Towards Mining Semantic Maturity in Social Bookmarking Systems

Martin Atzmueller[1], Dominik Benz[1], Andreas Hotho[2], and Gerd Stumme[1]

[1] Knowledge & Data Engineering Group, University of Kassel,
34121 Kassel, Germany
{atzmueller,benz,stumme}@cs.uni-kassel.de
[2] Data Mining & Information Retrieval Group, University of Würzburg,
97074 Würzburg, Germany
hotho@informatik.uni-wuerzburg.de

Abstract. The existence of emergent semantics within social metadata (such as tags in bookmarking systems) has been proven by a large number of successful approaches making the implicit semantic structures explicit. However, much less attention has been given to the *factors* which influence the "maturing" process of these structures over time. A natural hypothesis is that tags become semantically more and more mature whenever many users use them in the same contexts. This would allow to describe a tag by a specific and informative "semantic fingerprint" in the context of tagged resoures. However, the question of assessing the quality of such fingerprints has been seldomly addressed.
In this paper, we provide a systematic approach of mining semantic maturity profiles within folksonomy-based tag properties. Our ultimate goal is to provide a characterization of "mature tags". Additionally, we consider semantic information about the tags as a gold-standard source for the characterization of the collected results. Our initial results suggest that a suitable composition of tag properties allows the identification of more mature tag subsets. The presented work has implications for a number of problems related to social tagging systems, including tag ranking, tag recommendation, and the capturing of light-weight ontologies from tagging data.

1 Introduction

Social metadata, especially collaboratively created keywords or tags, form an integral part of many social applications such as BibSonomy[3], Delicious[4], or Flickr[5]. In such social systems, many studies of the development of the tagging structure have shown the presence of *emergent semantics* (e.g., [3]) in the set of human-annotated resources. That is, the semantics of tags develop gradually depending on their usage.

[3] http://www.bibsonomy.org
[4] http://www.delicious.com
[5] http://www.flickr.com

Due to this important observation, one can regard this development as a process of "semantic maturing". The basic idea is that knowledge about a set of cooccurring tags is sufficient for determining synonyms with a certain reliability. The underlying assumption is that tags become "mature" after a certain amount of usage. This maturity will then be reflected in a stable semantic profile. Thus, tags that have arrived at this stage can be regarded as high-quality tags, concerning their encoded amount of emergent semantics.

In this paper, we utilize folksonomy-based tag properties for mining profiles indicating "matured tags", i.e., high-quality tags that can be considered to convey more precise semantics according to their usage contexts. The proposed properties consist of various structural properties of the tagging data. e.g., centrality, or frequency properties. For a semantic grounding, we analyze the applied tagging data with respect to tag-tag relations in Wordnet, for assessing the "true" semantic quality. Our contribution is thus three-fold: We provide and discuss different tag properties that are useful in determining semantic maturity profiles of tags. These are all obtained considering the network structure of folksonomies. Additionally, we obtain a detailed statistical characterization of semantic tag maturity profiles in a folksonomy dataset. Finally, we provide a list of useful indicators for identifying "mature tags" as well as synonyms in this context.

Applications of the obtained knowledge concern the construction of lightweight ontologies using tagging knowledge [18], tag recommendation [14,19], or tag ranking [16]. All of these utilize selection options and/or ranking information about sets of tags, for initial setup and refinement. Tag ranking approaches, for example, can benefit from a "maturity ranking" for filtering purposes.

The rest of the paper is structured as follows: Section 2 discusses related work. After that, Section 3 introduces basic notions of the presented approach, including folksonomy-based tag properties, and the applied pattern mining method. Then, we describe the mining methodology in detail, discuss our evaluation setting and present the obtained results. Finally, Section 5 concludes the paper with a summary and interesting directions for future research.

