

Leveraging traditional classification by social software approaches

Abstract of a thesis
(09.08.2007)

Author: Rolf Sint

Mentor at the University : O.Univ.-Prof. Dr. Wolfgang Pree

Mentor at Salzburg Research: DI Georg Güntner

Date of Assignment: July, 2007

Estimated Completion Date: December, 2007

Abstract

In his thesis entitled “*Leveraging traditional classification by social software approaches*”, the author focuses on the combination of two diverging approaches for the classification of content objects in the internet: Firstly, the traditional classification approach based on controlled vocabularies and secondly, the social tagging approach. The thesis will investigate typical application scenarios for both approaches, identify weaknesses and strengths and come up with a solution for combining the two approaches with the goal to semi-automatically enrich the classification space and efficiently support search and retrieval of content objects. Besides the theoretical part, the thesis will develop a prototypical implementation of an application allowing to align free user tags with traditional “closed” classification approaches and thereby enrich the classification scheme. Based on an evaluation initial conclusions about the efficiency and success of the proposed approach will be drawn.

A controlled vocabulary is a organized list of words and phrases. It consists of predefined terms that have been preselected by the designer of the controlled vocabulary. A lot of applications currently use terms from a controlled vocabulary to classify content objects in this “traditional” way to support navigation and search. The advantage of this approach is that the search can make use of broader and narrower terms relationships within the controlled vocabulary, if such relationships are defines. A controlled vocabulary built in such a way (i.e. a thesaurus or a taxonomies) represents a domain knowledge model for the application or an organization. The disadvantage of this method is that it needs a lot of maintenance effort because the controlled vocabulary must be adapted regularly by the knowledge engineer.

However, nowadays the general tendency in the Internet is different: User communities publish content objects and use free associated tags to annotate them (this is commonly referred to as “social software” or “social tagging”). Popular examples of this approach are Flickr, del.icio.us, and YouTube. In case of the social tagging approach no controlled vocabulary exists and therefore nearly no centralized maintenance is necessary. The disadvantage of this system is that free tagging may be unreliable and inconsistent: Without a controlled vocabulary there are no relationships between the tags and therefore the search for broader and narrower terms is impossible. Using an uncontrolled vocabulary for search, the results will only include the terms entered by the user.

In his thesis entitled “*Leveraging traditional classification by social software approaches*”, the author focuses on investigating and implementing an approach for content classification which combines the advantages of both methods. The thesis describes how web communities can help to develop a knowledge model designed for company environments by free tagging of content which needs fewer maintenance than a central controlled vocabulary. Furthermore, a way is introduced in which free classification tags from the user community can be used to extend a controlled vocabulary and implement relations between the different entries. A de-

monstrator will show the practical relevance of this method. This demonstrator shows how the enrichment of a controlled vocabulary can be supported by means of free tags and how the enriched controlled vocabulary can be used to search and retrieve content on different websites. The thesis of this approach is, that search and retrieval can be effectively supported because also narrower, broader, synonyms, translations and related entries relationships from a specific entry of the controlled vocabulary can be used in the queries.