# Enriching Tagging Systems with Google Query Tags

Christoph Trattner, Denis Helic
*IICM & KMI,Graz University of Technology*
*Inffeldgasse 16c, 8010 Graz, Austria*
*E-mail: ctrattner@iicm.edu, dhelic@tugraz.at*

Seid Maglajlic
*RailNetEurope, Austria*
*Annagasse 12/5, 1010 Vienna, Austria*
*E-mail: seid.maglajlic@rne.at*

**Abstract.** *As recent research shows, efficient navigability of tagging systems is only possible if the number of tags grows hand in hand with the number of tagged resources. However, the number of resources grows typically faster than the number of tags. In this paper we analyze how enriching of user tags with tags generated from Google queries influences navigability in tagging systems. The analysis dataset comes from an online encyclopedia called Austria-Forum. The first results are promising and show an increase in the number of resources that can be efficiently reached by navigation.*

**Keywords.** Tagging systems, tags, resources, tag clouds, navigation, social networks

## 1. Introduction

There is a widespread opinion that tagging systems in general and tag clouds in particular are useful for navigating the resources in social tagging systems. However, recent research has questioned this assumption [3, 4, 5]. As the first research results show, navigation efficiency of single tags, as well as tag clouds leaves much to be desired. Among others the research studies have identified the following reasons for such disappointing navigation characteristics of tagging systems.

Firstly, at the level of single tags [3] showed that the number of resources in a tagging system increases more rapidly than the number of tags assigned to those resources. This leads to a situation where each tag is assigned to more and more resources and as such loses its potential as an efficient navigational tool, simply because of a fact that clicking on a particular tag confronts users with an increasingly larger list of resources that exceeds users' cognitive limits as well as technical limitations of the system.

Secondly, at the level of tag clouds [4] showed that user interface restrictions produce a tag cloud network that is sparely connected with a large number of tag clouds that are not strongly connected with each other. In other words, there is a huge number of isolated tag cloud "islands" that are not connected with the main navigational component in a tagging system. Once the users have navigated to such "islands" they cannot escape from it by purely tag cloud supported navigational means. Rather, they need to go back to the homepage, or use search mechanism to find further resources of interest. The problem becomes even more serious when cognitive limits of the user are taken into account, e.g. when pagination is used to show resource lists after a tag selection. Again, the reason for this behavior of tag cloud networks is similar to that found on the single tag level. Namely, the number of tags per resource seems to be simply too small to achieve more efficient navigational properties [5]. In addition, as [4] data analysis showed the above recognized problems are substantial for tagging systems in their beginning phase - where the tags to resource ratio is lower than in a mature tagging system.

A simple approach to support the tag creation process by users would be to enrich resources with automatically generated tags. One obvious approach would be to apply full-text processing of resources and extract "significant" keywords by using standard information retrieval measures such as $tf * idf$. However, such an approach yields only tags which are already present in the text of the resources and does not, in our opinion, create an added value in a tagging system.

Another approach, presented by [1] seems to be much more promising. Essentially, one collects the Google queries that had in their result lists the resources in question and adds the query terms as the new tags. Since Google ranking is based on PageRank algorithm that takes into account the network structure of the Web, i.e.

the anchor text of links to found resources, one would assume that the quality of these newly added tags is highly satisfactory. [1] proves this natural assumption. Thus, tags added in this fashion provide an added value to the underlying tag database.

In this paper we investigate to what extent such Google query added tags could improve navigability of tagging systems especially in the beginning phase. We analyze the tagging data from an emerging tagging system called Austria-Forum, Google added tags, and the combined dataset and measure the following properties for each of these tagging datasets:

1. Number of resources and tags
2. Tag cloud resource coverage
3. Largest strongly connected component of the tagging network

The rest of the paper is organized as follows. In Chapter 2 we present our approach to data analysis in tagging systems. Chapter 3 shortly presents the Austria-Forum system, its tagging dataset, as well as the approach to acquiring new tags from Google queries. In chapter 4 we present the results of the tagging data analysis. Finally, Chapter 5 discusses shortly the results and provides a couple of ideas for future work.

## 2. Approach

In this paper we model the tagging data as a pair of the form $(r,t)$ where $r$ is a resource from the resource set $R$ and $t$ is a tag from the tag set $T$. Henceforth we call such a pair *bookmark* and denote it with $b$. The set of all bookmarks in a tagging system is denoted with $B$. For a complete formal definition of tagging data taking also users into account see [4].

The main navigational aid in tagging systems is a tag cloud. We define a tag cloud $TC$ as a particular selection of tags from the tag set $T$, i.e. a tag cloud is a subset of $T$. Typical criteria for selecting tags include most frequent tags, related tags, or similar tags. The number of tags $n$ in a tag cloud is the cardinality of the $TC$ set.