2 Related Work

While the phenomenon of collaborative tagging was discussed in its early stages mainly in newsgroups or mailing lists (e.g. [17]), a first systematic analysis was performed by [10]. One core finding was that the openness and uncontrolledness of these systems did not give rise to a "tag chaos", but led on the contrary to the development of stable patterns in tag proportions assigned to a given resource. [5] reported similar results and denoted the emerging patterns as *"semantic fingerprints"* of resources. [18] presented an approach to capture emergent semantics from a folksonomy by deriving lightweight ontologies. In the sequel, several methods of capturing emergent semantics in the form of (i) tag taxonomies [12], (ii) measures of semantic tag relatedness [6], (iii) tag clusterings [22] and (iv) mapping tags to concepts in existing ontologies [1] were proposed.

Most of the above works provided evidence for the *existence* of emergent tag semantics by making certain aspects of it explicit; however, the question

which *factors* influcence its development were seldomly discussed. Despite that, a common perception seemed to be that a certain amount of data is necessary for getting a "signal". Golder and Hubermann [10] gave a rough estimate that *"after the first 100 or so bookmarks"*, the proportions of tags assigned to a resource tended to stabilize. This suggested the rule "the more data, the better semantics". This assumption was partially confirmed by Körner et al. [15], who analyzed the amount of emergent semantics contained in different folksonomy partitions. More data had a beneficial effect, but the *user composition* within the partitions turned out to be crucial as well: Sub-folksonomies induced by so-called "describers", which exhibit a certain kind of tag usage pattern, proved to contain semantic structures of higher quality. Halpin [11] showed that the tag distribution at resources tends to stabilize quickly into a power-law, as a kind of "maturing" of resources. In contrast, our work targets the maturing of tags themselves.

Hovever, to the best of our knowledge none of the aforementioned works has systematically addressed the question if there exists a connection between *structural* properties of tags and the quality of semantics they encode (i.e. their "semantic maturity"). In this work, we aim to fill this gap.

3 Preliminaries

In the following sections, we first briefly present a formal folksonomy model and a folksonomy-based measure of tag relatedness. Then, we detail on the structural and statistical tag properties serving as a basis for mining maturity profiles. After that, we briefly summarize the basics of the applied pattern mining technique.

3.1 Folksonomies and Semantic Tag Relatedness

The underlying data structure of collaborative tagging systems is called *folksonomy*; according to [13], a folksonomy is a tuple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, respectively. Y is a ternary relation between them, i.e. $Y \subseteq U \times T \times R$. An element $y \in Y$ is called a *tag assignment* or TAS. A *post* is a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, and a non-empty set $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$.

Folksonomies introduce various kinds of relations among their contained lexical items. A typical example are cooccurrence networks, which constitute an aggregation indicating which tags occur together. Given a folksonomy (U, T, R, Y), one can define the post-based *tag-tag cooccurrence graph* as $\mathbb{G}_{cooc} = (T, E, w)$, whose set of vertices corresponds to the set T of tags. Two tags t_1 and t_2 are connected by an edge, iff there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The *weight* of this edge is given by the number of posts that contain both t_1 and t_2, i.e. $w(t_1, t_2) := \mathrm{card}\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\}$

For assessing the semantic relatedness between tags we apply the *resource context similarity* (cf. [6]) computed in the vector space $\mathbb{R}^R$. For a tag t, the vector $v_t \in \mathbb{R}^R$ counts how often the tag t is used for annotating a certain resource $r \in R$:

$$v_{tr} = \mathrm{card}\{u \in U \mid (u, t, r) \in Y\}\,.$$

Based on this representation, we measure vector similarity by using the cosine measure, as is customary in Information Retrieval: If two tags t_1 and t_2 are represented by $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^X$, their cosine similarity is defined as: $\text{cossim}(t_1, t_2) := \cos \measuredangle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1||_2 \cdot ||\boldsymbol{v}_2||_2}$. In prior work, we showed that this measure comes close to what humans perceive as semantically related [6].

3.2 Folksonomy-Based Tag Properties

For folksonomy-based tag properties, we can utilize aggregated information such as frequency, but also properties based on the network structure of the tag-tag co-occurrence graph. The properties below are based on prior work in related areas. They are abstract in that sense, that none of them considers the textual content of a tag. Therefore, all properties are language independent since the only operate on the folksonomy structure, on aggregated information, or on derived networks. Below, we describe the different folksonomy-based properties, and also discuss their intuitive role regarding the assessment of tag maturity.

