Tackling the Curse of Prepayment – Collaborative Knowledge Formalization Beyond Lightweight

Valentin Zacharias[1], and Simone Braun[1]
[1] FZI, Research Center for Information Science, Haid-und-Neu Strasse 10-14, 76131 Karlsruhe, Germany
{zach, braun}@fzi.de

Abstract. This paper argues for collaborative incremental augmentation of text retrieval as an approach that can be used to immediately show the benefits of relatively heavyweight knowledge formalization in the context of Web 2.0 style collaborative knowledge formalization. Such an approach helps to overcome the "Curse of Prepayment"; i.e. the hitherto necessary very large initial investment in formalization tasks before any benefit of Semantic Web technologies is visible. Some initial ideas about the architecture of such a system are presented and it is placed within the overall emerging trend of "people powered search".

Keywords: Semantic Web, collaborative knowledge formalization, web 2.0, Semantic Wikis, People Powered Search

1 Introduction

The Curse of Prepayment is the Chicken-Egg problem of Semantic-Technologies: that Semantic technologies promise great functionality only after a large amount of knowledge is formalized. And that no one is willing to invest large amounts of money or time in formalization until the great functionality is visible or at least foreseeable.

Recently there has been a great interest in approaches that attempt to tackle this problem by adapting Web 2.0 ideas to make knowledge formalization collaborative, and very easy, cheap, and simple (e.g. [1,2]). In this way these approaches enable end users to successfully contribute to the creation of semantic structures. However, most of these approaches are restricted to very lightweight formalisms – there seems to be a lack of ideas how to extend these approaches to more powerful formalisms. This paper argues that the critical point that stops these approaches from adequately addressing heavyweight formalisms is – again - the Curse of Prepayment: that with these approaches an investment in (more) heavyweight formalization shows no immediate benefit. For example it is trivially possible to edit an OWL Full document in any Wiki by just uploading its XML representation, but there is nothing enabled by the continued development of this document; nowhere is it visible what kind of functionality is made possible by this formalization.

We present the "Collaborative Incremental Augmentation of Text Retrieval" as one approach that can be used to tackle this challenge. It stipulates to enable endusers to

collaboratively and incrementally extend a conventional search engine in the direction of question answering. In section 2 this paper starts with an examination of current approaches in this area and their attempts to tackle the Curse of Prepayment; the chicken-egg problem of Semantic technologies. In section 3, the five properties of simple, collaborative, incremental, partial and immediate are presented as critical in this respect. Section 4 then details the challenges of extending this kind of knowledge formalization to more heavyweight formalisms. Collaborative, incremental augmentation of text retrieval is introduced as one possible answer to this challenge in section 5; some ideas on its realization are contained in section 6. Finally the paper concludes with a short summary and a discussion of connections to related work.

2 Web 2.0 Knowledge Formalization and The Curse of Prepayment

The Curse of Prepayment is also often referred to as the Chicken-Egg problem of Semantic Web technologies: Semantic Web technologies promise great functionality once a large amount of knowledge is formalized. However, because knowledge formalization is difficult, often not well supported, and cumbersome, the investment beforehand needed to see any functionality is very large (cf. [3]). This is problematic, because users cannot learn from seeing the final effects of their changes, are not motivated from seeing growing functionality, and because organization may hesitate to make investments in new technologies when any visible success is very far off.

This is not a new observation and numerous approaches have emerged to address it – of particular interest here are approaches that try to harness Web 2.0 ideas for this task[1]**.** The assumption of these systems can be summarized as "*Maybe formalization can be made so simple and useful and distributed over so many people that people will do it for free*". These approaches can be roughly separated into three groups:

- **Social Semantic Tagging Systems:** Based on the observation that a large number of people are successfully creating structured data with tagging applications, these approaches try to extend these systems with a bit more structure, a bit more formality. Our own SOBOLEO[2] system [4], GroupMe [5], Int.ere.st [6], BibSonomy[3] [7], Fuzzzy[4] [8] and gnizr[5] are examples for these kinds of systems.
- **Semantic Wikis:** The second group of systems starts from the observation that people are spending large amounts of time creating semi-structured data in wikis. These system then try to give people the tools and the support such that they can create data with more structure, more formality. The Semantic Media Wiki[6] [9],

[1]Not mentioned here, but also important are research threads based on machine learning (automatically acquiring structure) and exposing pre-existing structure (e.g. exposing relational databases as SPARQL endpoints)

[2] http://www.soboleo.com

[3] http://www.bibsonomy.org

[4] http://www.fuzzzy.com

[5] http://gnizr.googlecode.com/

[6] http://semantic-mediawiki.org/

Freebase[7], IkeWiki [10] and MyOntology [2] are example for these kinds of systems.

