Quality Metrics for Tags of Broad Folksonomies

Céline Van Damme
(MOSI, Vrije Universiteit Brussel
celine.van.damme@vub.ac.be)

Martin Hepp
(E-Business and Web Science Research Group, Bundeswehr University
mhepp@computer.org)

Tanguy Coenen
(STARLab, Vrije Universiteit Brussel
tanguy.coenen@vub.ac.be)

Abstract: Folksonomies do not restrict its users to use a set of keywords preselected by a group of experts for interpersonal information retrieval. We assume that the quality of tag-based information retrieval and tag suggestions when tagging a resource can be ameliorated if we are able to automatically detect the tags that have an intersubjective meaning or tags that are understood and used by many members of a group. In this paper, (1) we suggest three possible tag quality measures for broad folksonomies (2) by means of an analysis of a del.icio.us dataset, we provide preliminary evidence that the suggested metrics return useful sets of intersubjectively valid tags and (3) as an additional evaluation, we asked individuals to judge the tag sets obtained through the metrics.
Key Words: broad folksonomies, tag quality, metrics, del.icio.us
Category: H.3.5, J.4.

1 Introduction

Folksonomies involve their user community into the creation process of categories by inclusion of their tags. The absence of a controlled vocabulary, allows the community members to produce any category or tag that enters their mind. Since users are not restricted to a controlled vocabulary, the question arises concerning the quality of the tags and the information retrieval capacities of folksonomies. [Golden and Huberman 2006] showed that users primarily tag for a personal purpose. Tags used for a private purpose can in some cases be useful for the whole community, e.g. many people could say that they have *toread* a certain book when annotating a book at Librarything. However an annotated picture with the tag *ourdog* does not imply that it is a picture of everyone's dog.

We assume that the quality of tag-based information retrieval and tag suggestions when annotating new objects can be improved if we are able to auto-

matically judge the quality of a tag in terms of the intersubjective comprehension it engenders. We define intersubjective comprehension as the degree that a tag is understood by many members of a group.

In order to create metrics for an automatic tag quality judgement, we have to distinguish between the two kinds of folksonomies: broad and narrow folksonomies as classified by [Vander Wal 2005]. The difference between these types lies in the number of people that tag the object. In the case of broad folksonomies, a resource is tagged by many people (e.g. web pages) whereas in the case of narrow folksonomies there are only a few persons involved, in most situations only the author or creator of the resource (e.g. pictures on Flickr).

1.1 Related Work

Besides [Guy and Tonkin 2006][Sen et al. 2007] [Lee et al. 2007], research on quality of tags is scarce. In [Guy and Tonkin 2006] the authors focus on how they can help the user providing good and consistent tags. They suggest giving a sort of tag education to the user to improve the quality of the tags, for instance train the user to use singular terms. The authors in [Sen et al. 2007] as well as in [Lee et al. 2007] propose to extend tags with a kind of rating mechanism. In [Lee et al. 2007] the user has to tag a resource as well as add a positive (e.g people like) or negative context (e.g. people do not like) to each tag (e.g. people do not like war). Positive and negative contexts are respectively indicated with a plus and minus sign. Different tag rating scenarios are tested and discussed in [Sen et al. 2007]. The authors conclude with a number of design guidelines for the creators of websites that contain a tagging mechanism.

Still, asking individuals to rate tags as a quality measure is time-consuming. An interesting alternative would be to automatically detect intersubjective tags and regard intersubjectivity as an indicator of quality.

1.2 Contribution and Overview

In this paper, (1) we suggest three possible tag quality measures for broad folksonomies,(2) by means of an analysis of a del.icio.us dataset, we provide preliminary evidence that our suggested metrics return useful intersubjective tag sets and (3) as an additional evaluation, we asked individuals to judge the tag sets obtained by applying the metrics.

The structure of this paper is as follows: in section 2, we give an overview of the metrics. We elaborate on the implementation issues of the metrics applied to a del.icio.us dataset in section 3. In section 4, we discuss the evaluation of the metrics. We give some limitations of the research in section 5 and end with a conclusion and discussion of future research in the last section.

2 Metrics

In this section, we present three metrics to automatically select the intersubjective tags from a set of tags used to annotated a web resource by the principles of broad folksonomies. For each metric, we give a description, explain how it can be calculated and motivate why we propose it.

2.1 Metric 1: High Frequency Tags

2.1.1 Description

We select tags with the highest 5 frequencies.

2.1.2 Calculation

For each tagged resource, we order the tags and count the frequency of each distinct tag. We then choose the tags with the highest 5 frequencies.

2.1.3 Motivation

Because many people express their thoughts about a particular resource through their selection of tags, we propose to analyze high frequency tags.

2.2 Metric 2: Tag Agreement

2.2.1 Description

We define tag agreement for resource x as the tags that are selected by more than 50% of the users who have tagged resource x.

