An Approach to Building High-Quality Tag Hierarchies from Crowdsourced Taxonomic Tag Pairs

Fahad Almoqhim, David E. Millard, Nigel Shadbolt

Electronics and Computer Science, University of Southampton, Southampton, United Kingdom

{fibm1e09,dem,nrs}@ecs.soton.ac.uk

Abstract. Building taxonomies for web content is costly. An alternative is to allow users to create folksonomies, collective social classifications. However, folksonomies lack structure and their use for searching and browsing is limited. Current approaches for acquiring latent hierarchical structures from folksonomies have had limited success. We explore whether asking users for tag pairs, rather than individual tags, can increase the quality of derived tag hierarchies. We measure the usability cost, and in particular cognitive effort required to create tag pairs rather than individual tags. Our results show that when applied to tag pairs a hierarchy creation algorithm (Heymann-Benz) has superior performance than when applied to individual tags, and with little impact on usability. However, the resulting hierarchies lack richness, and could be seen as less expressive than those derived from individual tags. This indicates that expressivity, not usability, is the limiting factor for collective tagging approaches aimed at crowdsourcing taxonomies.

Keywords: Folksonomies, Taxonomies, Collective Intelligence, Social Information Processing, Social Metadata, Tag similarities.

1 Introduction

One of the essential principles behind the success of Web 2.0 applications is to harness the power of Collective Intelligence (CI) [1]. Collaborative tagging is one of the most successful examples of the power of CI for constructing and organizing knowledge in the Web. Tagging is a process that allows individuals to freely assign tags to a web object or resource, whereas folksonomy (a set of user, tag, resource triples) is the result of that process [2].

In recent years, folksonomies have emerged as an alternative to traditional classifications of organizing information [3]. However, they share the inconsistent structure problem that is inherited from uncontrolled vocabularies, which causes many problems like ambiguity, homonymy (same spelling but different meanings), and synonymy (terms have the same meaning) [4,5]. As a result, many researchers have focused on resolving this problem by proposing approaches for acquiring latent hierarchical structures from folksonomies and building tag hierarchies [6,7,8]. Building tag hierar-

chies from folksonomies can be useful in different tasks, like improving content retrieval [9], building lightweight ontologies [10] and enriching knowledge bases [11].

However, current approaches to automatic tag hierarchy construction come with limitations [12,13], such as suffering from the "generality-popularity" problem or the limited coverage of the existing knowledge resources. In this research, rather than propose a new algorithm for analyzing folksonomies we seek to explore whether a slight change in the tagging process itself could improve the resulting tag hierarchies. The new tagging approach takes the form of an "is-a" relationship, where users should type two related tags; i.e. Tag1 is a tag for the resource and Tag2 is a generalization of Tag1. The research hypothesis of this paper is that this simple relationship (Tag1 is-a Tag2) can be gained with low user interaction cost and provides higher quality tag hierarchies, compared to ones constructed from flat tags.

2 Related Work

Recently there have been several promising approaches proposed for building tag hierarchies from folksonomies. These approaches can be seen in two directions: First, **knowledge resources based approaches,** which aim to discover the meaning of tags and their relationships by using some knowledge resources, like WordNet and online ontologies. However, such resources are limited and they can only handle standard terms [12]. Second, **clustering techniques based approaches**. First pair-wise tag similarities are computed and then divided into groups based on these similarities. After that, pair-wise group similarities are computed and then merged as one until all tags are in the same group. For example, Heymann and Garcia-Molinay [6] propose an extensible algorithm that automatically builds tag hierarchies from folksonomies, extracted from Delicious and CiteULike. Their claim is that the tag with the highest centrality is the most general tag thus it should be merged with the hierarchy before others. Benz et al. [8] improved Heymann's algorithm by applying tag co-occurrence as the similarity measure and the degree centrality as the generality measure.

C. Schmitz et al [14] adopted the theory of association rule mining to analyze and structure folksonomies from Delicious. P. Schmitz [15] adapted the work of [16] to propose a subsumption-based model for constructing tag hierarchical relations from Flickr. Plangprasopchok et al. [7] adapted affinity propagation introduced by Frey & Dueck [17] to construct deeper and denser tag hierarchies from folksonomies. Yet Strohmaier et al. [3] showed that generality-based approaches of tag hierarchy, with degree centrality as generality measure and co-occurrence as similarity measure, e.g. [8] show a superior performance compared to probabilistic models, e.g. [7].

