Position Paper: Ontology Learning from Folksonomies

Dominik Benz, Andreas Hotho
Knowledge & Data Engineering Group (KDE), University of Kassel,
Wilhelmshöher Allee 73, 34121 Kassel, Germany
http://www.kde.cs.uni-kassel.de

Abstract

The emergence of collaborative tagging systems with their underlying flat and uncontrolled resource organization paradigm has led to a large number of research activities focussing on a formal description and analysis of the resulting "folksonomies". An interesting outcome is that the characteristic qualities of these systems seem to be inverse to more traditional knowledge structuring approaches like taxonomies or ontologies: The latter provide rich and precise semantics, but suffer - amongst others - from a knowledge acquisition bottleneck. An important step towards exploiting the possible synergies by bridging the gap between both paradigms is the automatic extraction of relations between tags in a folksonomy. This position paper presents preliminary results of ongoing work to induce hierarchical relationships among tags by analyzing the aggregated data of collaborative tagging systems as a basis for an ontology learning procedure.

1 Introduction

A fundamental aspect of knowledge management is often the establishment of structure within a set of information resources, e.g., PDF documents, bookmarks or photographs. Most traditional approaches address this issue by decomposing the domain under consideration into interrelated classes or categories, which are intended to model exhaustively the underlying knowledge structure. Each available information resource is then assigned to one or more classes. Ontologies are a well-known formalism for this purpose. The hierarchical topic category structure of, e.g., a web directory like the Open Directory Project[1] can be seen as an example of a taxonomy, which constitute a core component of ontologies [Staab and Studer, 2004]. Their widespread use is however hindered by the expertise and cost required for their creation and maintenance.

Collaborative tagging systems feature another structuring paradigm: Each user can assign one or more arbitrary keywords (or *tags*) to each of his resources, facilitating a flat "by-keyword" access to personal or public resources. The resulting structure of users, tags and resources became known as *folksonomies* [Mathes, 2004]; refer to [Hotho *et al.*, 2006] for a formal definition. Due to their inherent simplicity and immediate usefulness, these systems are able to overcome the previously described knowledge acquisition bottleneck. However, this comes at the cost of a lack of precision (see [Golder and Huberman, 2006]), which is exactly the strength of ontological approaches.

As a first step towards unleashing synergies by automatically learning ontologies from folksonomies, this position paper proposes an algorithm to induce hierarchical relationships among tags. The algorithm has been tested with real-world user data from the social music sharing platform *Last.fm*[2], and the outcome has been evaluated against a gold-standard music style hierarchy taken from the comprehensive online music directory *MusicMoz*[3].

2 Inducing Hierarchical Relations among Tags

The goal of this work is to automatically induce a concept hierarchy, i.e., a tree structure, whose nodes (representing concepts) each consist of one or more tags from a folksonomy. Concept specificity increases with increasing depth in the tree, and there exists only a single type of relation, whose semantics resembles closely the one of the taxonomic relation [Bozsak *et al.*, 2002].

Data foundation The most often used information source is based on two types of so-called *tag-tag-cooccurrence networks*, which can be extracted from a folksonomy. Each existing tag corresponds to a node, and there exists a undirected edge with weight w_{ij} between two tags t_i and t_j if

- there were w_{ij} users who have used both t_i and t_j to annotate any of their resources (*user-based tag-tag-cooccurence, UTC*)
- there were w_{ij} resources both annotated with t_i and t_j by any user (*resource-based tag-tag-cooccurence, RTC*)

Classes of approaches Existing approaches based on tag cooccurrence information can be assigned to one of the following three classes:

- *Social Network Analysis:* [Mika, 2005] pioneered in applying centrality and other measures like the clustering coefficient coming from social network analysis to the UTC and RTC networks in order to identify broader and narrower terms. [Heymann and Garcia-Molina, 2006] proposed betweeness centrality as tag generality measure. The latter approach will serve as a basis for the proposed algorithm.

[1] http://www.dmoz.org

[2] http://www.last.fm

[3] http://www.musicmoz.org

- *Statistical approaches:* The work of [Schmitz, 2006] and [Schmitz *et al.*, 2006] is based on statistical models of tag subsumption, the latter is corroborated with the theory of association rule mining.
- *Clustering approaches:* Starting from a similarity measure between tags, clustering approaches like [Begelman *et al.*, 2006] identify groups of highly related tags. Depending on the chosen clustering algorithm, a hierarchical relationship between the tag clusters is established.

Proposed Algorithm The proposed algorithm is an extension of the work of [Heymann and Garcia-Molina, 2006]. It comprises the following steps:

1. Filter the tags by an occurrence threshold τ_{occ}
2. Order the tags in descending order by generality (measured by degree centrality [Hoser *et al.*, 2006] in the UTC network)
3. Starting from the most general tag, add all tags t_i subsequently to an evolving tree structure:
   - identify the most similar existing tag t_{sim} (using the weights w_{ij} in the UTC network as similarity measure)
   - decide whether t_{sim} and t_i are synonyms or form a compound expression (using an adapted statistical model of subsumption from [Schmitz, 2006] based on the RTC network)
   - if yes $\rightarrow$ merge t_{sim} and t_i, otherwise append t_i as a less general term underneath t_{sim}.

