Aspects of Broad Folksonomies

Mathias Lux
Klagenfurt University
Universitätsstrasse 65-67
9020 Klagenfurt, Carinthia, Austria
mlux@itec.uni-klu.ac.at

Michael Granitzer, Roman Kern
Know-Center
Inffeldgasse 21
8010 Graz, Styria, Austria
{mgrani, rkern}@know-center.at

Abstract

Folksonomies, collaboratively created sets of metadata, are becoming more and more important for organising information and knowledge of communites in the Web. While for a single user the difference to keyword assignment is marginal, the power of folksonomies emerges from the collaborative aspects. Folksonomies are already issue of research. Within this publication we analyse underlying statistical properties of broad folksonomies aiming to identify laws and characteristics, which allow inferring properties for folksonomy based retrieval. The actual benefit of folksonomies for retrieval and the derived methods are concluded from experiments with aggregated data from del.icio.us[1].

1 Introduction

Text retrieval on the internet has a quite short but rather successful history compared to the overall history of retrieval (see [8]). Since more than 10 years the internet offers ways to publish and store (hyper-)text in a distributed way and incrases the problem of retrieving the right piece of content at the right time. First approaches were based on textual search by indexing documents with searchable terms. Weighting schemes offered an improvement over Boolean retrieval methods. But in the World Wide Web (WWW) exploitation of domain characteristic offered far more possibilities: Link analysis approaches like PageRank have shown their show their superiority to approaches based solely on weighting schemes. Manual annotation of pages, as it is done for instance by *Yahoo!* or within the *Open Directory Project*, is another approach. These directories provide high quality yellow pages for prominent topics on the WWW, but do not cope to the size of the web.

With the rise of social software another type of directory emerged in form of *social bookmarking* applications (see e.g. [7]). Within social bookmarking systems, like for instance del.icio.us or Simpy[2], users bookmark and annotate web resources. The annotation abilities supported commonly among social bookmarking systems are simple: Users assign keywords, called tags, which are not restricted to a vocabulary, to a bookmark and optionally give a short description and title of the resource.

The sheer mass of users, bookmarks and tags leads to a vast and unorganized amount of data. However running social bookmarking systems have shown, that even from uncontrolled use of the system various patterns emerge. For instance the interconnection of resources, users and tags, which is called *folksonomy* (derived from *folk* and *taxonomy*), shows that tags often re-occur together based on the semantic interconnection of the tags (e.g. *webdesign* and *css* are assigned to the same resource more often than *webdesign* and *piano*).

This contribution analyses folksonomies and derives statistical properties of the interconnection of resources, users and tags in a sample of social bookmarking data. Based on the findings retrieval based on a folksonomy emerged from social bookmarking is discussed and conclusions for possible information retrieval methods are drawn.

2 Related Work

The actual phenomenon of social software has many different aspects. One important aspect therefore is what potential motivations for the success of social software are. In [3] several hypotheses and direct benefits like self organization, the notion of participation and a low participation barrier, are presented.

The emergent structure of folksonomies, which are also often called *collaborative tagging systems*, has been subject to several publications. In this context one has to distinguish between *narrow* and *broad* folksonomies. While in a nar-

[1] URI: http://del.icio.us

[2] URI: http://www.simpy.com

row folksonomy (e.g. Flickr[3]) only the owner of a ressource can tag it, in a broad folksonomy (e.g. del.icio.us) anyone can tag anything. In the following the focus is put on broad folksonomies, especially on social bookmarking systems.

Several research groups have investigated the growth, breadth and complexity of folksonomies in general. In [4] the amount and growth of the del.icio.us folksonomy is studied. Also different intentions for tagging tags are identified. They also show that while there are users, who use lots of different tags for social bookmarking, there are also users, who stick to a small set of tags. In [6] the tag quality is discussed. Language used in Flickr and del.icio.us for tags is investigated and approaches for increasing quality (regarding ambiguity, synonyms and different languages) are proposed.