Although several approaches based on clustering techniques have been tried to structure folksonomies, they come with limitations [12,13]. These include the suffering from the "generality-popularity" problem. For example, Plangprasopchok and Lerman [18] found, on Flickr, that the number of photos tagged with "car" are ten times as many as that tagged with "automobile". By applying clustering techniques, "car" is likely to have higher centrality, and thus it will be more general than "auto-

mobile". Therefore, while tag statistics are an important source for constructing tag hierarchies, they are not enough evidence to discover concept hierarchies.

The experiment in this paper aims to explore whether a key reason for these limitations is that the current tagging approach, flat tags, does not provide a source of enough semantic evidence for building high-quality tag hierarchies. Rather we propose a slight change to the current tagging approach to benefit more from the power of CI by moving from collective folksonomies to collective taxonomic tag pairs.

3 Tag Hierarchies Leaning from Taxonomic Tag Pairs

In the proposed 'tag pairs' approach (**Fig. 1**), the user is required to tag the resource in the form of an "is-a" relationship, where Tag1 (the left box) is a tag for the resource and Tag2 (the right box) is a generalization of Tag1. For example, "Tower of London" is a "tower", or "tower" is a "building". The users can tag as much as they want for each resource in this way. Although this tag pairs approach shares some of the issues of single tags, such as spelling errors, it also provides additional semantics between tags. The algorithm we have adopted in our work (**Table 1**) is an extension of Benz's algorithm [8], which itself is an extension of Heymann's algorithm [6].

Table 1. Pseudo-code for the proposed algorithm

Input*: user-generated terms (tag pairs)* **,** ***Output****: tag hierarchy*

1. Filter the tag pairs {tag1, tag2} by an occurrence threshold *occ*.
2. Order the tag pairs in descending order by generality (measured by degree centrality in the tag2– tag2 co-occurrence network).
3. Starting from the most general tag2, as the root node, and append tag1 as a less general term underneath tag2.
4. add all tags tag2*i* subsequently to an evolving tree structure:
   (a) Calculate the similarities (using the co-occurrence weights as similarity measure) between the current tag tag2*i* and each tag currently present in the hierarchy and add the current tag tag2*i* as a child to its most similar tag *tag_sim*.
   (b) If tag2*i* is very general (determined by a generality threshold *min_gen*) or no sufficiently similar tag exists (determined by a similarity threshold *min_sim*), append tag2*i* underneath the root node of the hierarchy.
   (c) Append tag1*i* as a less general term underneath tag tag2*i*.
5. Apply a post-processing to the resulting hierarchy by re-inserting orphaned tags underneath the root node in order to create a balanced representation. The re-insertion is done based on step 4.

The algorithm is affected by various parameters, including: occurrence threshold *occ* (the number of tag occurrences); similarity threshold *min_sim* (the number of tag co-occurrences with another tag); and generality threshold *min_gen* (the number of tag co-occurrences with other tags). Empirical experiments were performed to optimize these parameters. By incorporating tag pairs this variation of Benz's algorithm reduc-

es the reliance on co-occurrence to create relationships, as nearly a half of the resulting tag hierarchy is created directly by users.

4 Empirical Study

Since we are proposing a new tagging approach, our experiment must look at two distinct aspects. Firstly its usability, in terms of efficiency, effectiveness and satisfaction, and secondly its performance in building high-quality tag hierarchies, in terms of semantics and structure. The technique will be successful if it increases the quality of tag hierarchy structure and semantics without significant impacting the ease of use.

To test the proposed tagging approach and collect data for executing the empirical study, we created the TagTree System. It is a web-based prototype which allows participants to tag some online resources by using the two tagging approaches (tag pairs and flat tags). The TagTree System consists of four main components:

- **User Interface**: The user interface describes to users how to use the tag pairs approach, with an example, and requires them to tag five resources with the tag pairs approach and another five with the flat tags approach (**Fig. 1**). To give a fair balance between the two tagging approaches, each of them is used first by half of the participants before they swap to the second approach.
- **Tag Content Recording**: This component records the tag content, including: user ID (user session), tags and time spent for each tagging action by the user.
- **Tags Normalising**: Before hierarchy construction, tags are passed to the normalisation process for clearing, e.g.: *Letters Lower-case* and *Non-English Deleting*.
- **Tag Hierarchy Constructing**: This component uses the proposed algorithm to construct tag hierarchies from the tags.