Compared to the original algorithm, the first extensions consists of applying a computationally much less complex centrality measure (namely degree centrality) as tag generality measure. The original measure is based on betweenness centrality, whose computation requires $O(nm + n^2 \log n)$ time [Brandes, 2001], whereby n is the number of tags and m is the number of edges in the weighted cooccurrence network. This dimension becomes problematic when applied to real-world large scale folksonomy systems. As a further extension, tag synonymy and compound expressions (e.g., *"open"* and *"source"*) are considered.

3 Assessing the Quality of Learned Relations

Choosing a gold-standard based evaluation paradigm, it is a non-trivial task to judge the similarity between a learned concept hierarchy and a reference hierarchy, especially regarding the absence of well-established and universally accepted evaluation measures. As a detailed description of the similarity measures used is beyond the scope of this paper, the reader is referred to [Dellschaft and Staab, 2006] for an overview. Two of the described measures, namely taxonomic precision / recall / F_1-measure and the OntoRand-Index were adapted to compare two hierarchies on an instance-based level: The underlying idea is that two concept hierarchies are very similar if they structure the resources in question in a similar manner.

4 Preliminary Experimental Results

In order to validate the proposed algorithm, experiments were conducted with a dataset crawled from the social music sharing website *Last.fm*[4]. It consists of 978 resources

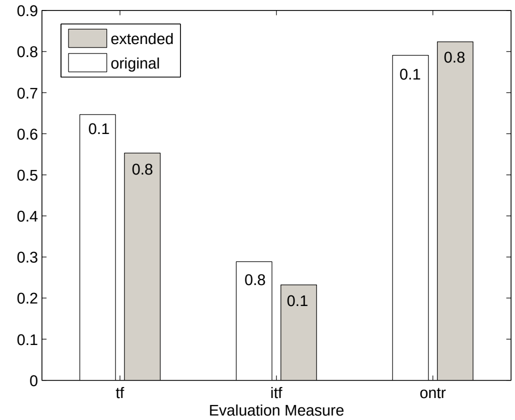

Figure 1: Experimental Results: Comparison of the performance of the proposed algorithm with the original version. The numbers in the bars correspond the optimal parameters for each algorithm as found in the first test phase.

(i.e., music artists), 3585 users and 7283 tags, connected by 162406 tag assignments. As a gold standard, a music style hierarchy (built by volunteer music fans) consisting of 548 styles was downloaded from *MusicMoz*[5]. Each artist from the Last.fm dataset was assigned to 1-3 MusicMoz style categories.

The experimental setup consisted of two phases:

1. parameter optimization for both the original and the proposed algorithm
2. comparison of the performance of both algorithms with the obtained optimal parameters, compared by the taxonomic F_1-measure (tf), the instance-based taxonomic F_1-measure (itf) as well as the extended OntoRand-Index $ontr$.

Figure 1 displays the results. For none of the given measures, there is a clear winner. An important issue when interpreting the differing assessments of the measures is their respective basis: The taxonomic F_1-measure (tf) compares two hierarchies based on matching concept names, while the instance-based taxonomic F_1-measure (itf) and the extended OntoRand-index are based on the assignment of information resources to each concept. It is obvious that the two latter measures are strongly influenced by the chosen assignment strategy.

Considering the fact that the proposed algorithm is computationally much less complex (see Section 2) compared to its original version, the results are acceptable. To get a better impression of the capabilities of the proposed algorithm, Figure 2 illustrates its outcome. Following paths from the hierarchy root towards the leafs, the styles become more and more specific. Starting from the *ROOT* node in the center of the image, one nice example is the path *rock* $\rightarrow$ *metal* $\rightarrow$ *death metal* $\rightarrow$ *progressive death metal* towards the lower left corner.

5 Conclusions and Further Work

This paper presented preliminary results of ongoing work on inducing hierarchical relationships among tags in a folksonomy as basis for an ontology learning procedure. Experiments with real-world data suggest that the proposed

[4]http://www.last.fm

[5]http://www.musicmoz.org

algorithm is able to produce a consistent hierarchical category scheme, which comes close to a handcrafted scheme. An open issue for future research is how to assess the quality of the gold-standard the outcome of the learning procedure is compared with. A deeper theoretical understanding of the interaction of the algorithm's building blocks (i.e., tag generality measure, tag similarity measure and tag subsumption measure) is needed in order to further improve the results. Another aspect that needs consideration is how the resources of the folksonomy are assigned to the resulting hierarchical structure.

References

[Begelman *et al.*, 2006] Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland, May 2006.

[Bozsak *et al.*, 2002] E. Bozsak, M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz, and V. Zacharias. Kaon - towards a large scale semantic web. In K. Bauknecht, A. Min Tjoa, and G. Quirchmayr, editors, *E-Commerce and Web Technologies, Third International Conference, EC-Web 2002, Proceedings*, volume 2455 of *LNCS*, pages 304–313, Berlin, 2002. Springer.

