Improving Search in Folksonomies: A Task Based Comparison of WordNet and Ontologies

Sofia Angeletou
KMi, The Open University
Milton Keynes, UK
S.Angeletou@open.ac.uk

Marta Sabou
KMi, The Open University
Milton Keynes, UK
R.M.Sabou@open.ac.uk

Enrico Motta
KMi, The Open University
Milton Keynes, UK
E.Motta@open.ac.uk

ABSTRACT
Search in folksonomies is hampered by the fact that the meaning of tags and their relations are not made explicit in the system. This is typically addressed by using knowledge sources (KS) to semantically enrich tagspaces, most notably WordNet and (online) ontologies. However, there is no insight of how the different characteristics of these KS contribute to search improvement in folksonomies. In this work we compare these two KS in the context of folksonomy search. We show that while WordNet leads to richer tag structures than online ontologies do, its fine-grained sense hierarchy renders these structures less effective in search compared to the ones generated from ontologies.

Categories and Subject Descriptors
I.2.1 Applications and Expert Systems - *Natural language interface,* I.2.8 Problem Solving, Control Methods, and Search - *Graph and tree search strategies, Heuristic method.*

General Terms
Algorithms, Design, Experimentation, Performance.

Keywords
Folksonomies, Search, Enrichment, Ontologies, WordNet.

1. INTRODUCTION
Folksonomies are well-known Web2.0 applications and popular medium to publish, annotate and share content. The basic entities of folksonomies are the **users**, who annotate (tag) the **resources** with **tags** (freely selected text labels).

Due to the lack of restrictions in tagging, **synonymy**, **polysemy** and **basic level variation** are frequently observed phenomena among folksonomies tags [2]. Their common underlying cause is that the **intended meaning of tags and their relations are not made explicit** and therefore are not known to the system. Ideally, to overcome all the above impediments, each folksonomy should be associated with a semantic structure that provides explicit meaning and relations for the tags. One solution to this is the use of **formal knowledge sources (KS)** to acquire concepts and relations for tags. The two most popular KS used for this purpose are WordNet and online ontologies. However, currently there is no insight on which of them can provide the optimal solution to folksonomy enrichment and search.

In this work we investigate how the usage of each KS addresses the tag polysemy, synonymy, and basic level variation and minimises their effects on folksonomy search.

K-CAP'09, September 1–4, 2009, Redondo Beach, California, USA.
ACM 978-1-60558-658-8/09/09.

2. METHOD
To establish clear measures for the comparison of WordNet and ontologies we identify which characteristics of a KS affect the solution of the above folksonomy issues.

To overcome tag **polysemy** a requirement for a KS is to have enough senses to match the possible senses of tags. Thus, we measure the number of senses per tag assigned when using each KS. To address tag **synonymy** a KS needs to provide enough synonym words for the senses. To explore the search problems caused by the **basic level variation** we inspect the structure of senses in each KS. Particularly, the number of more specific (subsense) and more generic (supersense) senses per sense and the additional resources retrieved per sub/super-sense. In order to obtain an overall tagspace coverage against a KS we measure the number of tags mapped to senses from each KS.

In a nutshell, in a tagspace with |R| resources, |T| tags and two KS, WordNet (W) and Ontologies (O), a tag T has a number of senses $S_{(T:KS)}$ in the respective KS. Each of them, has a number of subsenses $SUB(S_{(T:KS)})$ and supersenses $SUP(S_{(T:KS)})$. We are interested in measuring:

- $|T_{KS}|$, the number of tags covered by each KS.
- $\overline{|S_{(T:KS)}|}$, the mean number of senses per tag.
- $\overline{|SUB(S_{(T:KS)})|}$ and $\overline{|SUP(S_{(T:KS)})|}$, the mean number of subsenses and supersenses per sense.
- $Inc_{SUB(S_{(T:KS)})} = |R_{SUB(S_{(T:KS)})} - R_T| / |R_{SUB(S_{(T:KS)})} \cup R_T|$ and $Inc_{SUP(S_{(T:KS)})} = |R_{SUP(S_{(T:KS)})} - R_T| / |R_{SUP(S_{(T:KS)})} \cup R_T|$

are the ratios of resources tagged with the sub/super-senses of the query keyword T. These represent the increase in recall as they would normally be excluded from the results when using a keyword based search like folksonomies do.

