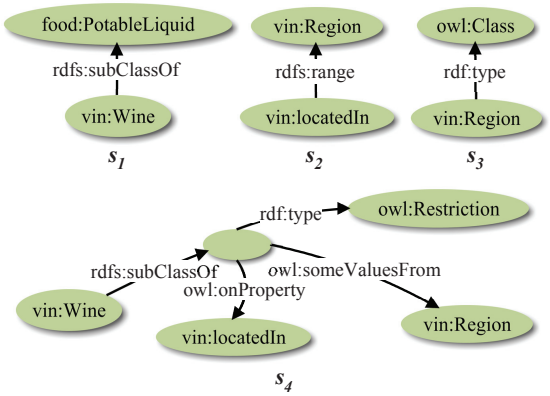

Figure 1: Four RDF sentences.

2. PROBLEM STATEMENT

Following [2], the RDF graph representation of an ontology has a unique finest partition that satisfies: two RDF triples that share common blank nodes are in the same part. Each part, being an RDF graph by itself, is called an *RDF sentence*. For example, Fig. 1 illustrates four RDF sentences. Then a summary of an ontology is defined as a subset of RDF sentences subject to a size constraint in terms of the total number of their constituent RDF triples.

In this work, we consider ontology summarization as an optimization problem by introducing an objective function that characterizes the *goodness* of a summary S of an ontology o w.r.t. a keyword query Q:

$$\begin{aligned}\text{Goodness}(o,S,Q) \triangleq &(1-\alpha-\beta)\cdot\text{Salience}(o,S)\\ &+\alpha\cdot\text{Relevance}(S,Q)\\ &+\beta\cdot\text{Cohesion}(S)\,,\end{aligned} \quad (1)$$

in which $\alpha,\beta,\alpha+\beta\in[0,1]$ are weighting coefficients. This function linearly combines three aspects, which will be detailed in the next section.

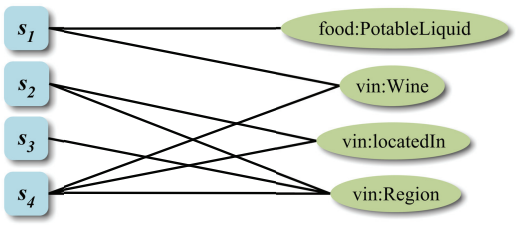

Figure 2: A sentence-term graph.

3. METRICS

In this section, we elaborate on the metrics we leverage to rank and select RDF sentences into a query-biased summary of an ontology for being used in the search scenario.

3.1 Salience

In an ontology, an RDF sentence describes one or more terms; a term is described by one or more RDF sentences. These lead to a bipartite graph that models such *description* relationship between RDF sentences and terms, which we call *sentence-term graph* (STG). Figure 2 illustrates an STG induced by the RDF sentences shown in Fig. 1. Note that we say an RDF sentence *describes* a term if, in the RDF sentence, the term has an occurrence that is not an *instantiation* (i.e. as the object of an RDF triple whose predicate is `rdf:type`, or as the predicate of an RDF triple). For instance, in Fig. 1, s_3 describes `vin:Region`, and thus, in Fig. 2, $\langle s_3, \texttt{vin:Region}\rangle \in \text{STG}$.

Different from [2] that operates on an RDF Sentence Graph, here we measure the salience of an RDF sentence by computing its centrality on the STG induced by its ontology. To this end, we adapt the well-known PageRank algorithm for this bipartite graph setting. Specifically, let $\text{PR}_r(s)$ be the ranking score of an RDF sentence s at state r, which is iteratively updated as follows:

$$\text{PR}_{r+1}(s) \triangleq \frac{1-\lambda}{n} + \lambda \cdot \sum_{\{\langle s',t\rangle \in \text{STG} \,|\, \exists \langle s,t\rangle \in \text{STG}\}} \frac{\text{PR}_r(s')}{\text{d}(s') \cdot \text{d}(t)}\,, \tag{2}$$

in which $\lambda \in (0,1)$ is a dumping factor, n is the total number of RDF sentences in o, and d returns the degree of a node in STG. To exploit PageRank, inspired by [1], we actually treat every path of length 2 from one RDF sentence to another in STG as a "link" between them; we assume a surfer, who performs random walks between RDF sentences, at each step either jumps to any RDF sentence or follows a "link" to some RDF sentence; the surfer selects targets always using a uniform probability distribution.

Finally, in PageRank it has been proved that $\text{PR}_r(s)$ converges toward a constant $\text{PR}^*(s)$ that does not depend on any initial values of PR. With $\text{PR}^*(s)$ that represents the centrality of s in the ontology and thus characterizes its salience, we define the *salience* of a summary S as:

$$\text{Salience}(o, S) \triangleq \sum_{s \in S} \text{PR}^*(s)\,. \tag{3}$$

3.2 Query Relevance

In the context of ontology search fed by a keyword query submitted by a user, the selection of RDF sentences should reflect not only the most salient part of an ontology but also how the ontology matches the user's needs carried by the query. We achieve this by constructing a textual representation for every RDF sentence and measuring how similar to the query it exhibits.

Specifically, let $\text{KWSet}(s)$ be the set of normalized (e.g. lowercased) keywords found in RDF sentence s. These keywords come from the local name of every term described by s and the lexical form of every literal that occurs in s. A query Q is also defined as a set of keywords. Thereby, the textual similarity between s and Q could be defined as the "precision" of $\text{KWSet}(s)$ in matching Q:

$$\text{TextSim}(s, Q) \triangleq \frac{|\,\text{KWSet}(s) \cap Q|}{|\,\text{KWSet}(s)|}\,. \tag{4}$$

Finally, the *query relevance* of a summary S is given by:

$$\text{Relevance}(S, Q) \triangleq \sum_{s \in S} \text{TextSim}(s, Q)\,. \tag{5}$$

3.3 Cohesion

Cohesion indicates the user-perceived unity of a summary. In ontology summarization, it can be achieved by having constituent RDF sentences referring to the same thing, i.e. they are connected by the terms they describe in common. A direct benefit from increasing such interconnections is that, if we visualize a summary as a merge of its constituent RDF sentences, this resulting RDF graph will be less disconnected so that it could better characterize the semantic relationships between terms. Such relationships function as additional information provided to users and may lead to more accurate relevance judgments in search.

Specifically, given $\text{Describes}(s)$ being the set of terms described by RDF sentence s, we define the *connection* between two RDF sentences:

$$\text{Connection}(s_i, s_j) \triangleq |\,\text{Describes}(s_i) \cap \text{Describes}(s_j)|\,. \tag{6}$$

Finally, the *cohesion* of a summary S is given by:

$$\text{Cohesion}(S) \triangleq \sum_{\substack{s_i, s_j \in S \\ s_i \neq s_j}} \text{Connection}(s_i, s_j)\,. \tag{7}$$

4. CONCLUSION AND FUTURE WORK

We have described a novel approach for summarizing ontologies, and have implemented it in a newly developed ontology search engine. In future work, we will have a more in-depth look at the hypothetical surfer's behavior of random walks on STG. We will also conduct an empirical study to compare this work with existing approaches.

5. ACKNOWLEDGMENTS

This work was supported by the NSFC under Grant 60973024. We would like to thank Dr. Xiang Zhang and anonymous reviewers for their invaluable comments.

6. REFERENCES

[1] L. Lempel and S. Moran. SALSA: The stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, April 2001.

[2] X. Zhang, G. Cheng, and Y. Qu. Ontology summarization based on rdf sentence graph. In *Proc. WWW*, pages 707–716, May 2007.