Toward Efficient Semantic Annotation: A Semantic Cloud Generation Scheme from Linked Data

Han-Gyu Ko[1] and In-Young Ko[1,2]

[1] Department of Computer Science
[2] Division of Web Science Technology,
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu, Daejeon, 305-701, Republic of Korea
{kohangyu,iko}@kaist.ac.kr

Abstract. As a bridge for evolution of the current Web toward Semantic Web, semantic annotation plays an important role to turn regular Web contents into meaningful ones. However, existing semantic annotation methods mostly use semantic terms in ontology created by domain experts. Therefore, they cannot cover the various subjects of contents, some of which frequently change. To deal with this problem by alternating ontology to Linked Data, we propose a semantic cloud generation scheme that finds and merges relevant terms from Linked Data for a given request. To reduce the complexity of handling a large amount of RDF data, we locate essential points at which to start searching for relevant concepts in Linked Data; we then iteratively analyze potential merges of different semantic data. In this paper, we describe the challenges of forming semantic clouds out of Linked Data and the approach of effectively generating semantic clouds by using the similarity link analysis method.

Keywords: Similarity link analysis, Semantic cloud generation, Linked Data, Semantic annotation, Semantic Web.

1 Introduction

Annotation on Web contents is a way of adding user-generated metadata, also known as *tags* to the contents by individual users. With the emergence of Web-based Social Network Services (SNS) and multi-media content sharing services such as YouTube, tags usually play an important role of efficiently finding appropriate contents from a large amount of content data [1]. However, merely allowing users to insert keyword-based tags may result in semantic ambiguity problems [2]. Users may experience difficulty of distinguishing between different semantics that can be represented by using the same keyword or phrase. For example, the word, 'apple' may be used in some tags to mean a company, Apple Inc. rather than a fruit. The semantic Web research community has developed semantic annotation approaches [1, 2] to overcome this problem.

Previous efforts for semantic annotation on Web contents have one essential problem. Most of them use semantic terms that are defined in an ontology created by

Q.Z. Sheng and J. Kjeldskov (Eds.): ICWE 2013 Workshops, LNCS 8295, pp. 315–319, 2013.

domain experts. Ontology, however, includes only a set of domain-specific terms and usually does not include new terms that reflect emerging trends and phenomena in the society. In addition, domain-specific ontologies do not cover various subjects of contents, some of which frequently change [3, 4]. This may restrict users' choice of adding semantic annotations to the social media and multi-media Web contents that include trendy issues and new topics. In order to tackle this limitation, we need a new knowledge source for semantic annotation.

There have been recent research efforts to use Linked Data to overcome the limitations of utilizing domain-specific ontologies. Linked Data is a large scale and evolvable Semantic Web data which contains more than 30 billion RDF triples in various datasets. It is driven by the idea of open access to structured data. Faviki [5] is a famous semantic annotation application. It uses Wikipedia concepts as common tags. When a user inputs a keyword, it recommends candidate tags that come from Wikipedia. Mirizzi et al. proposed an approach of generating semantic tag clouds from DBpedia [6]. Their approach finds relevant concepts by using Google Similarity Distance. However, both approaches depended on a single dataset such as DBpedia.

In this paper, we propose a semantic cloud generation scheme that searches through multiple Linked Data datasets based on a keyword given by a user and finds the essential concepts that are semantically relevant to the keyword. By using our approach, semantic annotation on Web contents can be done as follows. First, a user enters a keyword while accessing a Web content (e.g., watching a video content). Then, the user can see multiple groups of terms around the content. Since each group is made with coherent terms, the user can easily understand the semantics of the term group. We define a group of relevant terms as a *semantic cloud*.