To measure the above, we first enrich the tagspaces utilising the two KS. For each KS, we use one enrichment strategy and acquire one semantic structure. Second, we test each semantic structure on a search experiment utilising a query expansion mechanism.

Enrichment Strategy A uses WordNet to map tags to its noun synsets and each synset is considered a sense. For each sense we import all the synonym words which comprise its lexical information. To create a structure between senses, we import its ancestor synsets (hypernyms) till the root of the WordNet hierarchy as well as its first level of hyponyms. This strategy leads to semantic structure A.

Enrichment Strategy B exploits the online ontologies indexed in Watson[1]. This process of sense selection is less straightforward as there is no pre-defined ontological sense repository. All ontology concepts that contain the tag in their local name (id) or label(s) are selected. Because we explore multiple ontologies, the same concept may be defined in more than one thus leading to duplicates. To avoid the redundancies we use a clustering algorithm ([1]) that groups sufficiently similar entities together and merges them into a cluster of entities, which we consider one sense. The lexical information of the sense is comprised by the localnames and the labels of the entities. After the entity clustering, we iterate on the superclasses and subclasses of these entities to acquire the desirable sense hierarchy which leads to semantic structure B.

Once the structures A and B are created we implement a **Query Mechanism** to perform query expansion during a search task. Initially we map the query keyword to the relevant sense from each structure A and B. Then we acquire all the sub/super-senses of the sense and from them we extract their lexical information. Expanding the query with the lexical information, we search for resources tagged with the initial keyword query as well as the additional keywords obtained from the expansion.

3. Experiments

To evaluate the usefulness of WordNet and ontologies in folksonomy search we applied enrichment strategies A and B on a dataset from Flickr. This was a randomly selected sample of 12233 photos with 89446 tags and 13645 individual tags from the group Plant [directory]. We implemented the query mechanism in a web application and asked 11 users to ask both semantic structures A and B queries related to plants. The quantitative results from the enrichment and the querying are displayed in Table 1.

Quantitatively, in terms of enrichment, WordNet nouns provide higher coverage of tags, more senses and sub/super-senses for each tag compared to classes from online ontologies. In our search experiment, WordNet performed better in the subsense expansion, while ontologies performed better in supersense expansion despite the higher coverage of tags to WordNet.

[1]http://watson.kmi.open.ac.uk

Table 1: Quantitative results for two KS

KS	WordNet	Ontologies
$\lvert T_{KS}\rvert$	29%	15%
$\overline{\lvert S_{(T:KS)} \rvert}$	2.4	1.7
$\overline{\lvert SUB(S_{(T:W)}) \rvert}$	6.5	2.3
$\overline{\lvert SUP(S_{(T:W)}) \rvert}$	2.5	1.7
$\mathbf{Inc}_{SUB}$	25%	18%
$\mathbf{Inc}_{SUP}$	33%	39%

4. DISCUSSION

We comparatively investigated the effect of WordNet and ontologies on folksonomy enrichment and search and WordNet outperformed ontologies in most measures. Yet, in search ontologies performed equally well overall.

More tags of the dataset were mapped in WordNet, with more senses and more sub/super-senses per tag. In addition, the matching of tags to WordNet was more straightforward, due to the existing sense hierarchy and the universal naming patterns. On the other hand, for the matching of tags to ontological senses all the aforementioned issues need to be dealt with individually. The heterogeneity in modelling and naming, the redundant senses, the variable hierarchical granularity and the modelling errors impede the matching of tags to ontological entities.

But in a search scenario WordNet's hierarchy of senses resulted to a fine-grained structure which matched less well with the tagset compared to the variably granular and implicitly redundant structure acquired from ontologies. As a result of this, the query expansion through the WordNet structure returned relatively less results considering the significantly larger number of mapped tags to WordNet compared to ontologies.

ACKNOWLEDGMENTS

This work was funded by the NeOn project sponsored under EC grant number IST-FF6-027595.

REFERENCES

[1] Angeletou, S., Sabou, M., and Motta, E. *Semantically enriching folksonomies with FLOR*. In ESWC Workshop: CISWeb, Tenerife, Spain, 2008.

[2] Golder, S., and Huberman, B.*Usage Patterns of Collaborative Tagging Systems*. Journal of Information Science, 32:198-208, 2006.