Literature review

Eivind Elseth

February 24, 2012

Abstract

The tentative literature review for my thesis

Contents

1 Introduction

For my thesis I am going to develop a tool that will help gather information about businesses that might be of interest to tourists. The thesis will mainly focus on technologies relating to semantic web, and on lifting semantic data from non semantic sources. I will try to explore how much semantic data one can glean from user input, from users who don't have any in depth knowledge about the related technologies.

To achieve this I will create a tool that allows users to input information in natural language using the lexitags interface, and use see if these tags can be used to extract extra data automatically through existing web services. To answer this question I want to see if I can create a tool that, by utilizing the lexitags system, empower developers and business owners to produce semantic content. I want to perform a usability study to see if the tool can be successfully used by users without knowledge about semantic technologies. I hope to show that it is possible to get naive users to create and deploy semantic content. This could be important as it would show that that proper tools could make creating semantic content available to non experts. This in turn would be important with regards to the proliferation of semantic encoding on the internet.

2 Literature review

2.1 The semantic web

The internet is now an ingrained part of our everyday life, and the amount of content and services that are available through it is growing at an ever increasing rate. For all this information to be of use to humans it is necessary to have some interface through which to access the parts of it that are relevant to us. Shirky (2007) tells of the early attempts to organize the web, using ontologies and hierarchies created by experts. This soon got clunky as the number of documents increased, and this way of organizing information fell out of favor to be replaced by searching for information using keywords. It is this phase of organizing information we are in now. Berners-Lee et al. (2001) suggested that we could do better than this. With searching as it works today users have to manually check the results from the search engine, and compare the results from several documents, following links as is necessary. Instead of forcing users to go through this process, this new idea was to enrich the documents we put on the web with meta data that could be read and reasoned about by computers. By doing this we could move the tedious task of siphoning though websites looking for relevant information from users over to specialized software agents that could collect information on the topic and return the answer to the user.

2.2 Ontology and folksonomy

One of the central concepts in semantic web technology is that of the ontology. In philosophy Ontology is the branch dealing with the study of which things 'exists', and if it is possible to categorize these things. For artificial intelligence Gruber (1993) explained it as "an explicit specification of a conceptualization". That is than one commits to a given conceptualization of the domain in question, and formalize how we describe and reason about these conceptualizations. Pretorius (2004) also gives a good overview of the history of the term, and show several of the interpretations and formalisms. One can also go to Noy and Hafner (1997) to find comparisons of several of the early ontologies, including WordNet which will be important in this thesis.

Shirky (2007) criticizes the use of ontologies as a way of trying to enforce a structure on something that is by nature unstructured. He instead pushes the idea of common tagging. Part of the reason he criticizes the ontology approach is that it seem improbable that experts can know the needs of all the users a priori, and therefor that every ontology will prove to be inadequate. On the semantic web, ontologies are represented using collection of RDF(Resource Description Framework) triplets. These triplets are in the form of <subject, predicate, object>, much like simple declarative sentences (Berners-Lee et al., 2001). Each subject and predicate, and some objects, is represented by an URI(Universal Resource Identifier), which links to the resource that describes it.

While ontologies are formally constructed taxonomies, folksonomies are informal taxonomies generated by collecting tags or annotations from collaborative tagging systems on a given platform(Tang et al., 2009). Mika (2005) has given a more formal definition of folksonomy where he sees a folksonomy as a set of tags T, $T \subseteq A \times C \times I$, where A is the set of users tagging, C is the set of tags, and I is the set of objects being tagged. Gruber (2007) suggested that one should add the source of the tag, and some kind of rating system to help filter out junk tags. Kim et al. (2008) on the other hand suggests removing the objects being tagged from the ontology, seeing that the objects that are being described are not part of the tool to describe them. Instead they like Gruber want to add the source of the tag, the number or times a tag occur, and the tagging behavior of each user. Bang et al. (2008) explains the difference between ontologies and folksonomies by classifying them as a priori and a posteriori annotations. That is, ontologies are created by experts as ways on conceptualizing a domain, folksonomies on the other hand are samples of how people speak or think about a domain. Folksonomies grew as a subject of research as it became popular for users to tag content on the internet with keywords they felt were relevant.