An important property of a tag cloud is its resource set. We define the resource set $R_{TC_i}$ of a tag cloud $TC$ as a set of all resources $r$ such that $t \in TC$ and $b = (r,t) \in B$. Further, absolute coverage $acov_{TC_i}$ of a tag cloud $TC_i$ is the cardinality

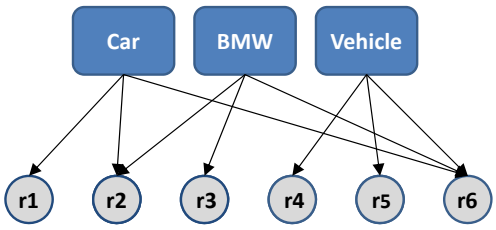



**Figure 1: A (simplified) tag cloud TC with acov = 6.**

of the set $R_{TC_i}$ and relative coverage $cov_{TC_i}$ of a tag cloud $TC_i$ is the value of absolute coverage normalized by the total number of resources in a tagging system, i.e. by the cardinality of the resource set $R$ (see Figure 1). Obviously, coverage is an important property for measuring navigability of tagging systems. This property tells us *how many resources can be accessed from a single tag cloud*.

Usually, some of the tags are assigned to a large number of resources - hundreds or even thousands of resources. When users click on such a tag the system presents a paginated list of resources. Typically, 10, 20, or 30 resources are presented to the users. To model these limitations and investigate their influence on navigability in tagging systems we limit the number of resources per tag in a tag cloud. We denote this parameter with $k$ and define the $k$-limited resource set $R^k_{TC_i}$ of a tag cloud $TC_i$ as a set of all resources $r$ such that $t \in TC$ and $b_{1...k} = (r,t) \in B$.

Now, let us model the navigation process in a tagging system. Typically, navigation in tagging systems acts upon the following pattern. When users access the homepage of the system a global tag cloud is presented. Upon clicking on a tag a paginated list of resources is shown. Once the users have selected a specific resource the system presents the resource and a resource-specific tag cloud. By selecting a tag from such a resource-specific tag cloud the navigation process is continued in the same manner.

For the modeling of such a navigation process one can choose the structure of a directed graph. Nodes in this graph are resource-specific tag clouds and edges are links between the tag clouds, where a tag cloud has links to all resource-specific tag clouds of resources from

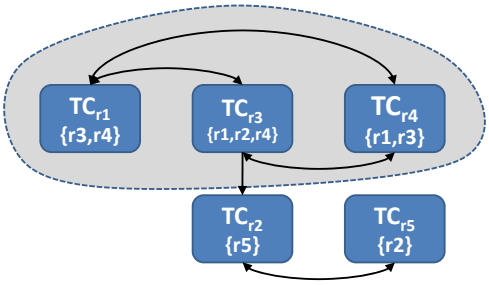



**Figure 2: Tag cloud network with the LSCC (gray cluster).**

its $k$-limited resource set $R^k_{TC_i}$. Formally, the $k$-limited link set $L^k_{TC_i}$ of a tag cloud $TC_i$ is a set of ordered pairs $(TC_i, TC_l)$ where $TC_l$ is a resource-specific tag cloud of a resource $r_l$ from $R^k_{TC_i}$. Finally, let $LS^k$ be union of $k$-limited tag cloud link sets of a tagging system. Then, the $k$-limited tag cloud network $N^k$ is a directed graph $N^k = (TCS, LS^k)$ (see Figure 2).

Apart from coverage we believe that one specific property, namely the largest strongly connected component (LSCC) of the tag cloud network, is of primary importance for the navigability of a tagging system. Informally, the LSCC is the largest sub graph in which every node can be reached by every other within this sub graph (see also Figure 2). Thus, this property of a tag cloud graph tells us *how many resources users can reach by navigating tag clouds on a global level*. Once the users are inside the LSCC they can, at least theoretically, reach all resources from that component.

## 3. Austria-Forum and Google Datasets

Austria-Forum[1] is JSPWiki based information system that manages a very large repository of information items, where new information items are easily published, edited, checked, assessed, and certified, and where the correctness and a high quality of each of these items is backed by a person that is accepted as an expert in a particular field. [4, 6]

The system was released at the beginning of October 2009 with about 30,000 pages and 60,000 media files available. At the moment of writing the system contains over 100,000 information items including pictures, movies, PDF and DOC-files. Since Austria-Forum can be seen as a collection of different encyclopedia systems within one system one challenge was to solve the issue of bad navigability due to weakly connected sets of articles. Therefore, a simple built-in tagging system was implemented [6], which applies resource-specific tag clouds for navigating within related resources.

Currently in Austria-Forum the tags related to the resource tags are calculated and included into the resource-specific tag cloud. Moreover the $n$ top tags with the highest frequencies are chosen for a particular tag cloud. In cases where $n$ tags are not available, only available tags are shown (or none if the resource has no tags at all). Lastly, due to the technical limitations the resource list is paginated and it is practically $k$-limited in length.