- **Semantic Games with a Purpose:** The third, much smaller, group is inspired by the success of the gwap platform[8], based on the "Games with a Purpose" paradigm [11]. This platform offers games that – as a side effect – also create structured data for the computer. OntoGame[9] is the approach that realized this for the Semantic Web [12]. This approach stands very much apart from the other approaches because (from a user point of view) the goal of the formalization is the formalization itself. This very interesting approach will nevertheless always only be able to address a small subset of needs for formalization and will not be discussed further in this paper.

In the authors' view there are five closely related properties that give these Social Semantic Tagging and Semantic Wikis a chance to tackle the curse of prepayment:

- **Simple:** Formalization is simple, can be done with little training, little effort and not only by logic experts. For example compared to an traditional ontology engineering tool the SOBOLEO and the Semantic Media Wiki are very easy to use.
- **Collaborative:** Formalization can be done jointly in a group – in this way the cost is spread over multiple persons; the prepayment needed from every person is reduced. All Web 2.0 knowledge formalization approaches have collaboration at their core.
- **Incremental:** Not everything needs to be formalized at once, formalization can be done incrementally. With the Semantic Media Wiki system the user can introduce typed relations incrementally as time is available.
- **Partial:** The tools can work with data stores that are only partly formalized, that contain data at different levels of formality. Again in Semantic Media Wiki, for example, typed relations can co-exist with internal links.
- **Immediate:** Formalized data can be used immediately, immediately brings some benefit to the user. With SOBOLEO or BibSonomy the user has an immediate advantage from adding just one 'broader' relation between tags, because his sped up.

Together these five properties can be summarized as: "*Making Every Penny Count, Immediately*". There is an immediate benefit for formalizing even small parts; and because these systems are simple and collaborative, formalizing these small parts is relatively cheap.

Hence in the authors' opinion **this immediate benefit for formalizing even small parts lies at the core of these systems' success.** The exact nature of this benefit differs between systems, examples are:

- **Tables and less redundant data:** The unique selling point of the Semantic Media Wiki: as soon as just a few attribute values have been specified, these can be used to create tables and overview pages that before had to be maintained manually.

[7] http://www.freebase.com/
[8] http://www.gwap.com/gwap/
[9] http://www.ontogame.org/

- **Hierarchical Organization:** In systems like SOBOLEO or BibSonomy tags can be organized hierarchically, this allows for more effective maintenance of the tag repository as well as for more effective navigation and retrieval. This works after having just one such relation.
- **Advanced Search:** For example in the SOBOLEO system adding just one synonym for a tag/concept will already improve the search experience, searching for this synonym will then also consider the documents annotated with the topic.

The immediate benefit is very important because it enables users to learn about the effects of their changes, it can motivate volunteer contributors to continue and finally it can also provide the justification for a continued investment of an organization.

3 The Challenges of Heavyweight Formalization

However, all the 'immediate benefits' presented in the previous section are benefits from very lightweight formalizations:

- **Tables and less redundant data:** The automatically generated overview tables envisioned for Semantic Wikipedia [9] only depend on simple RDF triples.
- **Hierarchical Organization:** The hierarchical organization in BibSonomy depends on just one taxonomic relation without a formal semantic.
- **Advanced Search:** The semantic search of the SOBOLEO system depends only on taxonomic broader-narrower relations and labels.

None of the mentioned systems can show a comparable immediate benefit from e.g. adding rules, disjunction statements, or elaborate models with many different relations between entities. Further, the most powerful of these, the arbitrary queries supported by Semantic Media Wiki can only be used by users with relatively advanced knowledge about the data model and the query language.

Extending the mentioned systems in the direction of more heavyweight formalisms faces many challenges, such as (partially based on [13]):

- **Usability / Debuggability:** Formalisms such as OWL or First Order Logic are harder to understand, in particular faults are much harder to identify.
- **Robustness:** A single faulty statement added to a knowledge base with a millions of axioms can make the knowledge base inconsistent and thereby invalidate all conclusions. Unless this problem is tackled, open collaborative knowledge formalization is impossible.
- **Performance and the Language Expressivity / Performance Tradeoff:** Current reasoners for OWL Full or FOL could not support a continuously updated knowledge base of even a fraction of the size of Wikipedia; hence restrictions on language expressivity, not-sound or incomplete algorithms or some use of non-declarative languages would be needed.
- **Mixed Formality:** Incremental and partial formalization also means that the data store is never fully formalized; always contains data at different levels of formality. Again a challenge for current reasoning approaches.