The analysis of folksonomies based on a general model is discussed in [9]: The general model of folksonomies is outlined as ontology, which incorporates the notion of the users. The author of [9] also investigates co-occurrence analysis based on association matrices. In [1] characteristics of co-occurrence are investigated. Major finding based on a dataset from del.icio.us is that the power law applies for frequency and rank of co-occurring tags. Based on this finding a time based model for the creation of folksonomies is proposed, which is described in detail in [2].

In [11] a probabilistic model for the creation of folksonomies in accordance to the latent semantic indexing model is proposed. probabilities Frthermore probability of co-occurrence of tags, users and resources are derived from this model. The proposed model differs from the one of [1] in so far, as in this model the conceptual dependency between users, tags and resources is integrated in a more general way.

Our brief survey shows that that the potential of folksonomies has already been pointed out by several research groups. However currently there is no generally agreed model for folksonomies available and although the emergence of superimposed structure has been already discussed, specific hypothesis have not been supported through analysis of big datasets. The most promising approach however has been published in [2], but empirical evidence was based only on a very small sample of del.icio.us, which focused on a set of three tags and the tags co-occurring with this three tags. The potential of folksonomies for retrieval has also been discussed, but the possible benefit of folksonomies has not yet been evaluated.

3 Methodology

The power law is defined by $y = \alpha \cdot x^{\beta}$ with coefficient α and exponent β. The power law is very prominent as it emerges in many systems relying on human interaction. Examples are the richness of people, the size of cities and the popularity of movies. An interpretation is that there are few very powerful samples, like very rich people, big cities or very popular movies (usually referred to as the head), while the majority of samples is rather weak, e.g. poor people, small cities or minor (usually referred to as tail).

As in folksonomies human interaction is heavily involved and other papers also indicate the emergence of power law distributions there, we focused on finding evidence for the existence of power law distributions in folksonomies. Fitting a sample to a power law is a highly discussed topic and several approaches exist. Beyond estimating the linear and exponential coefficients α and β the more interesting question is, whether the data fits the estimated coefficients or if the data does not obey a power law at all. Several approaches have been discussed in statistics for estimating this *goodness-of-fit*, as for example in [5]. In our experiment we used a χ^2 (chi-square) test to estimate the goodness of fit criterion and linear least squares (LLS) estimates (see [10]) for fitting the power law. We apply the following test procedure for analysing power law behaviour:

1. Obtain data samples for a tag t as $D_t = \{f(d_1), \ldots, f(d_n)\}$ using co-occurrence analysis [4] where $f(d_i)$ is the frequency of the i-th element.

2. Order the data samples D_t by frequency.

3. Estimate power law parameters using LLS of $\log y = \log \alpha + \beta \log r(x)$ where y is the frequency of event x and $r(x)$ is the rank of event x in the data sample (i.e. the $r(x)$ most often occurring tag).

4. Calculate the χ^2 statistic for each rank and estimate the significance that the data samples are generated by the estimated power law. [5]

4 Emergent Power Law Distributions in Folksonomies

As outlined in section 2, most folksonomy models assume a power law distribution among terms, similar as Zipf ([12]) has stated according to word distributions in classical text retrieval systems. Therefore we analysed whether this assumption holds for the whole folksonomy. Furthermore we analysed simple statistical properties in order to derive heuristics for tags, which may be used further in retrieval algorithms.

[3] URI: http://www.flickr.com

[4] see section 4 for details on obtaining data samples

[5] Note that one rank has to contain at least 5 samples as prequisite for the χ^2 test. To satisfy this, the tail was summed up into one single rank according to [5])

For the experiments a sample of the del.icio.us folksonomy was acquired. Through continuous aggregation of recent bookmarks that have been bookmarked by at least 2 people[6] every 7th minute we obtained a sample of 3,234,956 bookmarks having 9,241,878 tag associations of 356,838 different tags by 84,121 different users. The full size sample was used in section 4.2. A subset of this sample, which was generated by taking the results of aggregation up to a certain date, was used in section 4.1 and in section 5. The subset consisted of 838,804 bookmarks having 2,408,935 tag associations of 135,473 different tags by 26,919 different users.

