Maintenance of Human and Machine Metadata over the Web Content

Karol Rástočný and Mária Bieliková

Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology, Ilkovičova 3, Bratislava, Slovakia
{name.surname}@stuba.sk

Abstract. Semantics over the Web content is crucial for web information systems, e.g. for effective information exploration, navigation or search. However, current coverage of the Web by semantics is insufficient. Web information systems mostly create their own content based metadata (e.g., identified keywords) and user collaboration metadata (e.g., implicit user feedbacks) in a form of information tags – structured information with semantic relations to the tagged content. By information tags web information systems build a lightweight semantics over the Web content, in which they can store knowledge and information about the content and interconnections between information artifacts of the content. Crucial problem of information tags lies in dynamicity of the Web whose content is continually modified. This together with influence of time can lead to invalidation of information tags which are closely related to tagged content. We address this issue via maintenance approach based on automatically and semi-automatically generated rules that respect changes on the Web and time aspect. The maintenance utilizes a rule-based engine which watches changes in the tagged content, identifies dependencies among maintenance rules and builds optimal strategy of rules application. We evaluate proposed maintenance approach in two domains – programing repositories and digital libraries, which use shared information tags repository.

Keywords: metadata, information tag, maintenance, lightweight semantics.

1 Research Context

The Web was originally proposed as a hypertext – a repository of interconnected textual documents by links (references) straight from the document content [1]. This original idea has shifted from linked documents to linked data [2] nowadays. But these data are still mostly made accessible just in a human readable format (e.g., wrapped in the textual web page) which is not effective for a machine processing.

The problem is addressed by the Sematic Web initiative which is strictly oriented to data that are obviously stored in triple stores or ontologies [3]. Web information systems can use data from the Semantic Web repositories for an inference of new knowledge and to support users' information seeking and processing activities [4, 5]. Emergence of the Semantic Web for a support of users' activities often needs an interconnection between the Semantic Web ontologies and the “wild” Web content. The interconnection is provided by semantic annotations which annotate parts of a natural text with their formal representations in ontologies [6] (e.g., a word “Berlin” in a natural text can be annotated by the URI of the “Berlin” entity in an ontology).

Although some semi-automatic ontology learning approaches [7] and approaches for creating semantic annotations [8] exist already, domain experts have to make a non-trivial effort to propose rules for ontology extraction from the “wild” Web, to filter out misidentified entities and to maintain ontologies. As the result of this complication only a small part of the Web is covered by ontologies [9]. An improvement can be achieved by lightweight ontologies which usually contain only basic elements as terms or concepts and “is-a” and “part-of” relations [10]. Lightweight ontologies can be extracted from the webpage content and from users’ activities as annotating [11] and sharing which have become popular with an emergence of the Web 2.0 [12].

Similarly to user annotating activities, web information systems assign metadata to the Web content. These metadata describe particular aspect of an information artifact – a part of a webpage. We look on them as on structured tags which are generated by systems. These tags are based on the Web content (e.g., extracted concepts) or on users’ activities (e.g., relevant terms identified in often read document parts).

Well-structured tags are created by users, too. These human tags can be processed by systems and they contain valuable users’ information and knowledge e.g., explicit users’ ratings or keywords. We group metadata of described type and well-structured human tags under term *information tags*. Formally, an information tag is a triple *(type, anchoring, body)*, where *type* defines a type and a meaning of the information tag, *anchoring* identifies a tagged information artifact and *body* represents structured information those structure corresponds to the type of the information tag.

An advantage of information tags over freeform human annotations is that information tags are already in machine readable format. In addition information tags are in a semantic relation to tagged aspect of an information artifact (they are assigned to tagged data with specific purposes), so they provide a lightweight semantics over the Web content. But existing systems obviously store their information tags in private repositories in a form which is understandable only for them. More crucial problem is a dynamicity of the Web whose content is continually modified. This with influence of a time affects validity and topicality of information tags which are closely related to tagged content. To support building of a lightweight semantic based on information tags, we propose information tags maintenance approach based on automatically and semi-automatically generated rules that respect changes on the Web and a time aspect.