In the opinion of the authors, however, all of these challenges are trumped by the Curse of Prepayment – the question about the immediate benefit of formalizing even small parts of a data store. What is to be gained from spending some time and/or

money from bringing a part of a data store to a highly formal level, how is this immediately visible to the editors? Knowing an answer to this question may then also allow to find answers to the tradeoffs implied by the challenges above, e.g. this may provide the justification to remove certain powerful but slow features from the knowledge representation language or help decide whether to keep soundness or completeness of the reasoning algorithms used (in cases where both cannot be achieved).

An answer to the Curse of Prepayment for more heavyweight formalism must provide a way to profit from these formalizations that is useful, understandable and immediately visible to the user. This answer needs to realize the five properties of simple, collaborative, incremental, partial, and immediate for heavyweight formalisms.

One way to utilize heavyweight formalism is the creation of question answering systems, i.e. systems that do not just point a user to a document but that rather provide direct answers to questions. However, so far it has been impossible to create question answering systems that can answer the majority of arbitrary user questions, leading to almost constant disappointment of users. A further problem is that the creation of question answering systems for even small domains is a very costly and time consuming process. Also by now users are used to keyword based queries and there is evidence that they prefer keyword based queries to full question answering [14].

The proposed approach stipulates the collaborative creation of a question answering system by incrementally extending a text retrieval system. In this way the question answering functionality can harness the highly formal knowledge, the information retrieval engine prevents disappointment of the users, and the collaboration distributes the cost down.

4 Collaborative, Incremental Augmentation of Text Retrieval

Collaborative, incremental augmentation of text retrieval means the stepwise extension of normal text retrieval in the direction of questions answering. One for one, frequent queries that users already pose to a system are identified and the data store is extended to allow the computation of direct answers to these questions. For examples the maintainers of a site notice that queries of the form “<country name> size” are often entered. They then extend the search engine to detect this pattern and add formalizations needed to directly answer it.

The stepwise augmentation of text retrieval is already visible in modern search engines. For example posing the query “weather Karlsruhe” to Yahoo returns not just pages containing this string but an actual weather report for the city of Karlsruhe. Searching with Microsoft Search and the query “5 EUR in yen” returns the amount of Yen that 5EUR can buy with this days exchange rate. Google even allows developers to extend its search via the subscribed links feature[10]. For example, users subscribed to a Wikimedia Data[11] search extension that pose the query “distance from Paris to Karlsruhe” get the correct result of 443km; a result created through a specific file that

[10] http://www.google.com/coop/subscribedlinks/

[11] http://www.google.com/coop/profile?user=01659747360823524154O

contains the locations of cities based on Wikipedia entries. Yahoo also allows for the extension of its search engine in a related way through the SearchMonkey[12] platform.

Shown above is another example of augmentation of test retrieval – here from the ask.com search engine in response to the query "china size".

This stepwise augmentation of text retrieval in the direction of question answering has a number of advantages:

- **Reasonable Expectations:** No current question answering technique can answer the majority of arbitrary formulated natural language queries. For this reason current question answering systems will answer most queries incorrectly – something very few users are willing to accept. With augmented text retrieval question answering is an added bonus that appears only in relatively well understood cases. It thereby avoids the trap of constantly disappointing the user's expectation.
- **Incremental and Partial:** Functionality to answer queries can be added step by step, possibly depending on the progression of the overall formalization of the data store. No large up-front investment is needed.
- **Immediate:** As soon as the functionality to answer one kind of queries is complete, it can become part of the search engine and improve the user experience.
- **Accepted Interface:** That the system builds on what is currently probably the most accepted interface for information search.

These advantages mirror many of the desired properties identified in the previous sections. What is missing from these systems, however, is the notion of simple and collaborative participation in the creation of these answers. Google's Subscribed Links and Yahoo's SearchMonkey do this to a certain extent, but only for developers that are willing to learn the respective protocols and formats.

We hence propose the collaborative incremental augmentation of text retrieval as the next target for collaborative (Web2.0 style) knowledge formalization approaches. We propose to show the immediate benefit of higher levels of formality by enabling users to incrementally extend an information retrieval engine into a question answering system.