Centrality Properties In network theory the centrality of a node $v \in V$ in a network G is usually an indication of how important the vertex is [20]. Because important nodes are usually well-connected within the network, one can hypothesize that this connectedness corresponds to a well-established semantic fingerprint. On the other hand, high centrality might correspond to a relatively "broad" meaning – in the context of our study, we avoid the latter by restricting ourselves to single-sense tags (see Section 4). Applied to our problem at hand, we interpret centrality as a measure of maturity, following the intuition that more mature terms are also more "important". We adopted three standard centralities (degree, closeness, betweenness). All of them can be applied to a term graph $\mathbb{G}$:

- According to *betweenness centrality* a vertex has a high centrality if it can be found on many shortest paths between other vertex pairs:

$$bet(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

  Hereby, σ_{st} denotes the number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t passing through v. As its computation is obviously very expensive, it is often approximated [4] by calculating the shortest paths only between a fraction of points.
  It seems intuitive, that tags with a high betweenness centrality are closer to important (semantic) hubs, and therefore more mature themselves. In essence, higher values should indicate semantic maturity.
- A vertex ranks higher according to *closeness centrality* the shorter its shortest path length to all other reachable nodes is:

$$clos(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v, t)} \tag{2}$$

$d_G(v,t)$ denotes hereby the geodesic distance (shortest path) between the vertices v and t. A tag with a high closeness value is therefore more close to the core of the Folksonomy. Therefore, it seems intuitive to assume, that more central tags according to this measure should have a higher probability of being more mature.

- The *degree centrality* simply counts the number of direct neighbors $d(v)$ of a vertex v in a graph $G = (V, E)$:

$$deg(v) = \frac{d(v)}{|V| - 1} \tag{3}$$

Comparred to the other metrics, degree centrality is a local measure since it only takes into account the direct neighbourhood of a tag within the network. According to the degree, a tag could be linked to both semantically mature and non-mature tags. In this sense, it seems intuitive to assume that other factors need to be taken into account; then an estimation of the effect of the degree centrality can be considered.

Frequency Properties One first idea about tag maturity considers the fact that tags that are used more often *can* get more mature, since they can exhibit a more specific fingerprint. However, this does not guarantee maturity of tags. Therefore, we consider the frequency of a tag as a candidate for the analysis.

- We capture the *resource frequency* property *rfreq* which counts the number of resources tagged by a given tag t according to

$$rfreq(t) = \text{card}\{r : \exists(u, t', r) \in Y, t = t'\} \tag{4}$$

For the semantic assessment of tag, an intuitive hypothesis could be that the semantic profile of a tag gets more concise when more and more resource are tagged with it. However, this is not necessarily a criterion for mature tags since the development of the semantic profile could still be relatively fuzzy.

- The *user frequency* property *ufreq* counts the number of users that applied the tag t:

$$ufreq(t) = \text{card}\{u : \exists(u, t', r) \in Y, t = t'\} \tag{5}$$

Similar to the resource frequency, more users should help to focus the semantic profile of a tag due to the refinement of its usage patterns.

3.3 Pattern Mining using Subgroup Discovery

Subgroup discovery [21,2] aims at identifying interesting patterns with respect to a given target property of interest according to a specific interesting measure. In our context, the target property is given by a quality indicator for tags. The top patterns are then ranked according to the given interesting measure. Subgroup discovery is especially suited for identifying local patterns in the data, that is, *nuggets* that hold for specific subsets: It can uncover hidden relations captured in small subgroups, for which variables are only significantly correlated in these subgroups.

Formally, a database $D = (I, A)$ is given by a set of individuals I (tags) and a set of attributes A (i.e., tag properties). A *selector* or *basic pattern* $sel_{a=a_j}$ is a boolean function $I \to \{0, 1\}$ that is true, iff the value of attribute a is a_j for this individual. For a numeric attribute a_{num} selectors $sel_{a \in [min_j; max_j]}$ can be defined analogously for each interval $[min_j; max_j]$ in the domain of a_{num}. In this case, the respective boolean function is set to true, iff the value of attribute a_{num} is in the respective range.