4.1 Tag Co-Occurrence

In a first step the statistical characteristics of tag co-occurrence were investigated. Focusing on a tag $\bar{t}$, which is assigned to a set of resources $R_{\bar{t}}$, all other tags assigned to resources $r \in R_{\bar{t}}$ are considered as co-occurring tags, whereas the frequency of the co-occurrence depends on how often a tag has been assigned by users. The more often a tag is co-assigned (co-occurs with $\bar{t}$), the higher is its frequency. This follows the line of analysis as stated in [1].

Our main interest in analysing co-occurrence lies in the proposed Yule-Simon model with time dependent memory of Cattuto [1], which is based on the assumption that the co-occurrence of tags is distributed by a power law. While in [1] this is shown for a few, highly frequent tags, the questions arises whether this assumption holds for a whole folksonomy. Our analysis shows, that for around 80% of the tags of a folksonomy the co-occurring tags follow a power law distribution, which approves Cattuto's assumption. We found that for around 90% of the estimated power law exponent $\beta \in [-1.5, -0.5]$, which shows that for most tags co-occurrence follows a model with similar parameters.

4.2 Resource and User based Tagging Characteristics

A second analysis approach was to estimate (i) the *resource statistic* as user frequency for a given tag over the assigned resources (investigating users distribution for resources tagged by a specific tag) and (ii) the *user statistic* as resource frequency for a given tag over the users having assigned the tag (investigating resource distribution for users having assigned a specific tag). The data sample has been partitioned into two groups:

1. D_1 containing tags with a cumulative user (resource) frequency larger than 30 with $\|D_1\| = 15,835$ and
2. D_2 containing tags with cumulative user (resource) frequency lower than 30 with $\|D_2\| = 341,000$

[6]URI: http://del.icio.us/recent?min=2

Power Law Distributions: The power law analysis has been restricted to D_1 to focus on more frequently used tags. For the resource statistic, resources are ranked by the frequency of users tagging the resource with a tag $\bar{t}$; for the user statistics users are ranked by frequency of resources tagged with a tag $\bar{t}$.

The resource statistic for D_1 set showed, that around 18.4% of the ranked resources follow a power law distribution with statistical significance of 99%. The values of β are mostly found in the interval $[-0.5, -0.1]$ with some outliers ranging to a minimum exponent of -2.28. Regarding the user statistic, around 13% are following a power law distribution with a significance level of 99%. Exponent values β are mostly found in the interval $[-0.5, -0.1]$ with outliers to a minimum of -2.58. The characteristics of the user statistics are similar to the characteristics of the resource statistic.

We did not find any correlation between the power law distributions in the user and in the resource statistic by taking into account their statistical characteristics only. However a more detailed look on tags, which are distributed by power law [7] in both statistics, which are around 5.6% of D_1, revealed that all tags are "meaningful" descriptive tags like for example *RFC*, *Technorati*, *X86* etc. The list contained also most of the high frequency tags like *web2.0*, *webdesign* etc.; no misspellings or unpopular tags were found. We argue that those tags, which follow a power law w.r.t. users *and* resources are high quality tags (i.e. tags describing ressources with high accuracy) for most of the users involved in the investigated social bookmarking system.

Following the interpretation of the power law, those 18.4% of the tags, which follow a power law distribution in the resource statistic, are assigned by a lot of users to few resources (head of the distribution) and to a lot of different resources by a few users (tail of the distribution). So the question arises, whether there is a disagreement on the assignment of tags between users on the tail or if the tagging results from (semantically) different sub community of users. Therefore we compared users tagging high rank, mid rank and low rank resources, were each rank level takes on 33% of the area under the power law curve. As a result we found that only a small fraction of tags have overlapping user groups, which points towards sub communities (user groups sharing the same link selection and tagging behaviour) in the tail of the power law distribution. Furthermore we can assume that the model of preferential attachments applies on high ranked resources (aka rich get richer).