5 Realization

This section details some initial ideas on the architecture and layout of such a system in order to further explain the notion of collaborative incremental augmentation of text retrieval. The section starts with an overview of the question answering process followed by thoughts on the core reasoning architecture.

[12] http://developer.yahoo.com/searchmonkey/

Query processing starts with the user entering a query, as an example the user might enter “china size”. In order for the system to be able to process a set queries of the form “<country name> size” in a common way, it must first detect that some part of the query refers to a country. For this detection step the system uses the data already entered into the system by the users, i.e. the names and synonyms of countries. For the example query “china size” the output of this step might consist of the following:

```
china, fzi/col#Peoples_Republik_of_China
       a physicalThing, a country, a state ...
```

This indicates that the “china” part of the query could be matched to an instance with the URI “fzi/col#Peoples_Republik_of_China that is known to be of types physicalThing, country and state.

In the next step the system matches the processed query against a number of ’templates’ collaboratively maintained in a wiki-like system. These templates specify the relation between queries and internal queries. One such template might be:

```
<#a type:physicalThing> size
=>
        <-    #a a physicalThing
              #a size ?b
```

This defines that a query consisting of the reference to an entity of type “physicalThing” followed by the string “size” is translated into a query of the form shown above. This query mainly looks for a triple of the form #a size ?b, where #a is the country from the query and ?b is the variable representing the searched value. Obviously graphical editors would be needed to support the user in the creation of these templates.

In a next step the query created in this way is processed by the system using all information available. The result of this query processing is then presented together with the result from a normal information retrieval system. Additional (user maintained) templates might support the presentation of results.

The actual processing of the query can be done using any kind of formalization, such as OWL ontologies, FOL axioms, rules or even specialized heuristics created in procedural programming languages. We think that the best approach is not one based on a monolithic knowledge base using only one general purpose reasoned, but rather one build from relatively large heterogeneous reasoning modules; some using DL reasoners, some executing procedural scripts and some using parameterized heuristics. The important aspect is, however, that the elements used by these reasoning modules are created collaborative by the users and that these reasoning modules in their use in the augmented text retrieval then show the benefit of having these highly formal elements immediately.

In this way the proposed system can iteratively grow from an information retrieval system into a question answering system that can use all kinds of heavyweight knowledge for query processing. E.g. the example query introduced above could be processed using mapping rules that mediate between different vocabularies; or it could profit from OWL based reasoning that lead to the inference that a particular entity is a physical thing.

6 Related Work

The presented idea is part of the broader trend of 'People Powered Search'[13]; a trend that tries to unify the search paradigm exemplified by Google with the open, social collaboration of delicious[14] and Wikipedia[15]. Examples for other approaches within this trend are Mahalo[16] and Wikia Search[17] that understand result pages as akin to wiki pages that can be edited. Further examples are 50matches[18] that only searches pages bookmarked in social bookmarking services and sproose[19] that allows voting for results.

Question Answering systems and natural language interfaces have been developed for more than 30 years [15,16], with recent years seeing again a rise in interest in these systems (e.g. [17,18,19,20]); this recent rise fueled by the availability of a plethora of lexical resources, upper level ontologies, of the shelf grammars and parsers and advances in databases and knowledge representation [17]. With AskJeeves the recent years even saw a (now aborted) attempt to bring question answering to mainstream web search. Our proposed approach differs from this strand of research in the following ways:

- **Collaboration:** That the functionality of the system is created during use by its users (and not before)
- **Incremental:** That functionality to answer some queries directly is added step by step. This is only possible because an information retrieval engine forms the backup.
- **Existing Queries:** That users are not encouraged to 'speak to the machine'; that rather queries done anyway are detected.

7 Conclusion

This paper has presented collaborative, incremental augmentation of text retrieval as one answer to the question of what can be the benefit for formalizing parts of a data

[13] Also known as ‚Human Powered Search' or ‚User Powered Search'
[14] http://www.delicious.com/
[15] http://www.wikipedia.org
[16] http://www.mahalo.com/
[17] http://search.wikia.com
[18] http://www.50matches.com/
[19] http://www.sproose.com/

store with more than very lightweight formalisms. In this sense this idea goes beyond existing Web 2.0 style collaborative knowledge formalization approaches that obtain all their direct benefit only from very lightweight formalizations.

The advantages of this approach are (1) that a question answering system is build incrementally, without raising unreasonable expectations (2) that an improvement can be shown almost immediately, after only a small initial investment and (3) that it builds on what is currently probably the most accepted interface for information search.

