Deep Web Annotation Using Goal-Oriented Special Purpose Ontologies

(Position Paper)

Axel Hochstein, Michael Genesereth

Computer Science Department, Stanford University
Gates 250, Serra Mall, CA 94305, Stanford, U.S.A.
Gates 220, Serra Mall, CA 94305, Stanford, U.S.A.

axel.hochstein@cs.stanford.edu, genesereth@stanford.edu

Abstract

Current efforts in the semantic web area have not yet led to applications that deliver additional real-life value for web users. The main reason is a vicious circle of low populated concepts, low value for web users as well as missing incentives for annotating data. As a solution a pragmatic approach based on goal-oriented special purpose ontologies derived from the deep web in combination with annotation of instances in the deep web is presented. Finally, a use case is underpinning the idea.

Introduction and Motivation

Evolving semantic web technologies as well as first semantic web applications, bring forward the vision of a more powerful semantic web. However, real-life value for web users is still lacking. General purpose ontologies such as FOAF, SIOC and others do not yet generate a real add-on in comparison to typical web 2.0 pages. Also semantic knowledge bases such as OpenCyc, DBpedia or OpenCalais currently don't exceed the value of non-semantic knowledge bases such as Wikipedia.

Two reasons account for these facts: First, many concepts but only few instances are semantically described. Imagine a food retail consumer having the goal of creating his weekly shopping list. Concepts such as food item or shopping center don't really help unless instances of food items and shopping centers are attached to those concepts. Right now, there are no incentives for data owners to annotate their data, since web users don't use semantic applications. This will last as long as no value is generated by semantic web applications and no value is generated unless instance data is annotated – ending in a vicious circle.

The second reason for limited value of current semantic web applications is the lack of special purpose ontologies and is related to the first reason. In order to break the vicious circle described above, it is necessary to build semantic web applications that generate value, i.e. enable web users to achieve certain goals with a better performance than using traditional web applications. General purpose ontologies promote a top-down approach for building such applications with the promise of providing a solution for a broad range of goals within different domains. However, the effort for generating a critical mass of annotated instances increases with this approach, ending in web applications with low chances of generating value for a web user.

Goal-Oriented Deep Web Annotation

In this position paper a pragmatic research framework is presente, for generating limited scope but high value semantic web applications.

Instead of using a top-down approach, a bottom up approach is suggested by annotating existing instances to ontologies that are build based on existing data structures on the web. It is estimated that in 2002 the web of fixed web pages (surface web) consisted of 167 TB of data whereas the database driven websites that create web pages on demand (deep web) comprised 91,850 TB of data (Lyman & Varian 2003).

By retrieving the underlying concepts as well as the underlying data structures of the deep web, special purpose ontologies can be derived. Based on these ontologies, corresponding instances of the deep web can be retrieved and annotated, leading to highly populated concepts. Highly populated concepts lead to a high TermRank (Ding et al. 2005) making it easy to find the concept (e.g. with semantic search engines such as Swoogle) and increasing reuse of the concept.

A similar approach has been proposed and studied earlier (Handschuh et al. 2003a, Handschuh et al. 2003b, Volz et al. 2004). However, in addition to these studies it is suggested that in order to create semantic web applications with high real-life value, the deep web should be annotated in a specific enough, human-centered domain. The NAICS

(North American Industry Classification System) codes 51-81 (NAICS codes for service industries) on the lowest level are specific enough with the opportunity to extend to more generic domains, i.e. NAICS codes with fewer digits. Subsequently, for the chosen domain identify typical goals as well as corresponding plans of involved entities. For each plan or even for each action within these plans, crawl for sources of the deep web that potentially decrease the overall cost for reaching goals, derive corresponding special purpose ontologies, and annotate instances of the deep web. As specific enough domains have been chosen and highly populated concepts are generated chances for efficiently infer value-adding information increase with this approach.

As value is generated, web users within the chosen domain will increasingly use semantic web applications, building on derived special purpose ontologies. In return, this gives incentives for data owners to annotate their data to these special purpose ontologies, and as a consequence being considered by corresponding semantic web applications. Thus, reuse of these concepts even increases, more instances are annotated, and more powerful inferences can be made.

Use Case

As a use case we chose the domain with NAICS code 4451 “Grocery Stores” and as an involved entity we chose the grocery store consumer. Typical goals identified in a creative process are “consumer has food for one week in kitchen”, “consumer has food for today in kitchen”, “consumer learns about new products”, “consumer socializes”, “consumer saves money”, “consumer is healthy” and more. Corresponding plans for the goal “consumer has food in kitchen” are for example “create shopping list”, “drive to grocery store”, “pay for food” etc. For the plan “create shopping list” the following web pages from the deep web could be crawled in order to decrease cost of the plan:

- Web pages with reviews of recipes for getting highly rated recipes as well as corresponding ingredients (e.g. AllRecipes.com)
- Web pages with nutrition data of grocery items (e.g. Nutrient Data Laboratory)
- Web pages with coupon information (e.g. SmartSource.com)
- Web pages with reviews of grocery items (e.g. Amazon.com/Grocery)
- Web pages of local supermarkets for getting information about assortment and specials

Currently we are in the process of building special purpose ontologies for the selected domain as well as annotating instances of the deep web. For instance, the data structure of the Nutrient Data Laboratory served as a basis for developing a nutrition data ontology and was validated by several other nutrition data web pages. In addition based on data of the Nutrient Data Laboratory more than 7,000 instances of grocery items including their nutrition facts could be annotated to the concepts grocery item and nutrition fact leading to highly populated concepts.

Based on developed special purpose ontologies and its highly populated concepts, as a next step it is aimed to build value adding semantic web applications for the selected domain. In addition, built ontologies are extendable in the way that more generic domains such as retail trade (NAICS code 44-45) can be covered with the same approach leading to more powerful semantic web applications.

Conclusion and Further Research

A pragmatic approach for developing value adding semantic web applications has been presented in this position paper. We assume that with this approach critical masses can be generated that are necessary to create incentives for data owners to annotate data, leading to more powerful semantic web applications.

However, this approach still needs validation. Further research is planned for developing and implementing semantic web applications that are based on ontologies derived from the deep web as well as corresponding instances annotated from the deep web.

References

Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., and Kolari, P., 2005. *Finding and Ranking Knowledge on the Semantic Web*. Lecture Notes in Computer Science, Springer.

Handschuh, S., Staab, S., Volz, R., 2003a. *On deep annotation*. Proceedings of the 12th international conference on World Wide Web, ACM.

Handschuh, S., Volz, R., Staab, S., 2003b. *Annotation for the Deep Web*. IEEE Intelligent Systems, Volume 18, Issue 5, pp. 42-48.

Lyman, P., and Varian, H. R.. 2003. *How much information*. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on October 9, 2008.

Volz, R., Handschuh, S., Staab, S., Stojanovic, L., Stojanovic, N., 2004. *Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the Semantic Web*. Journal of Web Semantics, Volume 1, Issue 2, pp. 187-206.