Similar findings were made by analysing the power law distributions of the users statistic. Here a power law distribution indicates that a few users are tagging a lot of resources and that a lot of users are tagging few resources

[7]With a significance of 95%

with one tag. While this holds true for 13% of the tags, the question arises whether the same resources are tagged by different users which would hint towards a notion of favourite user tags. Again we calculated the overlap between resources used by high ranked, mid-ranked and low ranked users. Only a low fraction of resources overlap between head and tail, retaining the hypothesis of personal favourite tags since if a tag is assigned by a high rank user, it is assigned often to different resources. On the other hand, the tail indicates that users assign not only tags out of their favourite vocabulary. There maybe two reasons for this: (i) Users see a tag – for instance currently highlighted in a tag cloud – and use it one or two times or (ii) users assign tags which describe topics apart from their major interest (the tags are therefore not used commonly by the user).

Analysing Tags not following a Power Law. Since not all tags follow a power law distribution w.r.t to user and resource statistic, we can not analyse them based on their distribution. Therefore, we investigated some basic statistical properties of these tags. Interesting findings for data set D_1 and D_2 can be summarized as follows:

1. **Unique Assignements**: Around 57% of the tags in D_2 are used only once, that is by only one user for only one resource. Optimistically these tags can be seen as shortcuts for a user to a resource or pessimistically as misspellings of tags. In either case those tags are useless from a retrieval point of view.

2. **Personal Vocabulary**: In D_1 3.93% and in D_2 19% of the tags where used by only one user but assigned to more than one ressources. Those tags may be understood as personal vocabulary. Thus, useful for personal retrieval but useless for the rest of the community.

3. **Unpopular Vocabulary**: Around 12% of the tags in D_1 resp. 38.7% of the tags in D_2 are assigned to different resources by different users but only one time. These tags can be seen as unpopular vocabulary; tags only used by a small fraction of users not very often.

From a retrieval point of view, one can conclude that a large fraction of the folksonomy contains tags specific to single users or sub communities. In context of the overall success of folksonomies one can infer that one major success criterion of folksonomies is the possibility to use an implicitly defined personal or sub-community vocabulary embedded in a larger community context. It can be assumed that standard retrieval methods like inverted indices and ranking schemes like TF*IDF may be used upon the power law distributed tags efficiently, since word distribution in text also follows a power law [12]. Furthermore, to include those sub communities into retrieval, relationships between users and resources have to be taken into account using for example query exansion on co-occurring terms. What remains open is the influence of this large fraction of tags not distributed by power law and how to include the different usages of tags outlined above for retrieving information.

5 Retrieval Specific Aspects

Besides the analysis of tag co-occurrence and distributions, the retrieval performance of tags compared to the retrieval performance of title and descriptions has also been investigated. Besides the user id in our sample a bookmark contains the creation date, a number of tags, a title, automatically copied from the web page title on creation of the bookmark, and a description. The question therefore is, if tags are able to add information further to description and title for retrieval purposes.

We started our analysis by assuming that resources annotated with a tag $\bar{t}$ define the ground truth, i.e. those resources which should be found if a user searches for a term $\bar{t}$. Titles and description of resources have been indexed using the open source search engine Lucene[8] employing the included standard TF*IDF weighting scheme without adoption to the scenario. We did not take misspellings of tags or special wordings into account, but restricted the number of tags to the 6000 tags occurring most often in resources.

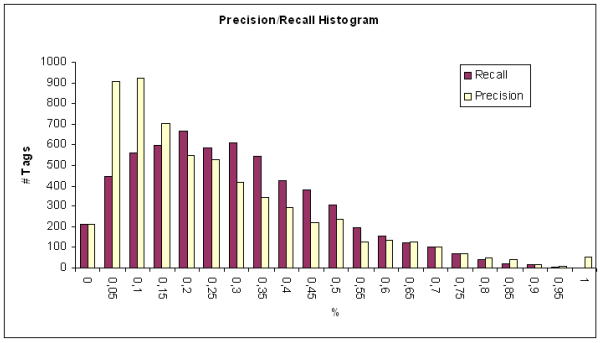

Figure 1. Histogram of Precision & Recall

For each tag $\bar{t}$ we searched the indexed title of resources and compared the results to the resources tagged with $\bar{t}$ by computing precision, recall and the F_1 measure. Figure 1 shows the histogram of the resulting Precision and Recall values [9] for queries on title and descriptions of resources.

As it can be seen, retrieval performance is mostly grouped around precision/recall values below 0.5 and that queries tend to have a higher recall and a lower precision.

