Generation of Semantic Clouds Based on Linked Data for Efficient Multimedia Semantic Annotation

Han-Gyu Ko and In-Young Ko

Department of Computer Science, Korea Advanced Institute of Science and Technology,
335 Gwahangno, Yuseong-gu, Daejeon, 305-701, Republic of Korea
{kohangyu,iko}@kaist.ac.kr

Abstract. The major drawback of existing semantic annotation methods is that they are not intuitive enough for users to easily resolve semantic ambiguities while associating semantic meaning to a chosen keyword. We have developed a semantic-cloud-based annotation scheme in which users can use semantic clouds as the primary interface for semantic annotation, and choose the most appropriate concept among the candidate semantic clouds. The most critical element of this semantic-cloud-based annotation scheme is the method of generating efficient semantic clouds that make users intuitively recognize candidate concepts to be annotated without having any semantic ambiguity. We propose a semantic cloud generation approach that locates essential points to start searching for relevant concepts in Linked Data and then iteratively analyze potential merges of different semantic data. We focus on reducing the complexity of handling a large amount of Linked Data by providing context sensitive traversal of such data. We demonstrate the quality of semantic clouds generated by the proposed approach with a case study.

Keywords: Semantic Web, Semantic Annotation, Linked Data, Semantic Cloud Generation.

1 Introduction

As users become the center of content creation and dissemination in the current Web environment, they are playing a more significant role in metadata generation. For instance, users may create tags that can be used to enhance content search results. However, the attempts to improve content searching by merely considering tags as plain text, have led to the problem of semantic ambiguity [5, 6]. Nevertheless, the Semantic Web research community has been utilizing the semantic annotation of contents as a way to overcome these limitations.

However, these previous efforts on semantic annotation of Web contents fail to fulfill the requirements of scalability and usability [5, 6]. Most existing semantic annotation tools use terms from ontologies created by domain experts. These ontologies do not, however, provide sufficient options to cover various kinds of semantics. That is, only domain specific terms are available. In addition, these ontologies do not necessarily reflect newly created knowledge in an up-to-date manner.

A. Harth and N. Koch (Eds.): ICWE 2011 Workshops, LNCS 7059, pp. 127–134, 2011.

In this paper, we propose a semantic-cloud-based annotation scheme that makes it easier to add semantic annotations to multimedia contents in resource-constrained environments, such as IPTV (Internet Protocol Television), since an interesting application area of semantic annotation is the increasing market of businesses that use multimedia contents on the Web [7]. The proposed approach uses semantic clouds as the primary interface for semantic annotation. In order to generate the semantic clouds, we first locate essential points to start searching for relevant concepts in Linked Data [1] and then iteratively analyze potential merges of different semantic data. Users can easily resolve semantic ambiguity and choose the most appropriate semantic cloud among a set of candidates.

2 Multimedia Semantic Annotation Scheme

In this section, we describe the proposed semantic annotation scheme. As the following figure shows, the semantic cloud generated from the Linked Data is used as the primary interface for semantic annotation.

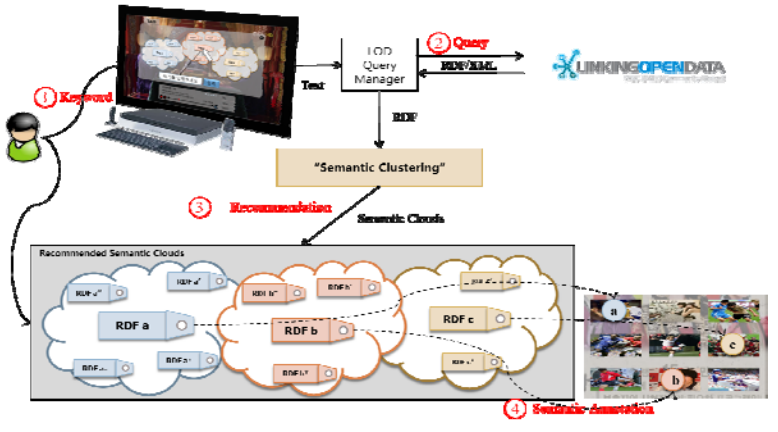


Fig. 1. Overview of the semantic annotation scheme applied to multimedia contents

When a user inputs a keyword while watching a multimedia content, the Linked Data query manager makes a query and obtains the relevant RDF nodes from the Linked Data. The proposed scheme generates recommended semantic clouds and the user then annotates the contents by choosing the most appropriate concept from them.

