ConTag: A Semantic Tag Recommendation System

Benjamin Adrian[1,2], **Leo Sauermann**[2], **Thomas Roth-Berghofer**[1,2]
[1](Knowledge-Based Systems Group, Department of Computer Science,
University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern)
[2](German Research Center for Artificial Intelligence DFKI GmbH,
Trippstadter Straße 122, 67663 Kaiserslautern Germany,
firstname.lastname@dfki.de)

Abstract: ConTag is an approach to generate semantic tag recommendations for documents based on Semantic Web ontologies and Web 2.0 services. We designed and implemented a process to normalize documents to RDF format, extract document topics using Web 2.0 services and finally match extracted topics to a Semantic Web ontology. Due to ConTag we are able to show that the information provided by Web 2.0 services in combination with a Semantic Web ontology enables the generation of relevant semantic tag recommendations for documents. The main contribution of this work is a semantic tag recommendation process based on a choreography of Web 2.0 services.
Key Words: Ontology, Web 2.0, Semantic Web, Social Software, Tagging
Category: H.1.1, H.3.3

1 Introduction

In this paper we describe ConTag, a recommendation system to tag or annotate documents with concepts of a Semantic Web ontology. In ConTag, Web 2.0 services providing text and term analysis functions such as phrase extraction, dictionaries, thesauri, classifications and term associations are used to extract the information content of a document. This approach shows that the convergence of Web 2.0 and Semantic Web is worthwhile regarding Web 2.0 tagging and Semantic Web ontologies. The information provided by Web 2.0 services combined with a Semantic Web ontology enables us to recommend semantic tags for documents.

In Section 2, we explain the state of the art of tagging in a Semantic Web environment. Section 3 describes the architecture of ConTag, including different possibilies of retrieving relevant similarities between document topics and ontology instances. Section 4 provides concrete implementation details. It illustrates the extraction of document topics based on Web 2.0 services and the recommendation of similar ontology instances as semantic tags. The evaluation in Section 5 confirms the statement that the information provided by Web 2.0 services in combination with a Semantic Web ontology enables the generation of relevant semantic tag recommendations for documents. Finally, Section 6 summarizes the approach and denotes future goals.

2 Related Work

ConTag generates tag recommendations based on an underlying Semantic Web ontology. The recommendations may be used, e.g in a Semantic Desktop application for classifying documents with a personal information model. Tag recommendations are generated by using existing Web 2.0 services. At the moment, we are not aware of any other system performing this task. Therefore we describe the state of the art of tagging in semantic environments.

The haystack project [Quan et al., 2003] was an early approach of Personal Information Management developed with Semantic Web techniques comparable to the Personal Information Model Ontology (PIMO) [Sauermann, 2006]. NEPOMUK - The Social Semantic Desktop[1] is a project using and building on experiences with *gnowsis* and the PIMO language/ontology. Tagging systems such as the bookmarking manager *del.icio.us*[2], the reference manager Connotea [Lund et al., 2005] or the photo sharing service *flickr*[3], enable users to annotate documents with self defined keywords called tags.

The studies [Golder and Huberman, 2005] and [Kipp and Campbell, 2006] point out patterns in tagging systems. Tags are more than just keywords but symbols for personal concepts. They also point out existing semantic difficulties such as managing polysemies and synonyms. In an analysis of tag usage, [Sen et al., 2006] demanded private tags in tagging systems to be used as personal concepts. Bridging the gap beteen tags and ontologies, the approach of [Schmitz, 2006] described the development of ontologies based on tag usages. The general problem of relating tags and ontologies based on social services is called Folksonomy [Wal, 2004]. In order to define tags in Semantic Web ontologies, Richard Newman introduced a first idea of a tagging ontology in [Newman, 2005]. Existing folksonomies are mined for association rules to retrieve semantic relations between tags using co-occurances [Schmitz et al., 2006]. PiggyBank [Huynh et al., 2005], CREAM [Handschuh and Staab, 2003] and Annotea [Kahan and Koivunen, 2001] provide RDF compliant tag or annotation repositories. [Bloehdorn and Hotho, 2004] describes techniques to optimize text classification using semantic information.

As a result of this state of the art analysis, it can be said that by now it is possible to annotate documents with tags, being symbols for personal concepts. These expressions may be stored as semantic relations in a semantic web ontology.