In order to generate semantic clouds, the proposed scheme firstly locates essential points (we call them as *spotting points*) to start searching for relevant terms in Linked Data and then iteratively analyze potential merges of relevant semantic terms. For example, semantic clouds for the input keyword, 'apple' can be generated with a number of spotting points such as 'Apple Inc.' the IT company, 'Apple Records' the music company, and 'Apple' the fruit. The proposed scheme organizes the temporal graphs of terms that describe the same and relevant concepts by parsing semantic relations such as `owl:samsAs`, `skos:broader`, and `skos:narrower` which are defined as *similarity links* in this paper. Then, the proposed scheme chooses the terms that have higher centrality values than others as spotting points.

2 Challenges in Semantic Cloud Generation from Linked Data

In this section, we describe the issue of generating semantic clouds from Linked Data in an efficient way. In order to achieve this goal, we have identified challenges and requirements for semantic cloud generation as follows:

1) **Restrict the number of clouds:** To ensure the usability of semantic clouds in choosing the right semantics to annotate on a Web content, the number of clouds generated should be restricted. According to educational measurement research result, the number of options should be four at most to make users efficiently choose among the options [7].

2) **Maximize the semantic cohesiveness of terms within a cloud:** To allow users to intuitively recognize the meaning of a semantic cloud, it is essential to include only the semantically relevant and coherent terms in the cloud.

3) **Minimize the semantic ambiguity between clouds:** To make users easily recognize the different meanings embedded in different semantic clouds, we need to minimize the semantic ambiguity between the clouds.

In this paper, we focus on explaining how to efficiently retrieve and merge relevant concepts from Linked Data to meet the above requirements.

3 The Proposed Scheme for Semantic Cloud Generation

We developed a similarity link analysis method to meet the requirements in generating semantic clouds efficiently for semantic annotation. The process of generating semantic clouds from Linked Data is incremental and iterative. There are two steps in the semantic cloud generation process as shown in Figure 1. Firstly, spotting points for finding relevant concepts are located in Linked Data. This is done by finding the representative RDF nodes that cover the concepts that are related to an input keyword. The representative RDF nodes that have high centrality characteristic will be selected as spotting points. After locating the spotting points, the proposed scheme traverses the neighboring RDF nodes, which are connected via semantic similarity links. The visited RDF nodes are then merged with the spotting points to form semantic clouds.

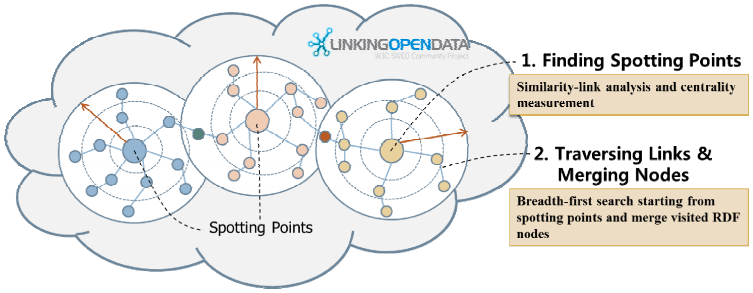

Fig. 1. Overall process of semantic cloud generation

3.1 Finding Spotting Points

If we can find a RDF node that is located in the center around many other nodes that are semantically relevant, we can efficiently reach to other relevant RDF nodes in a shortest distance. The process of finding the spotting points starts with querying Linked Data to obtain RDF triples. In order to find and group concepts that describe the same concept, we use the identity links such as `owl:sameAs` and `skos:exactMatch` relationships. We then find and merge semantically relevant concepts by using the similarity links such as `skos:broader` and `skos:narrower`. We use the similarity links along with the identity links in a complementary manner to group the relevant concepts retrieved

from different datasets in Linked Data. The identity links are used to identify the same concepts across different datasets that use different ontologies. Then, the similarity links are used to find relevant concepts in a subsumption hierarchy in each dataset.

To select the spotting points, we measure the centrality of the concepts retrieved. The concepts that have the highest centrality value are selected as representative concepts. We use the betweenness centrality measure because it considers both the connectivity of a node and the efficiency (shortest paths) of reaching to other concepts. The betweenness centrality is calculated as follows:

$$C_B(v) = \sigma_{st}(v)/\sigma_{st} \tag{1}$$

where σ_{st} is the total number of shortest paths from node s to node t, $\sigma_{st}(v)$ is the number of shortest paths that pass through v. The betweenness centrality can be normalized by dividing its value by the number of node pairs that do not include v - $(n-1)(n-2)$ for directed graphs, and $(n-1)(n-2)/2$ for undirected graphs.