In this annotation scenario, there are three technical issues to resolve: accessing and processing large-scale Semantic Web data, generating relevant semantic clouds and providing an efficient user interface that allows intuitive interactions. In this paper, we focus on the issue of semantic cloud generation from the large-scale Semantic Web data. In order to achieve this goal, we identify the requirements for well-organized semantic clouds as follows:

1) **Small number of clouds**: The number of options should be four at most [8]
2) **Balance of contents in the cloud**: Semantically relevant terms should also be included in the same cloud
3) **No ambiguity among clouds**: Semantic ambiguity among generated semantic clouds should be minimized so as to facilitate awareness of semantic differences

The proposed semantic cloud generation approach that satisfies these requirements will be introduced in the following section.

3 The Proposed Semantic Cloud Generation Approach

According to the statistics [13], Linked Data contains more than 28 billion RDF triples from 203 different datasets that are domain independent. Hence, Linked Data is a large-scale and heterogeneous Semantic Web data store. In order to generate semantic clouds from the Linked Data, we need to make our semantic cloud generation process incremental and iterative.

There are three steps in the cloud generation process. First, *spotting points* for the clustering should be located. This entails finding representative RDF nodes that cover the concepts related to an input keyword.

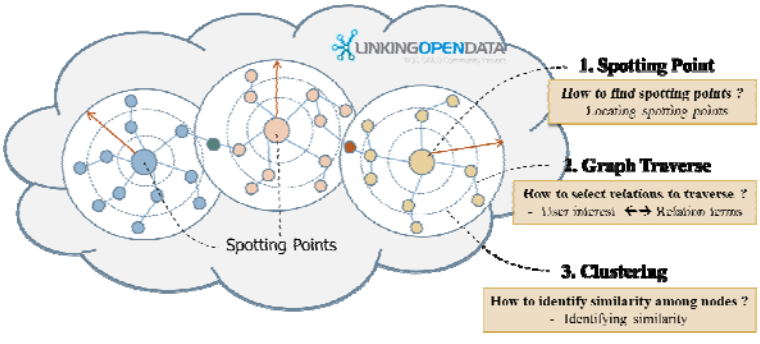


Fig. 2. Overall process of the semantic cloud generation

After locating the spotting points, the proposed approach selectively visits the neighboring nodes connected via relation terms interlinked with user context. This reduces the complexity of handling a large amount of Linked Data and also ensures the quality of semantic coherence of the generated semantic clouds by filtering out the less relevant relationships and the corresponding nodes.

In order to decide whether to include a visited RDF node in a cloud, it is necessary to measure the semantic similarity between the spotting point and the visited RDF node. Basically, the number of overlapping concepts could be the standard to measure the semantic similarity. The distance between a spotting point and the visited RDF node, which is measured by counting the number of hops from the spotting point to the RDF node, can be also considered to measure the semantic similarity.

3.1 Locating Spotting Points

The first step of finding and locating spotting points in the Linked Data is the most important process to generate high quality semantic clouds that satisfy the requirements discussed in the previous section. This is because the spotting points decide the representative semantics of the user keyword and the generated semantic clouds are dependent on the spotting points.

Locating spotting points starts with querying Linked Data to obtain the relevant RDF nodes. There exists two ways to make queries to Linked Data: via SPARQL or via Semantic Web search engines such as Swoogle [9], Falcons [10], or Sindice [11]. We chose the second method because we can obtain the relevant RDF nodes by simply adopting and using one of their Web services.