[1] `http://nepomuk.semanticdesktop.org`
[2] `http://del.icio.us`
[3] `http://www.flickr.com`

3 The semantic tag recommendation system ConTag

In order to generate tag recommendations we used concepts formalized in PIMO vocabulary. In PIMO, concepts are separated between the two classes `Thing` (e.g. persons, events, locations, etc.) and `ResourceManifestation` (music files, documents, etc). A relation `occurrence` connects `Thing`s to `ResourceManifestation`s, using the following semantic: *A thing occurs in a document*. Instances in a PIMO ontology are called *things*. Entities occurring in documents, are called *topics*. Expressing relevant similarities between things and topics may assume four different shapes in ConTag:

Equivalence A topic corresponds directly to a thing.

Classification If a topic's class corresponds directly to an ontology class, the topic is recommended as new thing of the ontology class.

Superordination If a topic's class does not correspond to any ontology class, the topic is recommended as new thing of a new ontology class.

Relation If a topic is semantically related to a thing without being equivalent, a suitable relationship between topic and thing should be proposed.

In the actual version of ConTag we focus on realising the similarity case *Equivalence*. Other semantic relations can be found in [Horak, 2006] and are discussed in future work.

Generally, the idea of using things as tags (instead of labels) entails some basic advantages. Things are identified by URIs and labeled by `rdfs:label` or alternative labels `pimo:altLabel`. This design overcomes existing semantic problems such as synonyms, homonyms, acronyms and different spelling, which current tagging systems suffer, by separating the tag's label from its identification. Additionally, things may possess a set of further describing RDF properties providing the capability to better retrieve similarities.

ConTag is based on a Semantic Tag Recommendation Process (see Fig. 1):

1. During the first step, *Normalisation*, the document's content is tranformed to RDF format to gain a fulltext description. We use the Aperture[4] framework to extract data and metadata such as author, creator and creation date.

2. During the second step, *Topic Extraction*, topics are extracted by requesting Web 2.0 services. This results in a *topic map* using SKOS vocabulary (Simple Knowledge Organisation System) [Miles and Brickley, 2005]. In succeeding lookup iterations, each topic entity is enriched by a set of semantic properties, such as definitions and synonyms.

[4] `http://www.aperture.sourceforge.net`

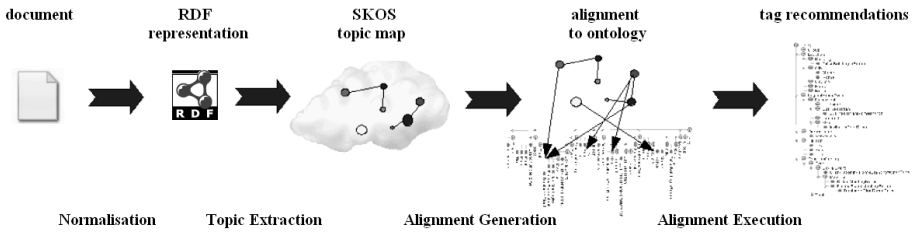

Figure 1: ConTag's Semantic Tag Recommendation Process

3. The *Alignment Generation* is based on document classification methods. For each topic in the topic map, several weighted alignment possibilities are computed to retrieve similar things.

4. The forth step is called *Alignment Execution*. The alignment scheme is visualized as tag recommendations. The user decides whether to accept or reject recommendations. Accepted recommendations are processed to: (1) create new `occurrence` relations in case of *Equivalence*, (2) create new instances in case of Classification, (3) create new classes in case of Superordination, and (4) create new relation types in cases of other semantic relations.

4 Implementation datails

The following sections describe parts of the Semantic Tag Recommendation Process, namely *Topic Extraction* and *Alignment Generation*. We used the RDF store Sesame 2 to manage ontologies in RDFS and topic maps in SKOS.

4.1 Topic Extraction

The topic extraction step is the most valuable step in the Tag Recommendation Process. It results in developing a document specific topic map by executing a Web 2.0 service choreography to extract document entities. The SKOS vocabulary distinguishes topics between instances and classes similar to PIMO language using relations (`broaderInstantive`, `narrowerInstantive`). Each topic possesses a name `prefLabel` and alternative labels `altLabel`. Each topic may be further explained by fulltext definitions written in natural language using `definition`.

The topic extraction step is based on querying Web 2.0 services. The choreography starts with extracting relevant keyphrases of the document. At the

moment Web 2.0 services such as Tagthe.net[5], Yahoo's Term Extraction service and Topicalizer[6] are used to extract relevant keyphrases. The results are stored into a document specific topic map.

In a succeeding iteration, for each topic in the topic map, three succeeding lookups request Web 2.0 services to gather for more information:

1. A *definition lookup* queries web dictionaries such as WordNet for existing definitions. These definitions are copied and attached to their grounding topics to be used in the succeeding hypernym extraction and to further provide explanations.

2. A succeeding *hypernym lookup* requests a self written hypernym extraction service called *DefTag*[7] to extract topic classes. These classes are stored as topics and link to instances using `broaderInstantive` and `narrower-Instantive` relations.

