Towards Automatic Content Quality Checks in Semantic Wikis
– Position Paper –

Denny Vrandečić
Insitut AIFB, Universität Karlsruhe (TH), Germany
denny@aifb.uni-karlsruhe.de

Abstract

Semantic wikis have shown to be feasible systems to enable communities to collaboratively create semantically rich content. They enable users to make the knowledge within the wiki explicit and thus accessible to the application in order to support the users in numerous ways, like improved browsing, or query answering. This also allows the wiki to automatically check the content. In this paper we present a number of approaches in order to provide facilities to ensure the content quality of a wiki, including the application of constraint semantics and autoepistemic operators in ways that are easy accessible for the end user. We discuss the social aspects of such an approach, and argue for the necessity of enabling end users to define the quality checks within the wiki itself.

Introduction

Wikis have proved to be systems that enable communities to collaboratively create knowledge. Wikipedia, the best known wiki, has shown that wikis can grow to a truly global scale. Wikipedia does not work solely because of its underlying software MediaWiki, but it is a rather complex socio-technical entity that works due to often implicit community processes and rules (Ayers, Matthews, & Yates 2008).

Semantic wikis have shown to be feasible systems to enable communities to collaboratively create semantically rich content. Additionally to classic wikis they also allow the community to add more structured information to the textual and multimedia content. Such structured content is successfully used in order to reduce redundancy and thus increase consistency within the wiki. In this paper we discuss how the structured data can additionally be used in order to further ensure the quality of the wiki content. Whereas the formal basis of these content checks is well understood, it is unclear how the social processes will play out. In this paper we sketch the planned implementation of these content checking features for the Semantic MediaWiki engine in order to gain early feedback from the scientific community.

The next Section introduces Semantic MediaWIki. Then the idea of automatic content checks are explained, followed by a discussion of social and usability issues. We close with a brief view on related work and an outlook.

Semantic MediaWiki

Semantic MediaWiki[1] is an extension to the MediaWiki wiki engine. It enables users to create triples by regarding a wiki page as the subject and a link target with a link type as the object respectively the predicate of a triple. E.g., users can create a link like `[[located in::California]]` on the page for Stanford University, thus stating the triple *"Stanford University" "located in" "California"*. Thus the wiki pages not only describe HTML pages, but also an RDF graph over the whole wiki.

Using that structure, the knowledge can be reused within the wiki in several new ways, e.g. to answer queries. A query language allows to create descriptions for query results. These descriptions are conjuncts of category and property statements, e.g. the description `[[Category:University]] [[located in::Californa]]` is a description consisting of two conjuncts, the first stating that the page has to be in the category *University*, the second that it has to have a link with the type *located in* and the target *California*. Within this paper we will build on this structural information in order to provide automatic content checks. A more detailed description of Semantic MediaWiki can be found in (Krötzsch *et al.* 2007).

Automatic content checks

For now, we plan to introduce the following features for automatic content checking within Semantic MediaWiki:

- **Concept cardinality**, i.e. how many results shall a query within the wiki have. Besides exact numbers also minimal and maximal cardinalities will be allowed. Note that this allows to state disjointness (by stating that the intersection has a cardinality of 0)
- **Domain and range constraints**, i.e. properties can be defined to only be used on pages in a certain categories or that they are allowed to point only to pages in certain categories.
- **Property cardinalities**, i.e. statements about how often a property can be used on a specific page or point to a specific page.

[1] http://www.semantic-mediawiki.org

The checking framework will allow to add further content checks as extensions. We expect that some wikis will introduce specific checks that only apply to the domain at hand.

It has to be noted that the above checks do not follow the usual OWL semantics. Concept and property cardinality in SMW do not relate to nominals and OWL property cardinalities, but follow rather the semantics of autoepistemic operators (Grimm & Motik 2005). Also domain and ranges are not based on the equivocal OWL operators but rather on a constraint semantics (Motik, Horrocks, & Sattler 2007). Within the wiki, we assume a closed world and unique names if not otherwise stated.

The deviation from OWL semantics has to be carefully considered when exporting data from the wiki. It is obvious that domain and range statements in the wiki should not be exported as OWL domain constructs, since that would lead to different semantics of the export and within the system.