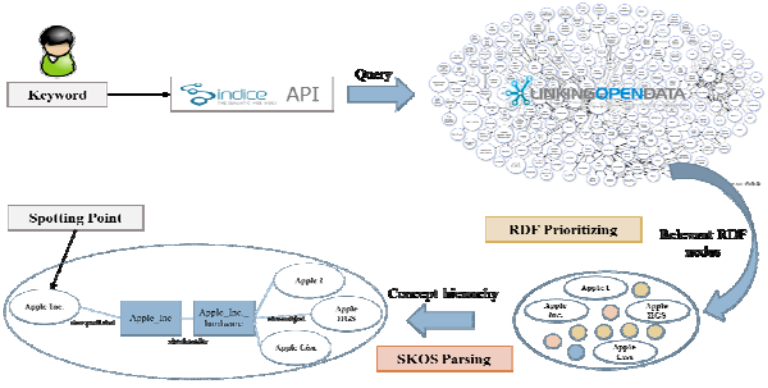


Fig. 3. Finding spotting points in the Linked Data

When we make a query about the keyword ‘apple’ via the Sindice API, it returns more than 400,000 RDF nodes. Rather than taking all the nodes, choosing some representative nodes could reduce the complexity of semantic cloud generation. Because Sindice ranks the resulting nodes by applying the principles of the PageRank algorithm as well as term frequency [12], taking the top n number of query results replaces RDF prioritizing.

We finally choose the most general concepts of RDF nodes as the spotting points by comparing their relative concept hierarchies because it ensures semantic unambiguity among the spotting points, thus supporting the requirement of no ambiguity among generated semantic clouds. The SKOS (Simple Knowledge Organization System) is a common data model for sharing and linking knowledge organization system. It provides useful relationship terms such as `skos:broader` and `skos:narrower` that can be exploited to find the relative concept hierarchies to group the RDF nodes, then choose the most general RDF node as the spotting point.

Figure 3 shows an example of finding a spotting point by parsing SKOS relationships. Some RDF nodes are extracted with the keyword ‘apple’ such as ‘Apple Inc.’, ‘Apple I’, ‘Apple IIGS’, and ‘Apple Lisa’. By parsing their SKOS relationships, we can recognize the most general concept of the RDF nodes; in this case ‘Apple Inc.’.

3.2 Selecting Relations to Traverse

The second step of the proposed approach is to select relations that link the relevant RDF nodes. We can thereby reduce the complexity for the semantic cloud generation by setting traversal bounds. The ideal method to select semantically relevant relations is to automatically associate the relation terms with user contexts such as interests and preferences.

We assume that user interests are interlinked with relation terms and they are defined by each user before semantic cloud generation. For example, in the case where 'movie' is a user interest, relation terms such as 'actor', 'director', 'rating', 'background music', and 'story' become the relations to traverse.

In addition, W3C recommends that Linked Data publishers use well-defined and popular terms such as FOAF, DC, SIOC, and SKOS in order to ensure interoperability among the Linked Data datasets. The proposed approach firstly traverses the relations and then takes into account the relations selected by users with consideration of their contexts. This facilitates visiting relevant nodes while reducing the complexity for clustering these nodes.

3.3 Identifying Similarity and Clustering

In the third step of the proposed approach, the semantic similarity between the RDF nodes is measured in order to decide whether to include the visited RDF nodes in the same cloud.

Similar to the term frequency in information retrieval, the number of query responses from the Semantic Web search engine is also used to measure the similarity between nodes. In the following equations, l_1 and l_2 are the labels of RDF nodes, $n(l)$ denotes the number of query responses for the RDF node l, and h is the number of hops to traverse.

$$TermFreq(l_1, l_2) = n(l_1, l_2) / n(l_1) + n(l_1, l_2) / n(l_2). \quad (1)$$

$$SemSim(l_1, l_2) = TermFreq(l_1, l_2) / w^h. \quad (2)$$

As the number of hops from a spotting point becomes larger, the value of semantic similarity exponentially decreases. For this reason the second equation that represents the semantic similarity between two RDF nodes take the inverse of the weight value w powered by h.