Fig. 1. The tag pairs (left) and the flat tags (right) tagging approaches

The Top 10 London Attractions[1], elected by visitlondon.com, were selected to be the resources used in the TagTree system for their popularity. The link of the TagTree system was sent to many people to take part in the study through email and social networks. After performing the normalization process, the dataset contains: 215 users, 333 tag pairs and 550 individual tags.

[1] http://www.visitlondon.com/things-to-do/sightseeing/tourist-attraction/top-ten-attractions

4.1 Evaluation Methodology

Evaluating taxonomy construction is a major challenge since there is not an approved evaluation dataset [3], nor an agreed methodology in the literature [19]. However, this subsection proposes a broad evaluation process to evaluate two things: 1) The quality of tag hierarchies constructed from our tag pairs approach, compared to tag hierarchies constructed from flat tags (Evaluation metrics: 1, 3 and 4). 2) The usability of the tag pairs approach compared to the flat tags approach, in terms of efficiency, effectiveness and satisfaction (Evaluation metric: 2). The proposed evaluation process consists of four evaluation metrics as follows:

Evaluation metric 1: Evaluation by Human Assessment (subjective). We chose a simple but effective approach, used by [3], for evaluating the consistency of our tag hierarchy relations. Each direct taxonomic pair (t1, t2) from the tag hierarchy is extracted and manually judged as a relation of either: "same as", "kind of/part of", "somehow related", "not related", or "unclear". The idea behind this approach is that a better tag hierarchy will have a higher percentage of pairs being judged as "kind of" or "part of", and a lower percentage of pairs being judged as "not related" or "unclear".

Evaluation metric 2: Usability Evaluation (subjective). We conducted an online survey based on the System Usability Scale (SUS) [20], a Likert scale questionnaire of 10 items that is a standardized tool and has been used and verified in many domains [21]. The survey yields a single score, from 0 to 100. Bangor et al. found that a product with SUS scores below 50 will mostly have usability difficulties, whereas scores between 70-89, though promising, do not assure high acceptance of usability [22].

Evaluation metric 3: Evaluation against Reference Taxonomy (objective). Two researchers, in the field of Semantic Web and Knowledge Engineering, were asked to create appropriate reference taxonomy of the experiment domain (**Fig. 2**). To perform the comparison between a produced taxonomy (PT) and reference taxonomy (RT), Dellschaft and Staab propose two measures: taxonomic precision (*tp*) and taxonomic recall (*tr*) for comparing concept hierarchies [23]. The main idea is to compare the positions of two common concepts (*c*) in both hierarchies (local measure), and then to compare the two whole hierarchies (global measure). The local measure of taxonomic precision (*tp*) and taxonomic recall (*tr*) are defined, respectively, as follows:

$$tp(c, PT, RT) = \frac{|ce(c,PT) \cap ce(c,RT)|}{|ce(c,PT)|} \quad (1) \qquad tr(c, PT, RT) = \frac{|ce(c,PT) \cap ce(c,RT)|}{|ce(c,RT)|} \quad (2)$$

Where (*ce*) is characteristic excerpts that contain the ancestors (super-concepts) and descendants (sub-concepts) of the concept which are present in both hierarchies. The global measure of taxonomic precision (TP) is defined, as follows:

$$TP(PT, RT) = \frac{1}{|Cp \cap Cr|} \sum_{c \in Cp \cap Cr} tp(c, PT, RT) \quad (3)$$

Where Cp is the concepts set in the produced taxonomy, and Cr is the concepts set of the reference taxonomy. To give an overall overview, taxonomic F-measure (TF) is computed as the harmonic mean of taxonomic precision and recall as follows:

$$TF(RT, PT) = \frac{2.\mathrm{TP(RT,PT)}.\mathrm{TR(RT,PT)}}{\mathrm{TP(RT,PT)}+\mathrm{TR(RT,PT)}} \quad (4)$$

Evaluation metric 4: Structural Evaluation (objective). It considers that a better tag hierarchy is a bushier and deeper hierarchy. To perform this evaluation, Plangprasopchok et al. [7] introduce a simple measure known as Area Under Tree (AUT). To compute AUT for a hierarchy, the distribution of nodes numbers in each level is computed first, and then the area under the distribution is calculated.

