Logsonomy — A Search Engine Folksonomy

Robert Jäschke ⋆ ‡ and **Beate Krause** ⋆ and **Andreas Hotho** ⋆ and **Gerd Stumme** ⋆ ‡
⋆ Knowledge & Data Engineering Group, University of Kassel, Wilhelmshöher Allee 73, 34121 Kassel, Germany
‡ Research Center L3S, Appelstr. 9a, 30167 Hannover, Germany

Abstract

In social bookmarking systems users describe bookmarks by keywords called tags. The structure behind these social systems, called *folksonomies*, can be viewed as a tripartite hypergraph of user, tag and resource nodes. This underlying network shows specific structural properties that explain its growth and the possibility of serendipitous exploration.

Search engines filter the vast information of the web. Queries describe a user's information need. In response to the displayed results of the search engine, users click on the links of the result page as they expect the answer to be of relevance. The clickdata can be represented as a folksonomy in which queries are descriptions of clicked URLs. This poster analyzes the topological characteristics of the resulting tripartite hypergraph of queries, users and bookmarks of two query logs and compares it two a snapshot of the folksonomy del.icio.us.

Introduction

Folksonomies are complex systems consisting of user-defined labels added to web content such as bookmarks, videos or photographs by different users. In contrast to classical search engines a folksonomy can be explored in different dimensions taking users, tags and resources into account. While search engines automatically crawl the web, the content of a folksonomy is determined by its users. As a consequence, the content selection and retrieval in folksonomies is a social process, in which users decide about relevance.

Search engines use relevance feedback by extracting user's click histories from log files and computing personalized rankings. However, many web searchers are not only interested in a ranked list of search results, but also in exploring community content.

We will discuss the realization of such search communities by building an anonymized folksonomy similar to a social bookmarking system from search engine logdata. As logdata contain queries, clicks and session IDs, the classical dimensions of a folksonomy can be reflected: Queries or query terms represent tags, session IDs correspond to users, and the URLs clicked by users can be considered as the resources that they tagged with the query terms. Search engine users can then browse this data along the well known folksonomy dimensions of tags, users, and resources. A similar point of view was represented in (Baeza-Yates & Tiberi 2007). We will call the resulting structure *logsonomy*.

Table 1: Datasets

dataset	$\lvert T\rvert$	$\lvert U\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$
del.icio.us host only URLs	430,526	81,992	934,575	14,730,683
del.icio.us complete URLs	430,526	81,992	2,913,354	16,217,222
AOL complete queries	4,811,436	519,250	1,620,034	14,427,759
AOL split queries	1,074,640	519,203	1,619,871	34,500,590
MSN complete queries	3,545,310	5,680,615	1,861,010	10,880,140
MSN split queries	902,210	5,679,240	1,860,728	24,204,125

Logsonomies open a wide field of exploration. What kind of semantics can we extract from them? Is a serendipitous discovery of information also possible? How do logsonomies differ from folksonomies? In this poster, we address these questions by analyzing the topological properties of two logsonomy datasets created from MSN and AOL clickdata sets and comparing our findings to the social bookmarking system del.icio.us. Table 1 gives more details. Each clickdata set is represented twice: in the first, queries remain queries, in the second, queries are split into single terms.

Topological Properties

In previous work (Cattuto *et al.* 2007) it was shown that folksonomies exhibit specific network characteristics. These characteristics help to explain why people are fascinated from this structure: *Short path lengths* lead to short ways between users, resources and tags, which allows for finding interesting resources by browsing the system randomly. *High clustering coefficients* show dense neighbourhoods which are tracked by the formation of communities around different topics. Finally, *cooccurrence graphs* show the building of user enabled shared semantics. We consider these network measures, adapted to the tripartite structure of our data.

Degree distribution The degree of a node in a tripartite graph reflects the number of hyperedges, which contain the specific node. It has been shown, that the distribution of the degree of nodes for tags and resources in a folksonomy follows a power law distribution (Hotho *et al.* 2006). For the distribution of resources and tags this is also the case in logsonomies. Thereby the split queries datasets show a more similar distribution to del.icio.us than the ones which

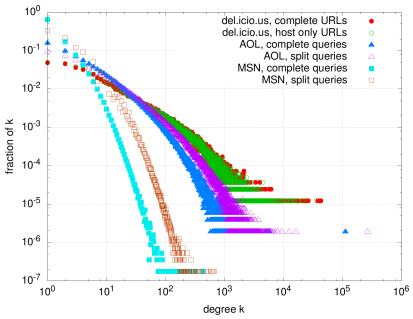

Figure 1: Degree distribution of users

contain full queries. We attribute this difference to the fact that full queries have less overlap among users.