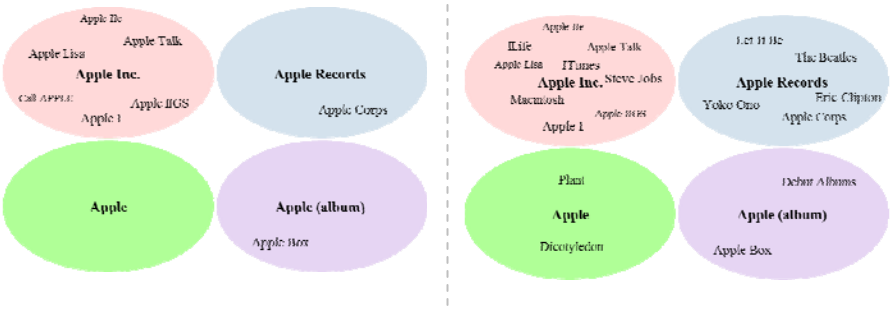
We need to carefully decide the threshold value h for clustering as well as the weight value w such that it includes semantically related concepts toward the keyword. Deciding each value is beyond the research scope of this paper, however.

4 A Case Study

There are three methods of semantic cloud generation. The first approach clusters RDF nodes according to their `rdf:type`. However, this method does not ensure high

quality semantic cloud generation. For instance, 'Apple Inc.' whose `rdf:type` is 'company' is separated from the groups 'Apple I', 'Apple IIGS', etc., whose `rdf:type` is 'Personal Computer', despite that there clearly is semantic relevance.

The next approach is using SKOS relationships. This method is useful to understand the relative concept hierarchy among RDF nodes. However, the obtained results fail to satisfy balance of content in each cloud.



(a) SKOS parsing (b) The proposed approach

Fig. 4. The result of Linked Data clustering toward the keyword 'apple'

As can be seen the above figure, the semantic clouds from the proposed approach provide better results, since the proposed approach also includes relevant RDF nodes which don't contain the keyword 'apple' via relation traversal.

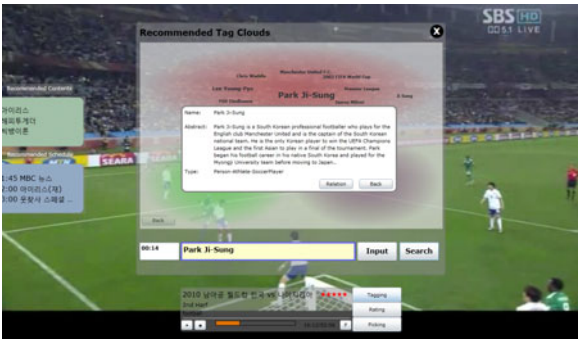


Fig. 5. Implementation of the semantic annotation method for a Web-based IPTV environment

The proposed approach was also applied to a Web-based IPTV environment. The proposed approach allows users to put annotations on multimedia contents by choosing the semantic options from the semantic clouds generated from Linked Data. The annotation results are used to provide semantic search capability, which enriches the search results for multimedia contents.

5 Related Work

In order to overcome the limitations such as semantic ambiguity of using tags as plain text, a semantic annotation scheme has been defined and proposed. Its definition is tagging ontology class instance data and mapping it into ontology classes [5]. The major benefit of semantic annotation is enhanced information retrieval, because it exploits ontologies to infer about data from heterogeneous resources, thereby resolving ambiguities such as 'Niger' the country and 'Niger' the river.

There are three semantic annotation methods, differentiated according to the level of automation: manual, semi-automatic, and automatic annotation. Because human annotators are often fraught with errors and this form of annotation is very costly, manual semantic annotation may cause knowledge acquisition bottleneck [2]. In addition, it is impossible to provide fully automatic creation of semantic annotations. In response, semi-automatic annotation approaches have been explored. The main issues to be resolved are difficulties in choosing appropriate indexing terms for annotating and dealing with unbalanced content arising from the different conventions used in indexing by different users [3]. Also, as the basic prerequisite for representation, most works uses an ontology defining the entity classes as a knowledge base [4].