A *subgroup description* or (complex) *pattern* $p = \{sel_1, \ldots, sel_d\}$ is then given by a set of basic patterns, which is interpreted as a conjunction, i.e., $p(I) = sel_1 \wedge \ldots \wedge sel_d$. A subgroup (extension) sg_p is now given by the set of individuals $sg_p = \{i \in I | p(i) = true\} := ext(p)$ which are covered by the subgroup description p. A subgroup discovery task can now be specified by a 5-tuple (D, C, S, Q, k). The target concept $C \colon I \to \Re$ specifies the property of interest. It is a function, that maps each instance in the dataset to a target value c. It can be binary (e.g., the quality of the tag is high or low), but can use arbitrary target values (e.g, the continuous quality of a given tag according to a certain measure). The search space 2^S is defined by set of basic patterns S. Given the dataset D and target concept c, the quality function $Q \colon 2^S \to \mathbb{R}$ maps every pattern in the search space to a real number that reflects the interestingness of a pattern. Finally, the integer k gives the number of returned patterns of this task. Thus, the result of a subgroup discovery task is the set of k subgroup descriptions $res_1, \ldots, res_k$ with the highest interestingness according to the quality function. Each of these descriptions could be reformulated as a rule $res_i \to c$.

While a huge amount of quality functions has been proposed in literature, cf. [9], many interesting measures trade-off the size $|ext(p)|$ of a subgroup and the deviation $c - c_0$, where c is the average value of the target concept in the subgroup and c_0 the average value of the target concept in the general population.

We consider the quality function *lift*, which measures just the increase of the average value of c in the subgroup compared to the general population:

$$lift(p) = \frac{c}{c_0}\,, \text{if } |ext(p)| \geq \mathcal{T}_{Supp}\,, \text{ and 0 otherwise}\,.$$

with an adequate minimal support threshold $\mathcal{T}_{Supp}$ considering the size of the subgroup. Usually, the analysis is performed using different minimal size thresholds in an explorative way. It is easy to see, that both types of quality measures are applicable for binary and continuous target concepts.

4 Mining Semantic Tag Maturity

For a given Folksonomy and its tagging dataset, we apply the following steps: Using the dataset, we construct the tag properties discussed in Section 3.2. As we will see below, the "raw" properties do not correlate sufficiently with semantic maturity. Therefore, we consider the dataset at the level of high-quality subgroups of semantically matured tags, and apply pattern mining using the *lift* quality function for this task. As an evaluation, we apply a gold-standard measure of semantic relatedness derived from WordNet [8].

4.1 Methodology

For the purpose of assessing the degree of semantic maturity of a given tag, a crucial question is how to measure this degree in a reliable and semantically grounded manner. In prior work [6] we identified folksonomy-based measures of semantic relatedness, which are among others able to detect potential synonym tags for a given tag. The most precise measure we found was the *resource context relatedness*, which is computed in the vector space $\mathbb{R}^R$. For a tag t, the vector $\boldsymbol{v}_t \in \mathbb{R}^R$ is constructed by counting how often a tag t is used to annotate a certain resource $r \in R$: $v_{tr} := \text{card}\{u \in U \mid (u, t, r) \in Y\}$. This vector representation can be interpreted as a "semantic fingerprint" of a given tag, based on its distribution over all resources. Our intuition for capturing the degree of maturity is based on the following argumentation chain:

1. The better the semantic fingerprint of a tag t reflects the meaning of t, the higher is the probability that the resource context relatedness yields "true" synomyms or semantically closely related tags $t_{sim1}, t_{sim2}, ...$ for t
2. If the most related potential synonym tag t_{sim1} is a "true" synonym of t (as grounded against the WordNet synset hierarchy), then the semantic fingerprint of t is regarded as semantically mature.
3. Otherwise, we consider the similarity in WordNet between t and t_{sim1} as an indicator for the maturity of the tag.