[Brandes, 2001] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[Dellschaft and Staab, 2006] Klaas Dellschaft and Steffen Staab. On how to perform a gold standard based evaluation of ontology learning. In *Proceedings of ISWC-2006 International Semantic Web Conference*, Athens, GA, USA, November 2006. Springer, LNCS.

[Golder and Huberman, 2006] Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Journal of Information Sciences*, 32(2):198–208, April 2006.

[Heymann and Garcia-Molina, 2006] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, Standford University, April 2006.

[Hoser *et al.*, 2006] Bettina Hoser, Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Semantic network analysis of ontologies. In *European Semantic Web Conference, Budva, Montenegro*, June 2006.

[Hotho *et al.*, 2006] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, June 2006. Springer.

[Mathes, 2004] Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata, December 2004.

[Mika, 2005] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, *The Semantic Web - ISWC 2005, Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer, 2005.

[Schmitz *et al.*, 2006] Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. iberna, editors, *Data Science and Classification. Proceedings of the 10th IFCS Conf.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270, Heidelberg, July 2006. Springer.

[Schmitz, 2006] Patrick Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland, May 2006.

[Staab and Studer, 2004] Steffen Staab and Rudi Studer, editors. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.

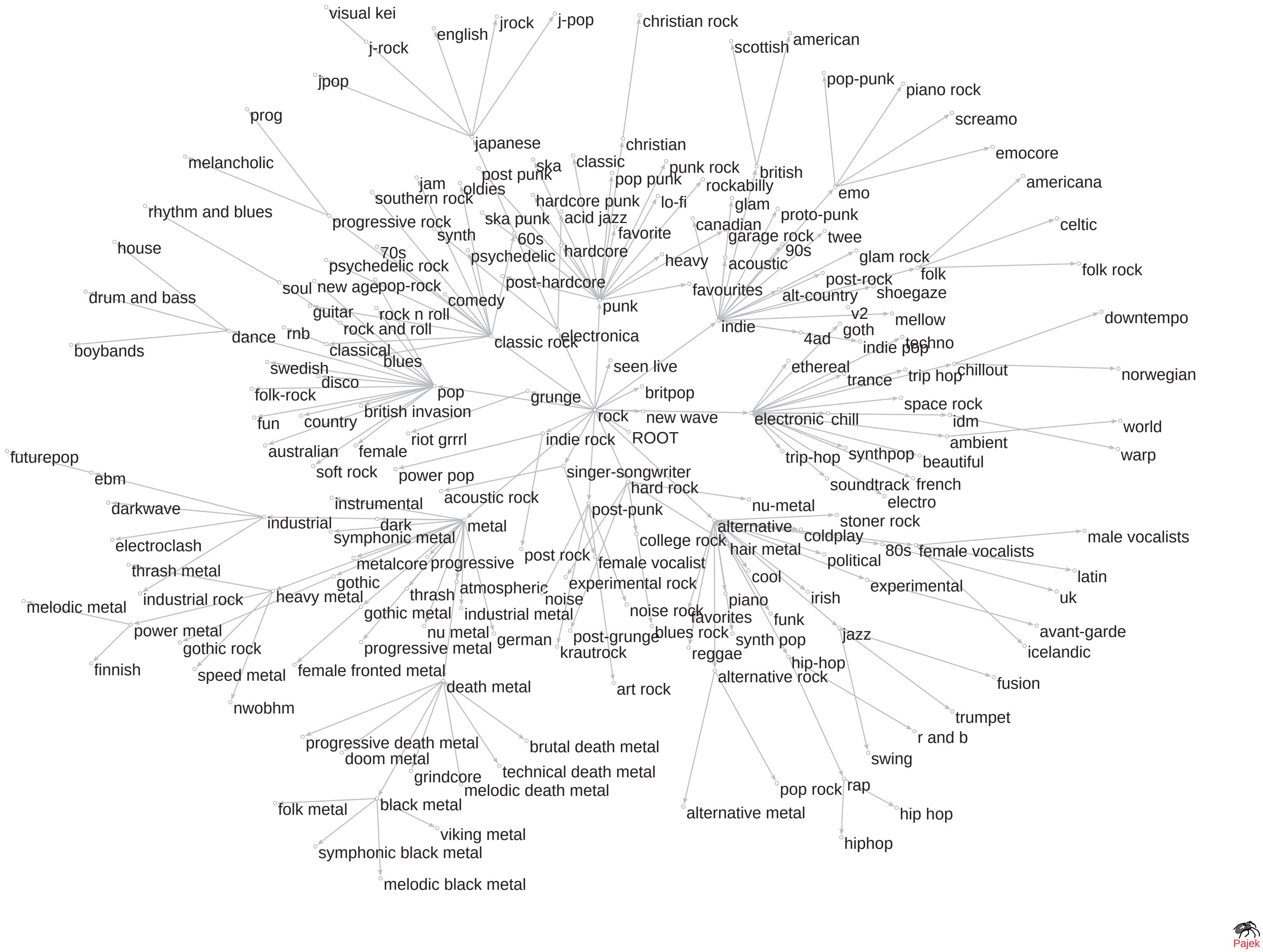

Figure 2: Music style hierarchy extracted from Last.fm dataset by the proposed algorithm.