2 Research Objectives

The main goal of our work is designing information tags maintenance approach which keeps information tags in a consistent form. To fulfill this goal we have to deal with:

- *Diversity of information tags formats and semantics* – each web information system generates information tags in different format and with different sematic relation to the source content.

- *Information tags accessibility* – web information systems have to be able find information tags assigned to the Web content.
- *Dynamicity of the Web* – the Web documents arise, are deleted and modified without a notice. In addition the Web users use the Web content differently over the time. These modifications, diversity in usage of the Web content and also time aspect invalidate information tags that have to be updated or deleted.

The first issue is not a direct part of the maintenance but, it falls to the scope of storages and data integration. Despite it, we have to deal with it for evaluation reasons. So we divide addressing of these issues to three parts whereby the last two parts represent main contribution of our research:

- *Information tags repository* – stores information tags in a flexible model which will be acceptable by wide range of web information systems. Current systems mostly preferred RDF-based models, e.g. Open Annotation (OA) model[1] which is currently in beta version but it is already used by a number of systems and projects [13–15]. Metadata in RDF-based format should be stored in triple stores that are good for inference but, they are not effective for manipulation with whole objects. But information tags have a meaning only as whole objects with information tags' anchoring and content together. We suppose that *information tags can be stored in a repository with RDF-based model, which stores information tags as a one entry and not fragmented to a set of entries (triples) and the repository still provides basic functionality of triple stores* (e.g., SPARQL querying).
- *Maintenance logic* – provides an automatic maintenance over information tags via maintenance rules that respect changes of the Web content and time aspect. We assume that if lightweight ontologies can be learned semi-automatically and automatically [10], their *maintenance rules can be semi-automatically and automatically learned by watching of a life cycle of information tags*. Because of some information tags are derived from other information tags, learned rules will not be independent and application of a one rule can lead to a complex cascade effect. These dependencies among maintenance rules can be identified, so we assume that *a rule-based engine can be used for the maintenance of information tags*.
- *Access provider* – provides an access to the information tags repository and notifies the maintenance logic about updates in the information tags repository and detected new versions of tagged documents. A detection of new versions of tagged document can be based on Memento framework [14].

3 Conclusions and Future Work Plan

We have proposed a repository of information tags, which stores information tags in a document database. Document databases store complete entry as a one document (object) and they often support indexes. Their properties predict them for fast access to whole information tags but, document databases do not provide functionalities of

[1] http://www.openannotation.org/spec/beta

triple stores. For this reason we proposed to employ MapReduce algorithm for effective evaluation of SPARQL queries over entries of document databases.

For evaluation purposes we have chosen web-scale MongoDB[2] document database and we have implemented proposed SPARQL evaluation algorithm for MongoDB. After that we performed several performance tests on single node deploy. We also repeated these tests with a repository based on classic triple store – Bigdata[3] which was chosen for its web-scale possibilities. We noticed that [16]:

- The repository based on MongoDB is at least *hundred times faster* in testing cases that manipulates with whole information tags than repository based on Bigdata.
- SPARQL query evaluation took *approximately same time* in both realizations.

Performed information tags repository evaluations are promising but, they are not representative in the web-scale. For final repository evaluation, larger performance tests with the repository distributed over several nodes have to be performed.

We focus our next work for proposition of the information tags maintenance itself and evaluation of hypothesizes related to the proposed maintenance approach:

- Rule-based engines can be used for the maintenance of information tags.
- Maintenance rules can be semi-automatically and automatically learned by watching of a life cycle of information tags.

We plan evaluate hypothesizes within domains related to two research projects currently realized at the Institute of Informatics and Software Engineering, Slovak University of Technology in Bratislava, that employ a lightweight ontology based on information tags for as a basis for semantics representation:

- *Personalized Conveying of Information and Knowledge* – this project is focused on support of enterprise applications development by viewing a software system as a web of information artifacts. In the project several agents collect and process documentations, source codes, developer blogs, developer activities, etc.
- *Traveling in Digital Space* –main goals of this project are collaborative learning and support of novice researchers to orientate in new research domains. The core part of an employed lightweight ontology contains domain concept maps built from learning materials and from captured user activities with a studied content.

We currently work on a proposition of the rule-based engine for maintenance information tags. The engine updates information tags' anchoring to a tagged content, invalidate (delete) information tags when they become outdated and update information tags' content if it will be possible.