2.2.2 Calculation

We first determine the frequency of each unique tag. Then, we calculate the number of users that have tagged each resource. The tag agreement is consequently calculated by dividing the tag frequency by the number of users that have tagged a resource, and multiply the result by 100 in order to have a percentage as a result. When all the users agree on a certain tag, this number should be equal to 100%. The closer to 0%, the less the users agree on that particular tag.

2.2.3 Motivation

Decisions in various areas of human acitvity are often taken on the basis of a majority: more than 50% of the people have to agree on a certain proposal in order to get it accepted. Tagging in the case of broad folksonomy can be seen as a way of voting for the semantic labeling of a resource and this is why we suggest tag agreement as a second metric.

2.3 Metric 3: TF-IRF

2.3.1 Description

For each tag we calculate its Tag Frequency Inverse Resource Frequency or TF-IRF weight and select tags with the highest 5 TF-IRF scores. We derived the TF-IRF metric from Term Frequency Inverse Document Frequency or TF-IDF, a common metric in the domain of automatic indexing for finding good descriptive keywords for a document. When selecting the appropriate tags for a certain document, the TF-IDF formula takes the intra as well as inter document frequency of keywords into account. The higher the TF-IDF weight, the more valuable the keyword.

2.3.2 Calculation

Corpus: To calculate the TF-IRF weights, we need a corpus of similar resources. This can be obtained through tag clustering. By calculating the co-occurrences of tag pairs and transforming the pairs (as nodes) and their co-occurrences (as weighted edges) into a graph, we can apply the Markov Clustering (MCL) Algorithm[Van Dongen 2000]. Results in [Van Dongen 2000] show that the MCL algorithm is very good and highly performant for clustering graphs. Therefore, we choose this algorithm to build the corpus.

Calculating TF-IRF: In order to transform TF-IDF into TF-IRF, we have to make some adjustments to the formula. We have to exclude the textual information or documents from our formula since tagged resources are not always textual (e.g. an mp3 audio file). The only data we can analyse are tags. As a consequence, we suggest the equation below to calculate the TF-IRF weight for a certain tag annotated to a resource. The formula is based on TF-IDF (with $t_{x,y}$ = frequency of tag_x for $resource_y$, T_y = total number of tags for $resource_y$, corpus = sum of resources and R_x= sum of resources that have tag_x).

$$TF - IRF(tag_{x,y}) = \frac{t_{x,y}}{T_y} * log(\frac{|corpus|}{R_x})$$

2.3.3 Motivation

In the domain of automatic indexing a lot has been written on how to select the most appropriate keywords. Research on automatic indexing dates back to the 1950's and consequently represents a large body of knowledge. We believe it is interesting to apply TF-IDF, which is one of the common techniques in this area, to broad folksonomies.

3 Data set

For the analysis we used the Del.icio.us dataset from [Laniado et al. 2007]. It contains more than 3.400.000 unique bookmarks from around 30.000 users retrieved in March 2007.

3.1 Preparing Steps

In order to be able to compare the results from the different metrics, we had to calculate each metric on the same collection of bookmarks. Since the TF-IRF metric required the creation of a corpus or a set of related resources, in this case bookmarks, we started the analysis by making the corpus.

3.1.1 Cleansing

Before we could apply the MCL algorithm to build the corpus, we had to do some data cleaning. We

- Removed all the English stop words from the tag set, since most of the high frequency tags of the dataset were in English.
- Stemmed the remaining tags by removing the end suffix. Words that have the same stem or root are considered to be referring to the same concepts (e.g. running and run have the same stem, i.e. run).
- Merged duplicate tags since a duplication of tags appeared after stemming.
- Disregarded all bookmarks that are tagged by less than 100 users. Because we only want to include bookmarks that are evaluated by a large number of users. We reduced the number of bookmarks in the collection to 3898.
- Calculated the co-occurrence of the tag pairs.

URL	Metrics[1]
http://www.imdb.com	M1: movie film refer database entertainment M2: movie M3: database cinema movie film refer
http://www.ifilm.com	M1: video movie film entertainment fun M2: video M3: video trailer film ifilm movie
http://www.apple.com/trailers/	M1: movie trailer entertainment apple film M2: movie M3: trailer apple quicktime importediefavorites movie

Table 1: Tag sets obtained by applying the metrics

3.1.2 Applying MCL Algorithm

We applied the MCL algorithm on all tag pairs and their corresponding frequencies obtained in previous step. We excluded the tag pairs with a frequency of less than 100. This means both tags have been used less than 100 times together to tag a particular resource. A lower threshold value did not result in clearly distinguishable clusters. We opted for the cluster which contained the following tags: entertainment, film and movie since these tags are common used terms. We decided to include a bookmark in the corpus if it had at least one tag with a frequency of 10 that belonged to this cluster. We opted for a number of 10 since we wanted to be sure that a link with the cluster existed. As a result, we obtained 127 bookmarks for this cluster.

3.2 Results

We calculated the tag sets for each metric and bookmark in the cluster. Some examples of the results are included in table 1. For each metric, we ordered the tags from the left to the right based on decreasing values. We noticed a close linkage between the tag sets obtained by the first and third metric. In some cases, the high frequency tags and TF-IRF metrics only differ by the order of the tags. In the other cases, there is a close overlap between metric 1 and 3 because they often share similar tags.