Figure 1 shows the distributions of users for the different datasets. Neither in the logsonomies nor in the folksonomy a power law distribution is reflected. While the curve of the AOL users shows a progression similar to the one of del.icio.us, the curve for the MSN users exhibits a steeper gradient. This is probably due to the nature of sessions representing the users in this dataset: sessions have a shorter life time and are more topic specific as opposed to unique, timeless user IDs as they exist in the AOL dataset. The probability of being strongly interlinked is therefore lower.

Connected components In both folksonomies and logsonomies a giant connected component exists, which comprehends most of the existing nodes. For instance, in del.icio.us with host only URLs the size of the GCC is 1,446,888. As the dataset contains in total 1,447,093 nodes, the GCC covers 99.99% of the whole hypergraph. In the AOL split query dataset the relation is similiar with 3,220,395 vs. 3,229,100 total nodes.

Small-world properties It has been shown in (Cattuto *et al.* 2007), that folksonomies exhibit small world characteristics considering average shortest path lengths and clustering coefficients. The average shortest path length denotes the mean distance between any two nodes in the graph. Because of complexity reasons, we have approximated the average path length by randomly selecting 4000 nodes and calculating the average path length of each of those nodes to all other nodes in its connected component. Compared to del.icio.us, all four datasets from MSN and AOL provide larger path lengths. Capturing the intuition of serendipitous browsing, it takes longer to reach other queries, users, or URLs within a logsonomy than it takes to jump between tags, users and resources in a folksonomy. In particular, the high values for MSN are likely to result from the fact that a user cannot bridge between different topics if he searched for them in different sessions. However, the path lengths still indicate the graph's small world properties: Comparing each logsonomy to its corresponding random graphs, the path lengths do not differ considerably. (For instance, AOL split queries: 3.62; corresponding binomial random graph 3.90).

The clustering coefficient characterizes the density of connections in the environment of a node. It describes the cliquishness, (i. e., *are neighbor nodes of a node also connected among each other*) and the connectedness of a node, (i. e., *would they stay acquainted if the node was removed*). Our results show that the cliquishness and connectedness coefficients of the original graphs are in general higher than the ones of the corresponding random graphs. This indicates that there is some systematic aspect in the search behaviour which is destroyed in the randomized versions.

Result I The analysis of the topological structure of logsonomies has shown, that the clicking behaviour of search engine users and the tagging behaviour of social bookmarking users is driven by similar dynamics: in both systems, power law and small world properties exist. Hence, logsonomies can serve as a source of finding topic-oriented, community driven content either by a specific search along the three dimensions or by means of serendipitous browsing.

Strength in the tag-tag-co-occurrence graph A first approach to studying the semantics in a logsonomy is the analysis of the tag-tag-cooccurrence graph. This graph consists of tags which are linked if they share the same user and resource. A weight for edges is introduced by counting in how many user-resource combinations these two tags appear together. The strength s_t of a tag t denotes the sum of the weight of its edges. Finally, the nearest neighbour connectivity of a tag t_i, denoted as $S_{nn}(t)$, is the sum of the strengths of each tag t_j connected to t_i, averaged by the total number of links of the tag t_i. For each tag t one can set $S_{nn}(t)$ in relation to its own strength.

Result II The strength distributions of the split versions of the logsonomies are similar to the del.icio.us dataset: $S_{nn}(t)$ of tags with low strength varies strongly, while for tags with higher strength the variation is much smaller and shows a dissassortative behaviour. The strength distributions of the complete queries differs strongly in shape and size. We assume that this structure stems from frequency effects rather than from inherent semantics as shown for tagging systems (Cattuto *et al.* 2007). In future work, we want to digg deeper into the semantics of logsonomies to investigate its application to query expansion and ranking methods.

References

Baeza-Yates, R., and Tiberi, A. 2007. Extracting semantic relations from query logs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 76–85. New York, NY, USA: ACM.

Cattuto, C.; Schmitz, C.; Baldassarri, A.; Servedio, V. D. P.; Loreto, V.; Hotho, A.; Grahl, M.; and Stumme, G. 2007. Network properties of folksonomies. *AI Communications Special Issue on "Network Analysis in Natural Sciences and Engineering"*.

Hotho, A.; Jäschke, R.; Schmitz, C.; and Stumme, G. 2006. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, volume 4011 of *LNCS*, 411–426. Budva, Montenegro: Springer.