5 Results and Analysis

Two data sets were extracted from the TagTree system. The first one was collected by the tag pairs approach, while the second one was collected by the flat tags approach. In the experiment, three tag hierarchies (**Fig. 2**) are produced as follows: 1) **H1**: By using the tag pairs algorithm and the *tag pairs* dataset. 2) **H2**: By using the Benz's algorithm and the *flat tags* dataset. 3) **H3**: By using the Benz's algorithm and using the *tag pairs* dataset in which {tag1 is-a tag2} is considered as flat tags, i.e. ignoring the "is-a" relations.

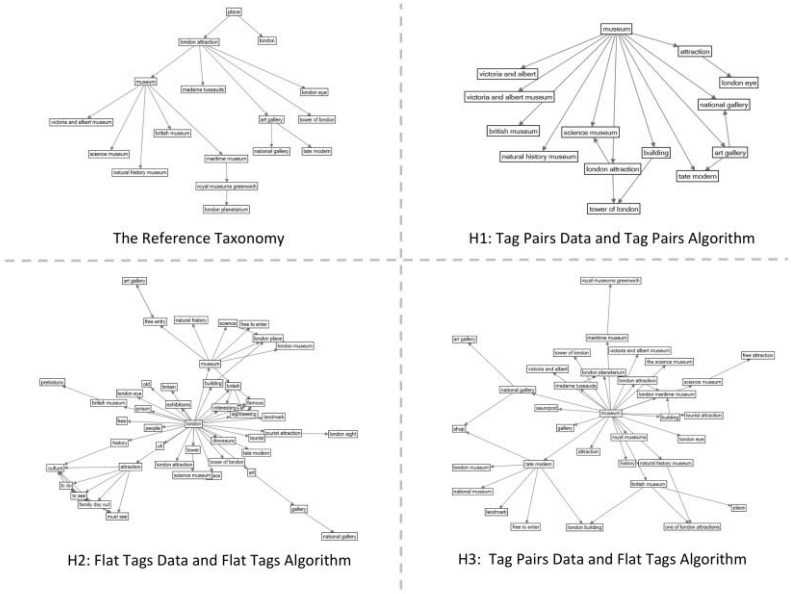

Fig. 2. The tag hierarchies used and produced in the experiment

5.1 Results of Semantic Evaluation

Fig. 3 shows the results of the semantic evaluation against the reference taxonomy, in terms of taxonomic precision (TP), taxonomic recall (TR) and taxonomic F-measure (TF). More similarity between a tag hierarchy and the reference taxonomy indicates that tag hierarchy has a higher quality.

The first observation that can be drawn from these empirical results is that there is a significant difference between H1 and H2. Our proposed extended algorithm yields tag taxonomies from our proposed tagging approach that is more similar to the reference taxonomy with taxonomic F-measure (TF) equal to 70.16%. Another important observation is that the quality of H3 is much better than the quality of H2, although both have been constructed by the same process (Benz's algorithm). However, H3 is built from tags originally collected from the tag pairs approach. This confirms our expectation and validates our research hypothesis that to make a small change to the current tagging approach can make a big change to the quality of the knowledge structure that can be built.

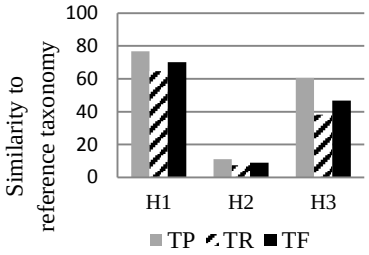

Fig. 3. Results of semantic evaluation against reference taxonomy

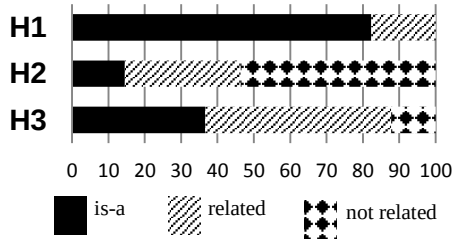

Fig. 4. Results of semantic evaluation by human assessment

Fig. 5. shows there is a large difference between the percentages of pairs being judged as "is-a" in H1 and others. Also, all the pairs in H1 are related. On the other hand, H2 is the worst since it has the lowest portion of "is-a" relation and the highest portion of "not related" relation between pairs. Furthermore, similar to the observation in Fig. 3, the quality of H3 is much better than the quality of H2.