Please note, that we are using purely folksonomy-based measures (i.e., resource context relatedness) as a proxy for semantic similarity, because WordNet is not available for all tags. Simply spoken, this approach regards a tag as semantically mature if the information encoded in its resource context vector suffices to identify other tags with the same meaning. Naturally, this requires the existence of a sufficiently similar tag, which cannot be guaranteed. Therefore, this is not a sufficient but a necessary criterion. However, we think that the approach is justified, because the process of maturing is not restricted to isolated tags, but takes place similar to a "co-evolution" among several tags belonging to a certain domain of interest. As an example, if the topic of *semantic web* is very popular, then a relatively broad vocabulary to describe this concept will emerge, e.g. `semantic_web`, `semanticweb`, `semweb`, `sw`, ... In such a case, the maturity of a single tag would "correlate" with the existence of semantically similar tags within the same domain of interest. In general, it is important to notice that our methodology is also applicable to narrow folksonomies when replacing the resource context relatedness with the tag context relatedness (see [6]).

4.2 Semantic Considerations

For assessing the semantic similarity between tags we apply WordNet [8], a semantic lexicon of the English language. WordNet groups words into *synsets*, i.e., sets of synonyms that represent one concept. These synsets are nodes in a network; links between these represent semantic relations. WordNet provides a distinct network structure for each syntactic category (nouns, verbs, adjectives and

adverbs). For nouns and verbs, it is possible to restrict the links in the network to (directed) *is-a* relationships only, therefore a subsumption hierarchy can be defined. The *is-a* relation connects a *hyponym* (more specific synset) to a *hypernym* (more general synset). A synset can have multiple hypernyms, so that the graph is not a tree, but a directed acyclic graph. Since the *is-a* WordNet network for nouns and verbs consists of several disconnected hierarchies, it is useful to add a fake top-level node subsuming all the roots of those hierarchies; the graph is then fully connected so that several graph-based similarity metrics between pairs of nouns and pairs of verbs can be defined. In WordNet, we measure the semantic similarity using the taxonomic shortest-path length *dist*; the WordNet similarity $wns = 1 - \frac{dist}{max_{dist}}$ is then normalized using the maximum distance max_{dist}.

In addition to the WordNet similarity, we consider two additional indicators:

- The *Maturity Indicator* (mat) is a binary feature and measures if a tag has reached a certain maturity according to the WordNet information, i.e., the indicator is true, if we observe a WordNet similarity $wns \geq 0.5$.
- The *Synonym-Indicator* (syn) is a binary feature that specifies, if a tag-pair is in a synonym relation, i.e., the WordNet similarity $wns = 1$.

Since we consider the semantic fingerprint of tags using folksonomy information, we restrict the analysis to WordNet terms with only one sense; otherwise advanced word-sense disambiguation would be necessary in order to compare the correct senses in the WordNet synsets.