3. A third *association lookup* requests services for *word association*s concerning each topic. This lookup considers four different services at the moment: (1+2) Two web services hosted by Ontok Wikipedia provide an access to *Wikipedia Online Encyclopedia*, a collaborative web dictionary system. (3+4) Two web dictionary services (*Moby Thesaurus II*, *WordNet Dictionary*) are requested using the DICT protocol to extract a set of synonyms for a given term.

The topic extraction step results in a document specific topic map written in SKOS. It describes each topic with definitions and word associations. See [Horak, 2006] for more information about the used services.

4.2 Aligning topics to things

The *alignment generation* searches for similarities between topics and things. It results in an alignment scheme which is visualized as a list of tag recommendations. In order to express and weight similarities with confidence ratios, we used an ontology alignment vocabulary[8].

Due to a topological analysis of PIMO ontologies and document topic maps we assume that an ontology contains more entities than a topic map. Additionally, ontologies contain class hierarchies, whether topic maps are rather flat structured. Therefore we focussed on aligning topics to things by applying hierarchical document classification techniques instead of using topolical ontology matching methods. In this paper, we describe a rather simple alignment approach. Other appraches can be found in [Horak, 2006].

[5] `http://tagthe.net`
[6] `http://www.topicalizer.com`
[7] `http://www.dfki.uni-kl.de/~horak/2006/contag`
[8] `http://phaselibs.opendfki.de/wiki/AlignmentOntology`

To retrieve *equivalencies* between topics and things, we compared feature vectors using string matchings. A thing's feature vector is an aggregation of existing describing properties such as `label` or `altLabel`. A topic's feature vector is a list of extracted labels namely `prefLabel` and `altLabel`. We used SPARQL select queries with regular expressions to match both vectors and then computed a string similarity using the dice metric [Rijsbergen, 1979] to gain a confidence ratio. If this confidence ratio exceeds a threshold, an equivalence relation is created.

To retrieve *classification* relations we compared topic classes with thing classes using the upper defined method. If this comparison results in any matching *equivalence*, all concerning topic instances are recommended as new instances of the equivalent ontology class.

Finally, the *Alignment Execution* visualizes the resulting alignment scheme as tag recommendations. Each thing being involved in an alignment relation is concerned to be a tag for the document.

5 Evaluation

In order to provide further evidence for our statement that information provided by Web 2.0 services in combination with a Semantic Web ontology enables the recommendation of relevant semantic tags for documents, we evaluated ConTag by user ratings according to Precision and Recall [Rijsbergen, 1979] ratios. We used an existing ontology with information about projects, employees and partners in PIMO language. Then, we interviewed eight persons, working on subjects being described in the ontology. They got a summary of the ontology content and eleven documents with tag recommendations. The documents were web sites about employees or projects, existing as things in the ontology. The interviewees rated the quality of the given tag recommendations with Precision and Recall ratios (see Fig.2). As a result they rated recommended things of classes Projects, Persons and Organisations with Recall ratios above 80%, in general. These things were based on *Equivalences*. Things of class Location were rated with Recall ratios above 60%. These things were based on *Classifications*. Precision was rated above 70%, in general.

These ratios validate that the information provided by Web 2.0 services in combination with a Semantic Web ontology enables the generation of relevant semantic tag recommendations for documents. Following this result it can be said that the convergence of Web 2.0 and Semantic Web is worthwhile regarding Web 2.0 tagging and Semantic Web ontologies

More detailed evaluation results concerning Precision and Recall progressions in different configuration scenarios and the distribution of tagging recommendations in dynamic and nested class hierarchies are described in [Horak, 2006].

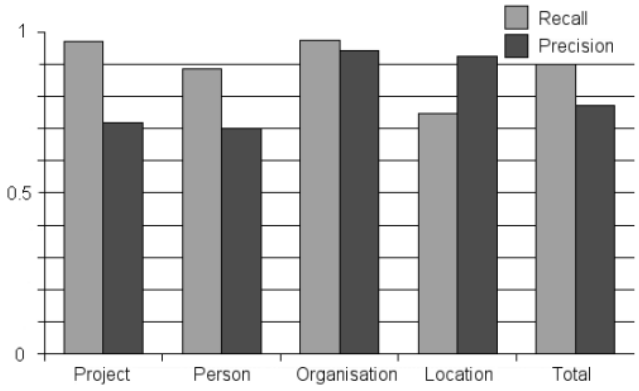

Figure 2: Precision and Recall ratios.