5.2 Results of Structural Evaluation

Fig. 6 shows the results of AUT on the three produced tag hierarchies. H2 yields the highest AUT result, which indicates H2 is bushier and deeper than the others, whereas H1 yields the lowest AUT result. However, according to the previous results, H2 has a very small degree (TF=8.83%) of similarity to the reference taxonomy and also a big amount (53.62%) of "not related" tags pairs. This indicates that H2 has many noisy tags, while the proposed tagging approach and algorithm succeed in avoiding them. Ideally, it is a better to have an approach that generates both high quality and expressive tag hierarchies. While our approach succeeded in tackling the lack of consistent structure in folksonomies, it generated a less expressive hierarchy.

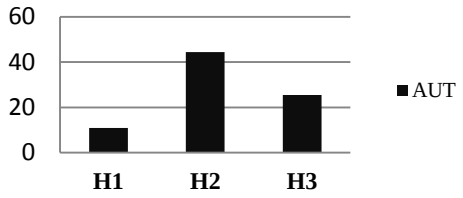

Fig. 6. Results of structural evaluation (AUT)

5.3 Results of Usability Evaluation

The results of this evaluation indicate that the average of SUS for the flat tags approach is 71.3%, with a standard deviation of 19.57, whereas the new approach obtains 54.6%, with a standard deviation of 16.22. First of all, this results show that the new approach is marginal acceptable since its average SUS score is over 50%. SUS scores are affected by the user experience by 15-16% between users who have "never" and "extensive" experience of the approach [24]. Consequently, the incipient SUS score of the tag pairs approach my get better over the time.

To measure the efficiency of the tag pairs approach compared to the flat tags approach, the time spent for each tagging action by users is recorded. The tagging action (ta) for the tag pairs approach means a pair of tags typed by the user, whereas for the flat tags approach means one tag or more typed by the user. The average time spent for using the tag pairs approach is 44.88 sec/ta, and 22.44 sec/tag. In contrast, the average time spent for using the flat tags approach is 71.90 sec/ta, and 36.37 sec/tag. This is a surprising result, as we expected the additional cognitive load of creating a tag pair to increase time taken, whereas our results show that users created a tag pair (containing two tags) in only slightly more time than it takes to generate a single tag.

6 Discussion and Conclusion

Current approaches to automatic tag hierarchy construction are limited, such as suffering from the "generality-popularity" problem or the limited coverage of the existing knowledge resources. Therefore, we proposed a slight change to the current tagging approach to cope with the lack of a consistent structure in folksonomies, and raise their semantic quality, whilst keeping the interaction cost of the process down. Our aim was to see if collecting tag pairs resulted in better quality hierarchy structure and semantics while minimizing the cost to usability.

The evaluation results of the produced tag hierarchies have shown that on the one hand the proposed tagging approach and algorithm have a superior performance in building high quality tag hierarchies when compared to ones built from the flat tags approach, but on the other hand they are not as rich, and therefore could be seen as less expressive. This problem might be caused by one or both of the following two reasons: First, the small size of the tagging resources and dataset in our experiment (we might expect to see a power law [25] in tag occurrence, and therefore the more

tags gathered the longer the tail of rare tags), and second, the inability of the tag pairs approach to capture the intermediate concepts between the high levels and leaves of the hierarchy. To solve the first problem, there is a need to run the experiment with a system that can motivate a larger number of participants for a longer time. And to solve the second problem, the approach itself need to be improved, e.g. adding other relations between tag pairs or asking users some specific questions to encourage them to provide intermediate concepts. However, this has the risk of making the approach complicated and losing the simplicity and flexibility features of folksonomies.

In terms of usability, the results have shown that the tag pairs approach is marginal acceptable. Users were even able to complete the task by using the tag pairs approach in quicker way compared to the flat tags approach. Although the tag pairs approach succeeded in avoiding noisy tags, it seemed to restrict users in their choice of tags, leading them to particular key taxonomic relations (this may a reason behind the surprisingly low average time taken to create tag pairs). This meant that the tags produced by the tag pairs approach had lower quantity and diversity than with flat tags.

Our results therefore indicate that expressivity, not usability, is the limiting factor for collective tagging approaches aimed at crowdsourcing taxonomies.