4.3 Dataset

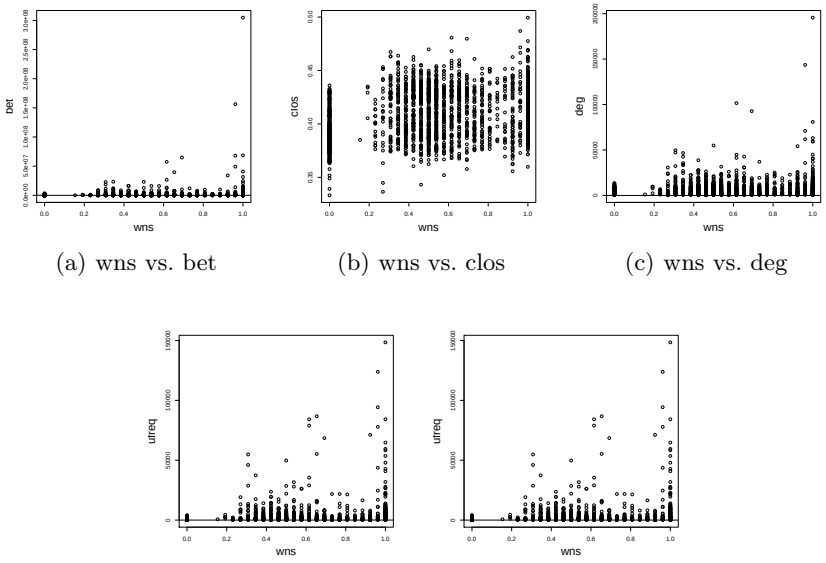
For our experiments we used data from the social bookmarking system del.icio.us, collected in November 2006. In total, data from 667,128 users of the del.icio.us community were collected, comprising 2,454,546 tags, 18,782,132 resources, and 140,333,714 tag assignments. For the specific purpose of our papers, some preprocessing and filtering was necessary: For the purpose of "grounding" the true semantic content of a tag t, we are applying vector-based measures to compute similar tags t_{sim}. Hence, we must assure that (i) the vector representation is dense enough to yield meaningful similarity judgements and (ii) there exist sufficiently similar tags t_{sim}. For these reasons, we first restrict our dataset to the 10.000 most frequent tags of delicious (and to the resources/users that have been associated with at least one of those tags). The restricted folksonomy consists of $|U| = 476,378$ users, $|T| = 10,000$ tags, $|R| = 12,660,470$ resources, and $|Y| = 101,491,722$ tag assignments. In order to assure the existence of sufficient "similarity partners" for each tag, we filter all tags whose cosine similarity to their most similar tag is lower than 0.05. As a last step, we only considered tags with exactly a single sense in WordNet in order to eliminate the influence of ambiguity. After all filtering steps, we considered a total of 1944 tags. We are aware that this is a strong limitation regardint the number of considered tags – however, because the problem at hand as well as our experimental methodology is sensitive towards a number of factors (like ambiguity or folksonomy-based similarity judgements), our focus is to start with a very "clean" subset. As a followup, it would of course be interesting to include more tags given the results on the clean subset are promising.

Table 1. Correlation between WordNet Similarity (wns), Maturity Indicator (mat), Synonym-Indicator (syn) and the different tag properties.

	bet	*clos*	*deg*	*rfreq*	*ufreq*
wns	0.15	0.20	0.20	0.21	0.18
mat	0.09	0.14	0.12	0.15	0.12
syn	0.12	0.14	0.13	0.15	0.15

We calculated all tag properties given the described co-ocurrence network, and discretized these using the standard MDL method of Fayyad & Irani [7] considering the WordNet similarity as a target class.

Statistical Characterization Figure 1 and Table 1 provide a first glance on the applied data. Each circle in Figure 1 represents one of the 1944 tags. Concerning the WordNet similarity (wns), we observe, that there is little correlation with the tag properties; Furthermore, we observe even lower correlations considering the two indicators *mat* and *syn*. Therefore, pattern mining using subgroup discovery is very suited for mining semantic tag profiles, since it also considers correlations in rather small subgroups described by combinations of different influence factors.



(a) wns vs. bet (b) wns vs. clos (c) wns vs. deg

(d) wns vs. rfreq (e) wns vs. ufreq

Fig. 1. Scatterplots for the WordNet Similarity (wns) vs. different tag properties.

4.4 Results

We applied pattern mining for the presented dataset using the tag properties as attributes, and the target concepts (wns, mat, syn) discussed above. Concerning the Wordnet Similarity (wns) and the *lift* quality function with a minimal subgroup size $n = 40$, we obtained the top patterns shown in Table 2. Lines 1-10

Table 2. Top patterns for target concept wns, split according to the different lengths of the patterns (mean in dataset: 0.54).