6 Conclusions and Outlook

In this paper we presented *ConTag*, a system to generate semantic tag recommendations for documents based on Semantic Web ontologies. We designed a process to normalize documents to RDF format, extract document topics using Web 2.0 services and finally match extracted topics to instances of a Semantic Web ontology. We use ontologies written in *PIMO language* to formalize instances and classes. Based on our evaluation, we provide evidence that it's possible to create relevant tag recommendations for documents by using Web 2.0 services in combination with a Semantic Web ontology. The implemented system is available under a GPL license for download at the first author's homepage[9].

In future work we want to look for additional similarity metrics to further enhance the alignment creation. The use of additional web services such as Google Glossary and existing tagging services providing accessible APIs is planned. At the moment ConTag concentrates on retrieving equivalent things occuring in documents and the ontology. In future work, we are going to further develop and enhance remaining similarity cases.

Acknowledgments

ConTag was awarded with the IBPM Award at IBPM Kongress 2007 (see `http://www.ibpmkongress.de`) for being the best diploma thesis in the category Document Management. We would like to thank all members of the knowledge management department for all their time, spent on discussions and support. This

[9] `http://www.dfki.uni-kl.de/~horak/2006/contag`

work was supported by the German Federal Ministry of Education, Science, Research and Technology (bmb+f), (Grant 01 IW C01, Project EPOS: Evolving Personal to Organizational Memories) and by the European Commission IST fund (Grant FP6-027705, project Nepomuk).

References

[Bloehdorn and Hotho, 2004] Bloehdorn, S. and Hotho, A. (2004). Boosting for text classification with semantic features. In *Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 70–87. Joint Session with WebKDD workshop. Reprinted in Proceedings of WebKDD 2004, LNCS 3932, Springer.

[Golder and Huberman, 2005] Golder, S. and Huberman, B. A. (2005). The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

[Handschuh and Staab, 2003] Handschuh, S. and Staab, S. (2003). Cream - creating metadata for the semantic web. *Computer Networks*, 42:579–598. Elsevier.

[Horak, 2006] Horak, B. (2006). Contag - a tagging system linking the semantic desktop with web 2.0. Diploma thesis, University Kaiserslautern, http://www.dfki.uni-kl.de/~horak/mypubs/ConTag.pdf.

[Huynh et al., 2005] Huynh, D., Mazzocchi, S., and Karger, D. (2005). Piggy bank: Experience the semantic web inside your web browser. In and Motta, E., Benjamins, V. R., and Musen, M. A., editors, *International Semantic Web Conference*.

[Kahan and Koivunen, 2001] Kahan, J. and Koivunen, M.-R. (2001). Annotea: an open RDF infrastructure for shared web annotations. In *Proceedings of the 10th International World Wide Web Conference*, pages 623–632.

[Kipp and Campbell, 2006] Kipp, M. E. I. and Campbell, D. G. (2006). Patterns and inconsistencies in collaborative tagging systems : An examination of tagging practices. In *Annual General Meeting of the American Society for Information Science and Technology*.

[Lund et al., 2005] Lund, B., Hammond, T., Flack, M., and Hannay, T. (2005). Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4).

[Miles and Brickley, 2005] Miles, A. and Brickley, D. (2005). SKOS core vocabulary specification. W3c working draft, World Wide Web Consortium.

[Newman, 2005] Newman, R. (2005). Tag ontology design. blog entry at http://www.holygoat.co.uk/projects/tags.

[Quan et al., 2003] Quan, D., Huynh, D., and Karger, D. R. (2003). Haystack: a platform for authoring end user semantic web applications. In *Second International Semantic Web Conference (ISWC2003), Proceedings*.

[Rijsbergen, 1979] Rijsbergen, C. J. v. (1979). *Information retrieval*. Butterworths, London, 2 edition.

[Sauermann, 2006] Sauermann, L. (2006). Pimo - a pim ontology for the semantic desktop. draft article at http://www.dfki.uni-kl.de/ sauermann/2006/01-pimo-report/pimOntologyLanguageReport.html.

[Schmitz et al., 2006] Schmitz, C., Hotho, A., Jaeschke, R., and Stumme, G. (2006). Mining association rules in folksonomies. In *Proc. IFCS 2006 Conference*, pages 261–270, Ljubljana.

[Schmitz, 2006] Schmitz, P. (2006). Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*.

[Sen et al., 2006] Sen, S., Lam, S. K. T., Rashid, A. M., Cosley, D., Frankowsk, D., Osterhouse, J., Harper, F. M., and Riedl, J. (2006). tagging, communities, vocabulary, evolution. In *Proceedings of CSCW 06*. ACM.

[Wal, 2004] Wal, T. V. (2004). Would we create hierarchies in a computing age? blog entry at http://www.vanderwal.net/random/entrysel.php?blog=1598.