As the obvious next step we plan to implement this idea as an extension of the SOBOLEO system. This is part of our ongoing project to support all stages of our proposed Ontology Maturing process model for collaborative knowledge formalization.

References

1. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: The Ontology Maturing Approach to Collaborative and Work-Integrated Ontology Development: Evaluation Results and Future Directions. In: Proc. of the ESOE-Workshop at ISWC'07, CEUR Workshop Proc. Vol. 292 (2007) 5-18
2. Siorpaes, K., Hepp, M.: myontology: The marriage of ontology engineering and collective intelligence. In: Bridging the Gap between Semantic Web and Web 2.0 (SemNet 07). (2007) 127–138
3. Friedland, N., Allen P., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, J., Angele, J., Staab, S., Mönch, E., Oppermann, H., Wenke, D., Porter, B., Barker, K., Fan, J., Chaw, S.Y., Yeh, P., Tecuci, D., Clark, P.: Project Halo: Towards a digital Aristotle. AI Magazine, 29(4) (2004) 29-48
4. Zacharias, V., Braun, S. (2007). SOBOLEO – Social Bookmarking and Lightweight Ontology Engineering. In: Proc. of the Workshop on Social and Collaborative Construction of Structured Knowledge at WWW'07, CEUR Workshop Pro. Vol. 273 (2007)
5. Abel, F., Henze, F.M., Krause, D., Plappert & D. Siehndel, P.: Group Me! Where Semantic Web meets Web 2.0. In: Proc. of the 6th Int. Semantic Web Conf. (2007)
6. Kim, H.L., Yang, S.-K., Song, S.-J., Breslin, J.G. & Kim, H.-G.: Tag Mediated Society with SCOT Ontology. In: Proc. of the 5th Semantic Web Challenge at ISWC'07 (2007)
7. Hotho, A., Jäschke, R., Schmitz, C.; Stumme, G.: BibSonomy: A Social Bookmark and Publication Sharing System. In: CS-TIW'06. Aalborg: Aalborg University Press (2006)
8. Lachica, R., Karabeg, D.: Metadata creation in socio-semantic tagging systems: Towards holistic knowledge creation and interchange. In: Scaling Topic Maps. Topic Maps Research and Applications 2007, Springer, (2007)
9. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. In: Journal of Web Semantics 5/2007, pp. 251–261, Elsevier (2007)
10. Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In: 1st International Workshop on Semantic Technologies in Collaborative Applications (STICA 06), Manchester, UK, June (2006)
11. Von Ahn, L.: "Games with a Purpose," Computer, vol. 29, no. 6, 2006, pp. 92–94.
12. Siorpaes, K., Hepp, M.: Games with a Purpose for the Semantic Web. Intelligent Systems, IEEE, (23):50-60,2008

13. Krötzsch, M., Schaffert, S., Vrandecic, D.: Reasoning in Semantic Wikis. In: Proc. of the 3rd Reasoning Web Summer School, Dresden, Germany, vol. 4636 of LNCS, pp. 310--329. Springer (2007)
14. Reichert, M., Linckels, S., Meinel, C., Engel, T.: Student's perception of a semantic search engine, In IADIS Cognition and Exploratory Learning in Digital Age, Porto, Portugal, oo. 139-147 (2005)
15. Ogden, W., Bernick, P.: Using natural language interfaces. In Helander, M., editor, Handbook of Human-Computer Interaction. Elsevier (1996)
16. Copestake, A., Jones, K. S.: Natural language interfaces to databases. Knowledge Engineering Review, 5(4):225–249. Special Issue on the Applications of Natural Language Processing Techniques. 1989
17. Cimiano, P., Haase, P., Heizmann, J., Mantel, M.: Orakel: A portable natural language interface to knowledge bases. Technical report, Institute AIFB, University of Karlsruhe (2007)
18. Kaufmann, E., Bernstein, A., Fischer, L.: Nlp-reduce: A "naıve" but domainindependent natural language interface for querying ontologies. In: 4th ESWC, Innsbruck, A (2007)
19. Bernstein, A., Kaufmann, E., Kaiser, C.: Querying the semantic web with ginseng: A guided input natural language search engine. In: 15th Workshop on Information Technologies and Systems, Las Vegas, NV, pp. 112–126 (2005)
20. Frank, A., Krieger, H.-U., Xu, F., Uszkoreit, H., Crysmann, B., Jörg, B., and Schäfer, U.: Question answering from structured knowledge sources. Journal of Applied Logic, Special Issue on Questions and Answers: Theoretical and Applied Perspectives, 5(1):20–48 (2007)