#	lift	mean	size	pattern
1	1.43	0.78	46	$ufreq > 13.0\%$
2	1.28	0.70	77	$rfreq > 3.0\%$
3	1.26	0.69	162	$clos > 64.7\%$
4	1.24	0.68	275	$deg > 6.0\%$
5	1.23	0.67	140	$bet > 1.0\%$
6	1.19	0.65	523	$ufreq > 1.0\%$
7	1.15	0.63	761	$rfreq > 0.1\%$
8	1.15	0.63	627	$bet > 0.2\%$
9	1.13	0.62	871	$clos > 47.0\%$
10	1.05	0.57	1519	$deg > 1.0\%$
11	1.32	0.72	51	$bet \in [0.2\%, 1.0\%]$ AND $clos > 64.7\%$
12	1.28	0.70	231	$deg > 6.0\%$ AND $ufreq > 1.0\%$
13	1.28	0.70	246	$deg > 6.0\%$ AND $rfreq > 0.1\%$
14	1.33	0.73	74	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $ufreq > 1.0\%$
15	1.30	0.71	119	$bet \in [0.2\%, 1.0\%]$ AND $deg > 6.0\%$ AND $rfreq > 0.1\%$

Table 3. Top patterns for the target concept "Maturity Indicator" (mean: 0.59)

#	lift	p	size	pattern
1	1.52	0.91	44	$ufreq > 13.0\%$ AND $clos > 64.7\%$
2	1.49	0.89	46	$ufreq > 13.0\%$
3	1.33	0.80	73	$rfreq > 3.0\%$ AND $clos > 64.7\%$
4	1.31	0.78	77	$rfreq > 3.0\%$
5	1.25	0.75	231	$deg > 6.0\%$ AND $ufreq > 1.0\%$
6	1.24	0.74	246	$deg > 6.0\%$ AND $rfreq > 0.1\%$
7	1.21	0.72	115	$bet \in [0.03\%, 1.0\%]$ AND $ufreq > 1.0\%$
8	1.21	0.72	275	$deg > 6.0\%$
9	1.20	0.72	162	$clos > 64.7\%$
10	1.18	0.70	588	$clos > 47.0\%$ AND $rfreq > 0.1\%$
11	1.36	0.81	74	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $ufreq > 1.0\%$
12	1.33	0.80	86	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $rfreq > 0.1\%$
13	1.30	0.77	105	$bet \in [0.03\%, 1.0\%]$ AND $deg > 6.0\%$ AND $ufreq > 1.0\%$
14	1.26	0.75	108	$clos \in [53.0\%, 64.7\%]$ AND $deg > 6.0\%$ AND $ufreq > 554$

show only basic patterns (one selector), while the lines 11-15 indicate more complex patterns. These results show that high betweenness and high closeness as intuitively expected. The influence of the degree centrality is not as prounounced as the other centralities, while higher degree also improves semantic maturity. Furthermore, a relatively high user frequency seems like the best indicator for high quality tags. Additionally, relatively high resource frequency is also a top indicator for semantic maturity.

If we consider the "maturity indicator" as the binary target concept, we obtain the patterns shown in Table 3. We observe similar influential properties as discussed above, however, the user and resource frequency combined with a medium or high closeness show the best performances.

Table 4. Top 5 patterns for the target concept "Synonym Indicator" (mean: 0.13)

#	lift	p	size	pattern
1	3.61	0.50	46	$ufreq > 13.0\%$
2	2.61	0.36	47	$bet \in [0.2\%, 1.0\%]$ AND $clos > 64.7\%$ AND $ufreq > 1.0\%$
3	2.53	0.35	77	$rfreq > 3.0\%$
4	2.40	0.33	51	$bet \in [0.2\%, 1.0\%]$ AND $clos > 64.7\%$
5	2.28	0.32	231	$deg > 6.0\%$ AND $ufreq > 1.0\%$

Looking at the "synonym indicator" results shown in Table 4, we observe, that the tag properties identified above have an even more pronounced influence, since the increase in the target concept (the lift) is between 2 and 3, indicating an increase in the mean target share of the synonym indicator in the subgroups by 100% to 200%. An example for a small subgroup containing only synonyms is described by the pattern: $bet \in [1326142, 1.0\%]$ AND $ufreq > 13.0\%$ consists of the tags "wallpaper", "templates" and "bookmarks".