For users tags are convenient, since Adding additional tags can make it easier for humans to search and browse collections. This is especially true for multimedia content, which we don't yet have good tools for searching in (Weinberger et al., 2008). Tags provide meta data about content in a way that makes sense to humans. From an information retrieval perspective this is interesting since it means that humans in some way add meaning to the content. There however are several problems with using tags as the basis of a semantic web. Tang et al. (2009) mention several. Tags are supposed to be written in natural language, and natural language has words that are synonymes(words that are written in different ways, but mean the same), homonymes(words with different meanings that are written in the same way), or polynyms(a word that can have several meanings) making it unsuitable for computer reasoning since they are ambiguous (Passant and Laublet, 2008). As Golder and Huberman (2005) mentions, users also operate on different levels of abstraction, which can make it harder to find interesting resources. In addition to this comes the problem of non dictionary words, both new, or compound words, or simply words that have been misspelled(Tonkin and Guy, 2006).

There has been done a lot of research into how one can lift semantic data out of these unstructured tags. Golder and Huberman (2005) has done research into the statistical analysis of tags. The analysis done here show that there seems to develop vocabularies of frequently used tags. This might help diminish the effect of misspelled and nonsense words. Similar findings were also reported by (Shirky, 2007)

There has also been done research into automatic clustering. Mika (2005) created clusters by creating weighted graphs, and compared using tag concurrence and actor interest as weights. Brooks and Montanez (2006) has done work an categorizing blogs entries by tags, to see if concurrence of tags indicated similar content. Using the most common tags did give some results, but only broad categories. The results were not better than extracting words that were given asserted to be relevant for the category.

(Tang et al., 2009) tries to go further that clustering tags, and tries to build an hierarchical model from a folksonomy. They use a probabilistic model that takes into account the frequency and concurrence of tags and tries to generalize it to an ontology. The method does get good results in creating the hierarchy, but does also show som inappropriate sub/super category inferences.

Weinberger et al. (2008) suggests a method for removing ambiguity from tags, by suggesting additional tags to the user when the tag entered can belong to one of several distinct sets

While there are many difficulties attached to merging the social and semantic web, and with lifting semantic data from tags, there are many researchers who stress the need for this (Passant, 2007; Mika, 2005; Gruber, 2007).

2.3 WordNet and lexitags

Lexitags (Veres, 2011) utilizes a different approach for getting semantic meaning out of tags that the approaches mentioned until now. Instead of analyzing existing folksonomies and try to lift semantic data out of these tags, the idea presented is to turn it around and make users attach meaning to the tags at input time. This is done by letting users disambiguate the tags by using WordNet synsets, an idea that was also mentioned by.

WordNet is a lexical reference system that stores words in sets of synonymes called synsets. The idea is to separate the word form from the word sense. The underlying assumption is that the user already knows English, is familiar with the concepts that are conveyed, and doesn't need definitions to understand, but can use synonymes to identify the meaning they want to convey(Miller et al., 1990).

In addition to storing these synsets WordNet also contains information about the semantic relationship between different concepts. The synonymy relationship is obviously contained within each synset, though it should be mentioned that the definition used in WordNet is not one where substitution never changes the truth value of a sentence. WordNet uses a weaker definition where two word forms can be seen as synonymes in relation to some semantic context. The antonymy relationship is another relationship between word forms. While the exact definition of antonymy is hard to pin down, the intuitive notion that an antonym to x is not-x will take us a long way(Miller et al., 1990).

WordNet also stores information about hyponymy and hypernymy, which is a relationship between concepts. A hypernym can be explained as a generalization of some concept, a hyponym on the other hand can be seen as a specialization of a concept. {tree} can for example be seen as a hyponym of {plant}, and the reverse relation is a hypernymy relation (Veres et al., 2010).