The proposed approach generates a few semantic clouds as the primary interface for semantic annotation from Linked Data, enabling users to intuitively recognize semantic options. Users can easily resolve semantic ambiguity and choose the most appropriate node among the candidate semantic clouds even in resource constrained environments.

6 Conclusion and Future Work

In this paper, we propose a semantic clustering approach that locates spotting points to start searching relevant concepts in Linked Data and then iteratively analyze potential merges of different semantic data. Using this approach, we attempt to reduce the complexity of handling a large amount of Linked Data by providing context sensitive traversal of Linked Data.

Through a case study, we showed that the proposed semantic cloud generation approach ensures high quality semantic clouds in terms of optimal number of choices, balance of contents, and no ambiguity among generated semantic clouds. Because it allows users put annotations on multimedia contents by simply using keywords and choosing the most appropriate concept among the generated semantic clouds, it can also be applied in resource constrained environments such as the small screen of smart phones and IPTV environments where it is difficult to use text input interfaces of remote controllers.

In future research we will carry out user studies to measure and prove the usability of the proposed semantic annotation approach as well as empirical studies to answer questions such as how many RDF nodes need to be considered at the phase of locating spotting point, how many hops need to be traversed to generate semantic clouds efficiently, and what is the most appropriate threshold value to decide whether a RDF node be included in the same cloud.

Acknowledgments. This research was partially supported by 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). This research was also supported by the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA-2011-11913-05005).

References

1. Christian, B., Tom, H., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
2. Bayerl, P.S., Lungen, H., Gut, U., Paul, K.I.: Methodology for reliable schema development and evaluation of manual annotations. In: Knowledge Markup and Semantic Annotation at the International Conference on Knowledge Capture 2003 (2003)
3. Vehvilaiinen, A., Hyvonen, E., Alm, O.: A Semi-Automatic Semantic Annotation and Authoring Tool for a Library Help Desk Service. In: Proceedings of the 1st Semantic Authoring and Annotation Conference 2006 (2006)
4. Kiryakov, A., Popov, B., Ognyanoff, D., Manov D., Kirilov A., Goranov M.: Semantic Annotation, Indexing, and Retrieval. ELSEVIER Journal of Web Semantics 2004 (2004)
5. Reeve, L., Han, H.: Survey of Semantic Annotation Platforms. In: ACM Symposium on Applied Computing (2005)
6. 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)
7. Ko, I.-Y., Choi, S.-H., Ko, H.-G.: A Blog-Centered IPTV Environment for Enhancing Contents Provision, Consumption, and Evolution. In: Benatallah, B., Casati, F., Kappel, G., Rossi, G. (eds.) ICWE 2010. LNCS, vol. 6189, pp. 522–526. Springer, Heidelberg (2010)
8. Lord, F.M.: Optimal Number of Choices per Item – A Comparison of Four Approaches. Journal of Educational Measurement 14(1), 33–38 (1977)
9. Ding, L., Finin, T., Joshi, A., Pank, R., Cost, S.R., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata eigine for the semantic web. In: Proceedings of the CIMK 2004 (2004)
10. Cheng, G., Ge, W., Qu, Y.: Falcons: Searching and Browsing Entities on the Semantic Web. In: Proceedings of the 17th International World Wide Web Conference, Beijing, China, April 21-25 (2008)
11. Tummarello, G., Delbru, R., Oren, E.: Sindice.com: Weaving the Open Linked Data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 552–565. Springer, Heidelberg (2007)
12. Delbru, R., Rakhmawati, N.A., Tummarello, G.: Sindice at SemSearch 2010. In: Proceedings of the 19th International World Wide Web Conference, Raleigh, North Carolina, USA, April 26-30 (2010)
13. W3C SWEO Community Project Linking Open Data, `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`