[8] http://lucene.apache.org/

[9] Bars for F_1 values are not given due to keep the figure simple

Therefore the results from searching terms in title and description do not match the tagged resources. Thus, for retrieval tags can be seen as additional source of information, extending description and title as well as adding more precise information.

Furthermore we investigated the question if descriptions and tags are used for describing similar semantics. Overall, from all resources only 20% had a description assigned. In a second test run following the above described test the description was included into the full text search additionally to the title. Analysing precision, recall and F_1 measures showed that there is no significant difference between the first test (title only) and the second test (title & description). Looking more into detail we found, that for around half of the tags, which were assigned to resources in combination with a description, there is a significant decrease in precision and a corresponding increase in the recall if descriptions are included. For the other half precision and recall stays the same. Thus, 50% of the available descriptions contain information similar to the information described by tags, whereas the remaining 50% can be seen as orthogonal information.

As conclusion of this section, tags add additional information to titles and descriptions of resources and thus can be considered as added value for retrieving resources. It seems that tagging and descriptions of resources are orthogonal to each other and thus sublement each other in a retrieval process.

6. Conclusion

Based on the investigations presented above we can assume that within a folksonomy structures emerge from unorganized human input. Our tests have shown that for a large fraction of tags in our sample (around 80%) the cooccurring tags are power law distributed, which supports the findings in [1] for a bigger folksonomy ($\geq 350,000$ tags compared to three tags in [1]).

While tags in general add information for retrieval (see section 5), a huge amount of tags seems to be inappropriate for retrieving resources or users. These tags can be categorized in misspellings, unpopular tags, shortcuts on resources or personal vocabularies. Especially the last option is an interesting one: There are strong indications for vocabularies which are only used by a small group of people and only on a few resources. This isolates those sub-communities from the overall community and creates semantic islands similar to the islands in the bow tie structure of the WWW, which complicates retrieval of resources of these sub groups or contributions to these sub groups for other sub communities.

We also have shown that descriptions added to resources can be divided in two equal groups: (i) Those containing similar information as tags of the folksonomy and (ii) those containing information orthogonal to the tags.

Most important fact is that all the findings lead to the understanding of emergent structures in folksonomies. All above described conclusions lead to the assumption that classical text retrieval methods like latent semantic analysis or TF*IDF can be adapted to folksonomy based retrieval in a meaningful way.[10]

References

[1] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences United States of America*, 104:1461, 2007.

[2] C. Cattuto, V. Loreto, and V. D. Servedio. A yule-simon process with memory. *Europhysics Letters*, 76(2):208–214, 2006.

[3] G. W. Furnas, C. Fake, L. von Ahn, J. Schachter, S. Golder, K. Fox, M. Davis, C. Marlow, and M. Naaman. Why do tagging systems work? In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pages 36–39, New York, NY, USA, 2006. ACM Press.

[4] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32:2, April 2006.

[5] M. L. Goldstein, S. A. Morris, and G. G. Yen. Fitting to the power-law distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 41(2):255–258, 2004.

[6] M. Guy and E. Tonkin. Folksonomies - tidying up tags? *D-Lib Magazine*, 12(1), January 2006. ISSN 1082-9873.

[7] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i) - a general review. *D-Lib Magazine*, 11(4), April 2005.

[8] A. N. Langville and C. D. Meyer. *Google's Pagerank and Beyond: The Science of Search Engine Rankings. The Science of Search Engine Rankings*. University Presses of CA, July 2006.

[9] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference*, LNCS, pages 522–536. Springer, 2005.

[10] E. W. Weisstein. Least squares fitting–power law. From MathWorld–A Wolfram Web Resource., 2007.

[11] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM Press.

[12] G. K. Zipf. Human behaviour and the principle of least-effort. *Addison-Wesley, Cambridge MA*, 1949.

[10] **Acknowledgements**: The project results have been partially developed in the AVALON (Acquisition and Validation of Ontologies) project, financed by the Austrian Research Promotion Agency (http://www.ffg.at) within the strategic objective FIT-IT under the project contract number 810803 (http://kmi.tugraz.at/avalon). The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.ffg.at), and by the State of Styria.