The next step is focused on a proposal of (semi-)automatic maintenance rules learning based on monitoring the life cycle of information tags. We plan to evaluate our approach by a comparison of information tags that are maintained by manually, semi-automatically and automatically created rules. We will also watch update activities provided by web information systems for an evaluation of accuracy and coverage.

[2] http://www.mongodb.org/

[3] http://www.systap.com/bigdata.htm

Acknowledgements. This work was supported by the grants VG1/0675/11, APVV-0233-10 and it is a partial result of the Research and Development Operational Program for the projects SMART, ITMS 26240120005 and SMART II, ITMS 26240120029, co-funded by ERDF.

References

1. Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H.F., Secret, A.: The World-Wide Web. Communications of the ACM. 37, 76-82 (1994)
2. Handschuh, S., Heath, T., Thai, V.: Visual interfaces to the social and the semantic web (VISSW 2009). In: 13th Int. Conf. on Intelligent UIs, pp. 499-500. ACM Press, NY (2009)
3. Shadbolt, N., Berners-Lee, T., Hall, W.: The Semantic Web Revisited. IEEE Intelligent Systems. 21, 96-101 (2006)
4. Ramachandran, V.A., Krishnamurthi, I.: NLION: Natural Language Interface for Querying ONtologies. In: 2nd Bangalore Annual Compute Conf., p. 4. ACM Press, NY (2009)
5. Elbassuoni, S., Blanco, R.: Keyword Search over RDF Graphs. In: 20th ACM Int. Conf. on Information and Knowledge Management, pp. 237-242. ACM Press, NY (2011)
6. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargasvera, M., Motta, E., Ciravegna, F.: Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art. Web Sem.: Science, Services and Agents on the WWW. 4, 14-28 (2006)
7. Hazman, M., R. El-Beltagy, S., Rafea, A.: A Survey of Ontology Learning Approaches. International Journal of Computer Applications. 22, 36-43 (2011)
8. Reeve, L., Han, H.: Survey of Semantic Annotation Platforms. In: ACM Symposium on Applied Computing, pp. 1634-1638. ACM, NY (2005)
9. Sabou, M., Gracia, J., Angeletou, S., Aquin, M., Motta, E.: Evaluating the Semantic Web: A Task-Based Approach. In: Aberer, K. et al.(eds.) The Semantic Web. LNCS, vol. 4825, pp. 423-437. Springer-Verlag, Berlin, Heidelberg (2007)
10. Giunchiglia, F., Zaihrayeu, I.: Lightweight Ontologies. Tech. report, Univ. of Trento, p. 10 (2007).
11. Šimko, M.: Automated Acquisition of Domain Model for Adaptive Collaborative Web-Based Learning. Inf. Sciences and Tech., Bulletin of the ACM Slovakia 2(4), 9 p. (2012)
12. Bieliková, M., Barla, M., Šimko, M.: Lightweight Semantics for the “Wild Web”. In: IADIS Int. Conf. WWW/Internet‘11, pp. xxv-xxxii (keynote). IADIS Press (2011)
13. Gerber, A., Hyland, A., Hunter, J.: A Collaborative Scholarly Annotation System for Dynamic Web Documents – A Literary Case Study. In: Chowdhury, G. et al. (eds.) The Role of Digital Libraries in a Time of Global Change. LNCS, vol. 6102, pp. 29-39. Springer-Verlag, Berlin (2010)
14. Sanderson, R., Van de Sompel, H.: Making Web Annotations Persistent over Time. In: 10th Annual Joint Conf. on Digital Libraries. pp. 1-10, ACM Press, NY (2010).
15. Yu, C.-H., Groza, T., Hunter, J.: High Speed Capture, Retrieval and Rendering of Segment-Based Annotations on 3D Museum Objects. In: Xing, C. et al. (eds.) Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation. LNCS, vol. 7008, pp. 5-15. Springer, Berlin (2011)
16. Bieliková, M., Rástočný, K.: Lightweight Semantics over Web Information Systems Content Employing Knowledge Tags. In: S. Castano et al.(Eds.): ER Workshops 2012, LNCS, Vol. 7518, pp. 327–336, Springer-Verlag, Berlin Heidelberg (2012)