The fact that WordNet separates sense and form is good for our purposes, as we are interested in the sense, not the form of the word. Mapping tags to synsets removes the ambiguity that arrises from multiple spellings. At the same time, since the mapping preserves the form of the tag this can still be kept for analysis if one finds that there are significant differences in how different forms of a synset is used(Veres, 2011). By enforcing this mapping to WordNet lexitags also gets access to the hierarchical knowledge therein, and can create lightweight ontologies by using hypernyms of the tags as SuperTags, a method introduced by Veres et al. (2010). The mapping to WordNet also add some perks. There is a mapping between WordNet and Schema.org[1], and between WordNet and the SUMO(Suggested Upper Merged Ontology)(Niles and Pease, 2003).

Using WordNet to ground the semantics of the tags was idea also suggested by Cattuto et al. (2008). But Cattuto et al. (2008) suggested using a post hoc analysis of the tags in a

[1] `https://github.com/mhausenblas/schema-org-rdf`

social network, instead of enforcing the mapping though an interface.

One critique of WordNet comes from Mika (2005) who points out that while WordNet can catch lexical sameness, it lacks cultural awareness. The example used was that of the tie between Noah and the ark. This tie would be obvious for most humans, and would most lightly be caught through clustering tags, but would not be caught by WordNet.

Passant and Laublet (2008) has also suggested a system where taggs are disambiguated by the user at input time. A difference between the systems is that Passant and Laublet suggested using URIs to online resources for disambiguation.

References

Bang, B. H. K., Dané, E., and Grandbastien, M. (2008). Merging semantic and participative approaches for organising teachers' documents. *Proc Conf. Educational Multimedia, Hypermedia & Telecommunications, pages*, pages 4959–4966.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–+.

Brooks, C. H. and Montanez, N. (2006). Improved annotation of the blogopshere via autotagging and hierarchical clustering. *Proceedings of the 15th international conference on World Wide Web (S. 625-632). Edinburgh, Scotland: ACM.*

Cattuto, C., Benz, D., Hotho, A., and Stumme, G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. *The Semantic Web-ISWC 2008*, pages 615–631.

Golder, S. and Huberman, B. A. (2005). The Structure of Collaborative Tagging Systems. *Audio, Transactions of the IRE Professional Group on.*

Gruber, T. (2007). Ontology of Folksonomy: A Mash-up of Apples and Oranges. *International Journal on Semantic Web & Information Systems*, 3(2):1 – 11.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. In *International Journal of Human-Computer Studies*, volume 43, pages 907–928.

Kim, H. L., Decker, S., Scerri, S., Breslin, J. G., and Kim, H. G. (2008). The state of the art in tag ontologies: a semantic model for tagging and folksonomies. *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, pages 128–137.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. *The Semantic Web–ISWC 2005*, pages 522–536.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Niles, I. and Pease, A. (2003). Mapping WordNet to the SUMO ontology. *Proceedings of the IEEE International Knowledge Engineering conference*, pages 23–26.

Noy, N. F. and Hafner, C. D. (1997). The state of the art in ontology design: A survey and comparative review. *AI magazine*, 18(3):53.

Passant, A. (2007). Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. *International Conference on Weblogs and Social Media.*

Passant, A. and Laublet, P. (2008). Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China.*

Pretorius, A. J. (2004). Ontologies-Introduction and Overview. *Semantic technology and applications research laboratory.*

Shirky, C. (2007). Shirky: Ontology is Overrated – Categories, Links, and Tags.

Tang, J., Leung, H., Luo, Q., Chen, D., and Gong, J. (2009). Towards ontology learning from folksonomies. *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 2089–2094.

Tonkin, E. and Guy, M. (2006). Folksonomies: Tidying up tags. *D-Lib*, 12(1).

Veres, C. (2011). LexiTags: An Interlingua for the Social Semantic Web. In *Proceedings of the 11th Interational Semantic Web Conference ISWC2011*, pages 1–12.

Veres, C., Johansen, K., and Opdahl, A. L. (2010). Browsing and Visualizing Semantically Enriched Information Resources. In *2010 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, pages 968–973. IEEE.

Weinberger, K. Q., Slaney, M., and Van Zwol, R. (2008). Resolving tag ambiguity. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia.*