Algorithmic Extraction of Tag Semantics

Mike Rohland[1] and Olga Streibel[1]

Networked Information Systems, Free University Berlin,
Königin-Luise-Str.24-26 , 14195 Berlin, Germany,
mike.rohland@googlemail.com, streibel@inf.fu-berlin.de

Abstract. Tagging as a process of giving meaning to information is a common technique for categorizing things on the Internet. We assume that, due to the simplicity and usability, tagging will play a major role in connecting things, events and situations on the future Internet. In this demo paper, we present our approach on algorithmic extraction of tag semantics (aETS). This work contributes to automatic, user-driven knowledge modeling.

1 Problems with "Semantics" based on Tags

According to [2], there are two main problems emerged in social tagging:
Ambiguity- which means having more than one meaning for a tag- reduces the precision of keyword based search in folksonomy tags. Therefore users searching for *atlas* retrieve relevant resources to *world atlas*, as well as results to *Atlas Mountains* in Africa. **Synonymy**- which means equal or similar meaning for tags- reduces the quantity of keyword based search in folksonomy tags. Users searching for *titan* should retrieve *atlas* as well, since due to mythology *Atlas* is an example of *titan.*(Fig.1)
In addition to folksonomies, there is a concept of Extreme Tagging Systems (ETS)[1]. ETS are an extension of common folksonomy since they allow to tag tags and to tag emerged relation between tags X,Y: X $\langle is-tagged-with \rangle$ Y. Extreme tagging extends folksonomy graph adding semantics to tags. It enables the use of user's own concepts in description of the meaning, e.g. the use of their "subjective" synonyms. This is useful for generating personal ontologies from the ETS graph[1]. However, these advantages bring obstacles in realizing ETS, that are: the high **user interaction** in the tagging process, and the **user-specific language** used for tags and relations description like users' synonyms.

To address the problems described above, different methods for the algorithmic extraction of semantic relations has been developed. In [3], a tripartite model of ontologies has been proposed. It enables the generating of broader/narrower semantic relations out of folksonomy data. The model has been refined for the extraction of polysemous relation in [4]. In [5], different algorithms for generating synonym relations have been compared. However, all of these methods are restricted to only one type of semantic relation. From the users point of view both relation types have an equal strong influence on the retrieval experience. Also as these methods were developed for folksonomies they do not take into account

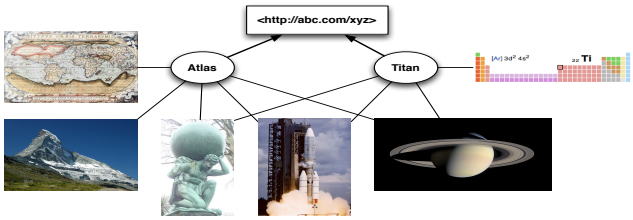

Fig. 1. Problems with "Semantics" based on Tags

the special characteristics of ETS. Hence, we focused on an integrative method that considers both types of relations and benefits from the ETS concept.

2 aETS Method

Our aETS method lessens the retrievability problems in folksonomies and user effort in ETS. It is based on a four-stepped process. Using the data contained in a folksonomy, our method develops an ontology build upon semantic relations between folksonomy tags (Fig.2)

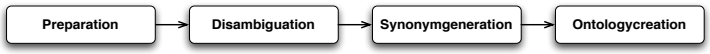

Fig. 2. Stages in aETS method

In the **preparation phase** the Jaro-Winkler-algorithm[6] is used on the set of tags to unify the spelling of the tag occurrences via gathering all different spelling forms of each tag. In the **disambiguation phase**, a bipartite graph is established for each tag that should be disambiguated (*dis-tag*). The bipartite graph contains *users:* u and *entities:* e as vertices. Tagging of an e with the *dis-tag* by u is represented by an edge in the graph. Applying one-mode projection[4] to the bipartite graph, results in new graph where vertices are represented by entities e and edges between them are transformed from tagging edges of the bipartite graph. The transformation reduces the graph by omitting u (entities tagged with the dis-tag by the same user are connected with the edge). It allows for representing tagged entities as a network of vertices due to users who tagged these entities with the dis-tag. Applying to the transformed graph the Girvan-Newman-algorithm[7] for cluster determination, determines clusters in graph; hence it determines possible meanings of the *dis-tag*. Entities that are included in one cluster are the considered as possible synonyms. In the **synonym extraction** step, the cosine similarity between each of the tags is used to determine synonyms. In the **generating ontology** step, the extracted semantic relations are recorded as a Semantic Web ontology.
We implemented the aETS method prototypically in Ruby and tested it relying

on Delicious[1] user tags. The first evaluation of aETS shows very promising results of extracting semantic relations out of tags. In order to validate the quality of retrieved relations and the usability of the generated ontology, we additionally evaluated the aETS results with test user group. The manual approach has shown that aETS is a very promising method for determining user-generated, language independent polysemes and synonyms. Table 1 shows selected aETS results whereas the interesting are the meaning sets for the *hra* tag. It shows three meaning sets for hra: the *insurance* context (in English hra is an abbreviation for *health reimbursement arrangaments*), the German *GmbH* context (part of Handelregister), and the Czech meaning for *radio* (hra is used to describe any performance).

Tag	Synonyms per meaning set
bridge	pics, amazing, image, top, 10, longest, wallpapers
	guide, linux, switch, howto, network, networking, kvm
	cars, city, vehicles, traffic, infrastructure, congestion, trucks, flow
	flex, fitness, pushup
green	hybrid, hydraulic, automobile, solar
	change, game, youth, fun, rules, activity
	garden
game	rpg, roleplaying, gaming
	smashbros, nintendo, brawl
hra	insurance, accounts, health
	GesellschaftsR, Unterlagen, GmbH, AG, AktienR
	rdio, esk, podcasting

Table 1. Exemplary meaning sets and synonyms detected by aETS for selected tags

References

1. Tanasescu, V., Streibel, O. Extreme Tagging: Emergent Semantics through the Tagging of Tags. Proc. Int. Workshop on Emergent Semantics and Ontology Evolution 292 (2007) 84-94
2. Golder, S.A., Huberman, B.A.: The Structure of Collaborative Tagging Systems. Jour. of Information Science 32(2) (2006) 198–208
3. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. The Semantic Web – Proc. Int. Semantic Web Conference 3729 (2005) 522–536
4. Au Yeung, C. Gibbins, N., Shadbolt, N.: Tag Meaning Disambiguation through Analysis of Tripartite Structure of Folksonomies. IEEE International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops (2007) 3–6
5. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems. Proc. of the 3rd Workshop on Ontology Learning and Population (OLP3) (2008) 39–43
6. Winkler, W. E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proc. of the Survey Research Methods Section(AMS) (1990) 354–359
7. Newman, M. E. J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2) (2004) 026113

[1] http://delicious.com/