3.2 Traversing Links and Merging Relevant Concepts

After finding the spotting points for the input keyword, a semantic cloud is generated by traversing all the links connected to a spotting point. We use the breadth-first search algorithm for the traversal. The topological distance between a spotting point and other nodes can be used to measure the semantic relevance. We can limit the distance of traversing links from a spotting point to improve the performance of generating a sematic cloud. In our experiment, we limited the maximum distance as 2. Based on our investigation, more than 80% of relevant concepts can be found by traversing the concepts in the distance, 2 from a spotting point [7].

4 Evaluation

In order to prove the effectiveness of the proposed semantic cloud generation scheme, we conducted both an empirical study and a user study. We collected 30 most popular keyword tags such as 'Art', 'Travel', and 'Wedding' used in Flickr[1], and measured how much (in terms of the number of RDF triples) of the relevant concepts for each tag can be retrieved and grouped together as a semantic cloud. We found that in average, 19% of the relevant concepts can be merged together sucessfully. In the case of the keyword, 'Nature', 75% of the cencepts are merged. The concept reduction ratio is proportional to the ratio of having similarity links in the retrieved concepts.

In addition, we performed a user study to validate the practical efficiency of the proposed scheme. We collected photo images that correspond to the Flickr's popular tags that are used for the evaluation. We showed our participants each of images with a set of semantic clouds that were generated by using our approach. We then let the participants chose one of the semantic clouds that they thought the most relevant one for the image. If they could not find an appropriate cloud, they could ask the system

[1] www.flickr.com/photos/tags/

for a new set of semantic clouds. In the preliminary study, most of participants were able to find the appropriate semantic clouds for the images in few clicks.

5 Conclusion and Future Work

In this paper, we proposed a semantic cloud generation scheme that locates spotting points to start searching for relevant concepts in Linked Data and then incrementally analyze potential merges of different semantic data. The proposed scheme analyzes the similarity links between concepts and measures the betweenness centrality of each concept in order to find spotting points. Because similarity links are widely used in various datasets in Linked Data, it is possible to reduce the number of candidates that users need to consider for each annotation. In addition, because the proposed scheme incrementally traverses Linked Data to find relevant concepts starting from a spotting point, it is possible to improve the performance of the semantic cloud generation.

In our future research, we will carry out more intensive user studies to measure and prove the usability of the semantic cloud based annotation considering semantically ambiguous situations. We will use some semantically ambiguous keywords and check if the users can efficiently annotate Web contents with the semantic clouds generated based on the keywords.

Acknowledgments. This research was supported by the WCU (World Class University) program under the National Research Foundation of Korea and funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007).

References

1. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. elsevier Journal of Web Semantics (2005)
2. Reeve, L., Han, H.: Survey of Semantic Annotation Platforms. In: ACM Symposium on Applied Computing (2005)
3. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: KIM–Semantic Annotation Platform. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 834–849. Springer, Heidelberg (2003)
4. Handschuh, S., Staab, S., Ciravegna, F.: S-CREAM – Semi-automatic CREAtion of Metadata. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 358–372. Springer, Heidelberg (2002)
5. Faviki, `http://www.faviki.com` (accessed May 8, 2013)
6. Mirizzi, R., Ragone, A., Di Noia, T., Di Sciascio, E.: Semantic tag cloud generation via DBpedia. In: Buccafurri, F., Semeraro, G. (eds.) EC-Web 2010. LNBIP, vol. 61, pp. 36–48. Springer, Heidelberg (2010)
7. Lord, F.M.: Optimal Number of Choices per Item – A Comparison of Four Approaches. Journal of Educational Measurement, 14(1), 33–38 (1997)