5 Conclusion

In this paper, we have presented an approach for mining semantic maturity of tags in social bookmarking systems. We applied pattern mining for identifying subgroups of tags with mature semantic fingerprints according to different tag properties. These were based on structural and statistical folksonomy properties and computed using the tag co-occurrence information and tag/user frequency information. We provided a detailed analysis of the different properties, and presented a case study using data from del.icio.us. The results indicate the influence of several properties with interesting orders of magnitude for the del.icio.us dataset. For example, the number of users plays a crucial role for the process of semantic maturing; however, the addditional consideration of centrality properties can help to identify subsets of tags with a higher degree of maturity.

For future work, we plan to extend our proposed methodology to larger tag sets, including less frequently used tags and especially the notion of semantic "immaturity". Furthermore we plan to include further tag properties, also including temporal aspects like the amount of time a tag is present in the system. Additionally, we aim to evaluate the method on more datasets from diverse social systems.

Acknowledgements

This work has partially been supported by the VENUS research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University, and by the EU project EveryAware.

References

1. Angeletou, S.: Semantic Enrichment of Folksonomy Tagspaces. In: Int'l Semantic Web Conference. LNCS, vol. 5318, pp. 889–894. Springer (2008)

2. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05). pp. 647–652. Edinburgh, Scotland (2005)
3. Benz, D., Hotho, A., Stumme, G.: Semantics Made by You and Me: Self-emerging Ontologies can Capture the Diversity of Shared Knowledge. In: Proceedings of the 2nd Web Science Conference (WebSci10). Raleigh, NC, USA (2010)
4. Brandes, U., Pich, C.: Centrality Estimation in Large Networks. I. J. Bifurcation and Chaos 17(7), 2303–2318 (2007)
5. Cattuto, C.: Semiotic dynamics in online social communities. The European Physical Journal C - Particles and Fields 46, 33–37 (August 2006)
6. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: The Semantic Web, Proc.Intl. Semantic Web Conference 2008. vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
7. Fayyad, U.M., Irani, K.B.: Multi-interval Discretization of continuousvalued Attributes for Classification Learning. In: Thirteenth International Joint Conference on Articial Intelligence. vol. 2, pp. 1022–1027. Morgan Kaufmann Publishers (1993)
8. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press (1998)
9. Geng, L., Hamilton, H.J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys 38(3) (2006)
10. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging Systems. Journal of Information Sciences 32(2), 198–208 (April 2006)
11. Halpin, H., Robu, V., Shepherd, H.: The Complex Dynamics of Collaborative Tagging. In: Proc. of WWW2007. pp. 211–220. ACM, New York, NY, USA (2007)
12. Heymann, P., Garcia-Molina, H.: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Tech. rep., Computer Science Department, Standford University (April 2006)
13. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: The Semantic Web: Research and Applications. LNAI, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
14. Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag Recommendations in Folksonomies. In: Proc. PKDD 2007. Lecture Notes in Computer Science, vol. 4702, pp. 506–514. Berlin, Heidelberg (2007)
15. Körner, C., Benz, D., Strohmaier, M., Hotho, A., Stumme, G.: Stop Thinking, start Tagging - Tag Semantics emerge from Collaborative Verbosity. In: Proc. of WWW2010. ACM, Raleigh, NC, USA (apr 2010)
16. Liu, D., Hua, X.S., Yang, L., Wang, M., Zhang, H.J.: Tag Ranking. In: Proc. of WWW2009. pp. 351–360. WWW '09, ACM, New York, NY, USA (2009)
17. Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata (December 2004)
18. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Proc. Intl. Semantic Web Conf. LNCS, vol. 3729, pp. 522–536. Springer (2005)
19. Sigurbjörnsson, B., van Zwol, R.: Flickr Tag Recommendation Based on Collective Knowledge. In: Proc. of WWW2008. WWW '08, ACM (2008)
20. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge Univ Pr (1994)
21. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97). pp. 78–87. Springer Verlag, Berlin (1997)
22. Zhou, M., Bao, S., Wu, X., Yu, Y.: An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations. pp. 680–693 (2008)