A promising content check on Wikipedia can be performed by comparing graphs from different languages. Due to interlanguage links most pages from different language versions of Wikipedia link to each others' pages in other languages. Since this is also true for categories, we assume that a semantic Wikipedia will also follow this for properties, thus mapping all concepts to and from different languages. This allows to create a bot to compare the emerging graph from different languages, and then create reports for the community that warns about differences. Such bots are already used for the interlanguage links themselves. Example: a vandal may switch the capital of a country in one or even two languages, but hardly in all of the more than 200 language versions of Wikipedia, thus creating a system where such a change can be discovered automatically.

Social and usability aspects

Making the above constraints in the wiki must be simple enough to allow contributors to actually apply these features. The given selection is based on the fact that they can be represented within the wiki simply and unambiguously, i.e. contributors will always know where to look for a specific piece of information. This is a necessary requirement in order to keep a wiki maintainable.

Property cardinalities and domain and range constraints can all be defined on their respective property page. Since all properties in SMW have a dedicated page, special properties can be used there, e.g. `[[domain::Person]]` on the property *social security number*. Concept cardinalities can be introduced on concept pages, i.e. pages that contain a query description.[2] Future work could look into learning and suggesting such constraints automatically.

A major decision to make is whether to allow contributors to introduce inconsistencies or not. When a page is modified, the wiki could check if that edit would turn it inconsistent and then cancel the edit. Disregarding if a real time check would be computationally feasible, it seems to conflict with the wiki paradigm. Instead of prohibiting edits that lead to inconsistencies we plan to introduce special pages that report discovered problems and allow to repair them efficiently.

Related work

The MOCA extensions fosters the convergence of the emerging vocabulary within a SMW instance (Kousetti, Millard, & Howard 2008). A convergent vocabulary is not only a requirement for the proper usage of a semantic wiki, but also for the automatic content checks described in this paper.

IkeWiki (Schaffert 2006), an alternative semantic wiki implementation, is embedded stronger in semantic technologies, and stays closer to the semantics defined by OWL. We have argued why we consider a deviation to be necessary.

Conclusions

We presented approaches towards enabling user communities to automatically detect quality problems within the content of their semantic wikis. This enables computers and humans to collaborate in novel ways in order to achieve high quality content in wikis more efficiently. We have argued that it is feasible to enable end users to define and manage such quality checks themselves. Our next step is to gather feedback from the workshop and the community, to implement the suggested approach, and to evaluate them in order to discover their effects on content quality and especially community development. We expect that these features will free up valuable time and effort of contributors who hitherto had to check such information manually, who in turn could engage in other projects towards improving their wiki.

Acknowledgements

Work presented in this paper has been partially funded by the European Union under EU IST FP7 project ACTIVE. We want to thank Markus Krötzsch, Anupriya Ankolekar, Sebastian Rudolph, and the Halo-Team at ontoprise for their valuable input and comments.

References

Ayers, P.; Matthews, C.; and Yates, B. 2008. *How Wikipedia works*. San Francisco, CA: No Starch Press.

Grimm, S., and Motik, B. 2005. Closed world reasoning in the semantic web through epistemic operators. In Grau, B. C.; Horrocks, I.; Parsia, B.; and Patel-Schneider, P., eds., *OWLED2005*.

Kousetti, C.; Millard, D.; and Howard, Y. 2008. A study of ontology convergence in a semantic wiki. In Aguiar, A., and Bernstein, M., eds., *WikiSym 2008*.

Krötzsch, M.; Vrandečić, D.; Völkel, M.; Haller, H.; and Studer, R. 2007. Semantic Wikipedia. *Journal of Web Semantics* 5:251–261.

Motik, B.; Horrocks, I.; and Sattler, U. 2007. Adding Integrity Constraints to OWL. In Golbreich, C.; Kalyanpur, A.; and Parsia, B., eds., *OWLED2007*.

Schaffert, S. 2006. Ikewiki: A semantic wiki for collaborative knowledge management. In Tolksdorf, R.; Simperl, E.; and Schild, K., eds., *WETICE'06 STICA*, 388–396.

[2] `http://semantic-mediawiki.org/wiki/Help:Concepts`