When applying the tag agreement metric on the dataset we received, we noticed that the average number of tags per bookmark where agreement exists was very low. The minimum and maximum values lay between 0 and 3, and the

[1] M1 = High Frequency Tags; M2 = Tag Agreement; M3 = TF-IRF

modus and median both had a value of 1. It was therefore not possible to select 5 tags for the tag agreement metric since there were on average 0.94 tags per bookmark that correspond to the definition. There were even 26 bookmarks that did not have any tags confirming to this pattern. After excluding these 26 bookmarks, the mean increased just slightly to 1.18. There was a very weak negative correlation between the number of tags retrieved by the tag agreement metric and the number of users that have tagged the object (ρ = -0.17). This means that an increase in the umber of users that tag a certain resource will slightly decrease the number of tags that comply with the tag agreement metric.

4 Preliminary Evaluation

To answer the question which metric is generating the best results, we decided to set up an online survey. To conduct the survey we created a tool in PHP that chooses a bookmark as well as its tag sets randomly from the MySQL database. In each session 10 bookmarks had to be evaluated. There were 101 bookmarks in the database, because we excluded the 26 bookmarks that did not have any tags for the tag agreement metric.

Since our cluster of bookmarks was selected based on the requirement of sharing one of the tags (entertainment, film and movie) with a frequency of 10, we asked an international group of 20 students to participate in the online survey. We asked them to select the tag set of which they thought it did the best job at describing a specific bookmark. In case of doubt, we told them to take the order of the tags into account. Indeed, a tag placed at the beginning, was more important than one which is located more to the right.

First, we gave the students a one hour presentation on tagging and folksonomies. We also introduced them to Del.icio.us and gave a brief demonstration of the system. Then, we invited them to a computer room to participate in the survey.

Due to randomness, 75 of the 101 bookmarks were evaluated and some of them were assessed several times. In total, 173 times a bookmark was evaluated. We did not obtain the logical number of 200 (20 students doing 10 evaluations), since (1) some of the websites were down during the survey and had to be removed from the result list and (2) not all the students pursued the survey until the end. On average, the students opted in 52.6% of the cases (n=91) for the high frequency tags metric and in 41% of the cases (n=71) for the TF-IRF metric. The tag agreement scored poorly: in 6.3% (n=11) they selected this option. A possible explanation for this might be the low number of tags.

5 Limitations of the Research

Although we obtained first preliminary results, there are certain limitations that apply to the online survey. We did not ask the students why did they opted for a certain tag set. The students were not asked whether the chosen tag set contained all tags, too many tags or not enough tags and the number of participants was too low.

6 Discussion and Conclusion

In this paper, we proposed three metrics to automatically detect the intersubjective tags for a resource tagged in the context of a broad folksonomy. We applied the three metrics to a Del.icio.us dataset and through an online survey we tried to find the most appropriate metric. Preliminary results show that the High Frequency Tag metric generates the best results. However, the TF-IRF metric also produces valuable results.

In the near future, we want to set up a large scale online survey to find out if the results suggested in this small-scal setup can be reproduced in a larger scale survey. Further, we want to find out what the characteristics are of the tags that comply with the metric. For instance, how many of the tags are general or specific.(3) In a next step, we plan to extent this research to the case of narrow folksonomies.

References

[Guy and Tonkin 2006] Guy, M., Tonkin, E.: "Tidying up Tags?"; D-Lib Magazine, 12, 1 (2006)

[Golden and Huberman 2006] Golder, S. and Huberman, B. A.: "Usage patterns of collaborative tagging systems"; Journal of Information Science, 32, 2(2006),198-208

[Laniado et al. 2007] Laniado, D., Eynard, D., Colombetti,M.: "Using WordNet to turn a Folksonomy into a Hierarchy of Concepts"; Proc. SWAP, CEUR, Bari (2007)

[Lee et al. 2007] Lee, S., Han, S.: "Qtag: introducing the qualitative tagging system"; Proc. Hypertext, ACM Publishing, Manchester (2007), 35-36.

[Luhn 1958] Luhn, H.P.: "The automatic creation of literature abstracts"; IBM Journal of Research and Development, 2 (1958), 159-165.

[Salton et al. 1974] Salton, G., Yang, C.S., Yu, C.T.: "A Theory of Term Importance in Automatic Text Analysis"; Journal of the American Society for Information Science, 26, (1974), 33-44

[Sen et al. 2007] Sen, S., Harper, F., LaPitz,A., Riedl,J.: "The Quest for Quality Tags"; Proc. Supporting group work, ACM Publishing, Sanibel Island (2007), 361-370

[Van Dongen 2000] van Dongen, S.: "Graph Clustering by Flow Simulation"; PhD thesis, University of Utrecht, (2000).

[Vander Wal 2005] Vander Wal, T.: "Explaining and Showing Broad and Narrow Folksonomies"; (2005) `http://www.personalinfocloud.com/2005/02/explaining_and_.html`