References

1. O'Reilly, T.: What is web 2.0: design patterns and business models for the next generation of software. [Online] Available at: http://oreilly.com/web2/archive/what-is-web-20.html (20 June 2013) (2005)
2. Vander Wal, T.: Folksonomy Coinage and Definition. (2007) Available at: http://vanderwal.net/folksonomy.html.
3. Strohmaier, M., Helic, D., Benz, D., Körner, C., Kern, R.: Evaluation of Folksonomy Induction Algorithms. ACM Transactions on Intelligent Systems and Technology 3(4) (2012) Article 74.
4. Golder, S., Huberman, B.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
5. Guy, M., Tonkin, E.: Tidying up tags. D-Lib Magazine 12(1) (January 2006) January, ISSN 1082-9873.
6. Heymann, P., Garcia-Molinay, H.: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. InfoLab Technical Report, Stanford (2006)
7. Plangprasopchok, A., Lerman, K., Getoor, L.: From saplings to a tree: Integrating structured metadata via relational affinity propagation. In : In Proceedings of the AAAI workshop on Statistical Relational AI, Menlo Park, CA, USA (2010)
8. Benz, D., Hotho, A., Stutzer, S.: Semantics made by you and me: Self-emerging ontologies cancapture the diversity of shared knowledge. In : 2nd Web Science Conference (WebSci10), Raleigh, NC, USA (2010)
9. Laniado, D., Eynard, D., Colombetti, M.: Using WordNet to turn a folksonomy into a hierarchy of concepts. In : 4th italian semantic web workshop: Semantic web application

and perspectives, Bari, Italy, pp.192-201 (2007)

10. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web 5(1), 5-15 (2007)
11. Zheng, H., Wu, X., Yu, Y.: Enriching WordNet with Folksonomies. In : 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'08), Osaka, Japan, pp.1075-1080 (2008)
12. Lin, H., Davis, J.: Computational and crowdsourcing methods for extracting ontological structure from folksonomy. In : 7th Extended Semantic Web Conference (ESWC'10), Heraklion, Greece, pp.472-477 (2010)
13. Solskinnsbakk, G., Gulla, J.: Mining tag similarity in folksonomies. In : 3rd international workshop on Search and mining user-generated contents (SMUC '11), Glasgow, Scotland, pp.53-60 (2011)
14. Schmitz, C., Hotho, A., Jäschke, R., Stumme, G.: Mining association rules in folksonomies. In : 10th IFCS Conference: Studies in Classification, Data Analysis and Knowledge Organization, Ljubljana, Slovenia, pp.261-270 (2006)
15. Schmitz, P.: Inducing ontology from flickr tags. In : Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland (2006)
16. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In : 22nd ACM Conference of the Special Interest Group in Information Retrieval, Berkeley, California, USA, pp.206-213 (1999)
17. Frey, B., Dueck, D.: Clustering by passing messages between data points. science 315(5814), 972-976 (2007)
18. Plangprasopchok, A., Lerman, K.: Constructing Folksonomies from User-Specified Relations on Flickr. In : 18th International World Wide Web conference, Madrid, Spain, pp.781-790 (2009)
19. Andrews, P., Pane, J.: Sense induction in folksonomies: a review. Artificial Intelligence Review, 1-28 (2012)
20. Brooke, J.: SUS-A quick and dirty usability scale. Usability evaluation in industry, 189-194 (1996)
21. Tullis, T., Stetson, J.: A comparison of questionnaires for assessing website usability. In : Usability Professional Association Conference, Minneapolis, USA, pp.1-12 (2004)
22. Bangor, A., Kortum, P., Miller, J.: An empirical evaluation of the system usability scale. International Journal of Human-Computer Interaction 24(6), 574-594 (2008)
23. Dellschaft, K., Staab, S.: On how to perform a gold standard based evaluation of ontology learning. In : The International Semantic Web Conference (ISWC2006), Athens, GA, USA, pp.228-241 (2006)
24. McLellan, S., Muddimer, A., Peres, S.: The Effect of Experience on System Usability Scale Ratings. Journal of Usability Studies 7(2), 56-67 (2012)
25. Halpin, H., Robu, V., Shepherd, H.: The Complex Dynamics of Collaborative Tagging. In : 6th International Conference on the World Wide Web, Banff, Canada, pp.211-220 (2007)