However, this approach works efficiently only if the tagging system is in a more mature state and provides a higher number of tagged resources and tags of a reasonable quality. Moreover, one need to achieve a high coverage factor together with a large LSCC in order to support efficient navigation processes. For that purpose we wanted to investigate the possibilities of adding Google query tags and how these new tags would influence navigability in Austria-Forum.

**Dataset AF (Austria-Forum):** The dataset consists of 32,480 bookmarks and 12,871 unique resources from AF (see Table 1). Note that $r_{AF}$ denotes a resource from the system, including also non-tagged resources. The set has been primarily created by the members of the editorial board. The system and the dataset are only in the beginning phase of the tagging process, which is reflected in the ratio of tagged to all resources (0.32).

**Table 1: Dataset AF as of 01-09-2010**

| #r | #b | $\frac{\#b}{\#r}$ | $\frac{\#r}{\#r_{AF}}$ |
|---|---|---|---|
| 12,871 | 32,480 | 2.52 | 0.32 |

**Dataset G (Google):** This dataset *also* consists of resources from Austria-Forum that was found with Google search engine. The tags from this dataset are filtered out of Google search queries used to find Austria-Forum resources. Technically, this information has been collected by parsing HTTP referrer information

[1] http://www.austria-lexikon.at

from the log files [1]. For example, for http://www.austria-lexikon.at/af/Wissenssammlungen/Symbole/Niederoesterreich_Landespatron we find the following referrer header in the log file: http://www.google.at/search?hl=de&q=heiliger+leopold+ursprung&btnG=Suche&meta=cr%3DcountryAT&aq=f&oq=. The following tags can be extracted from this HTTP header: **heiliger**, **leopold**, **ursprung**.

To filter out the noise tags the dataset was cleaned by applying a stop word filter based on a German stop word list provided by solariz.de[2] and a character filter from Bibsonomy[3]. A stemming approach was not used for data normalization in this case. We parsed the Google queries of the last 60 days.

The final dataset consists of 10,659 resources and 44,365 bookmarks (see Table 2).

**Table 2: Dataset G as of 01-09-2010**

| #r | #b | $\frac{\#b}{\#r}$ | $\frac{\#r}{\#r_{AF}}$ |
|---|---|---|---|
| 10,659 | 44,365 | 4.1 | 0.27 |

**Dataset AF+G (Austria-Forum and Google):** This dataset is a merge of the datasets AF and G.

**Table 3: Dataset AF+G as of 01-09-2010**

| #r | #b | $\frac{\#b}{\#r}$ | $\frac{\#r}{\#r_{AF}}$ |
|---|---|---|---|
| 20,688 | 76,287 | 4.36 | 0.53 |

## 4. Data Analysis

One important factor that influences navigation within tagging systems is the number of tagged resources. Since Austria-Forum provides 39,168 resources, at least the same number of bookmarks are needed within a tagging system to at least theoretically create a connected graph and even more to create a strongly connected one. Therefore as a first step the number of tagged resources is evaluated.

As shown in Table 2 over 10,659 AF resources could be bookmarked within the last 60 days with the help of Google query tags. Compared to the AF dataset (see Table 5) which shows an average increase of 399.35 tagged resources per 10 days for the last 200 days, an average of around 1,500 new bookmarks per 10 days for the last 60 days has been achieved within the Google dataset (see Table 4). Thus, the Google tagging approach bookmarked on average four times more resources than the human approach within AF the past 60 days. The notion $r_{new}$ is used to denote new resources.

**Table 4: Growth of Dataset G as of 01-09-2010**

| day | #r | $\#r_{new}$ |
|---|---|---|
| -60 | 1,698 | 1,698 |
| -50 | 3,160 | 1,462 |
| -40 | 4,710 | 1,550 |
| -30 | 6,245 | 1,535 |
| -20 | 8,055 | 1,810 |
| -10 | 9,368 | 1,313 |
| now | 10,659 | 1,291 |

**Table 5: Growth of Dataset AF as of 01-09-2010**

| day | #r | $\#r_{new}$ |
|---|---|---|
| -200 | 4,884 | 4,884 |
| -160 | 7,450 | 2,566 |
| -120 | 9,109 | 1,659 |
| -80 | 11,523 | 2,414 |
| -40 | 12,421 | 898 |
| now | 12,871 | 450 |

Another interesting question was whether the Google tagging approach is also increasing the number of tagged resources in the system? Therefore, the number of resources of the combined dataset AF+G was calculated showing an increase of more than 60% to a total of 20,688 resources (see Table 3).

A further important factor within tagging systems as shown by [3] is the number of tags within a tagging system. Since a small vocabulary size coupled with high tag frequencies lead to tags/tag clouds that are poorly navigable [4], an increase of tag vocabulary is preferred within a tagging system. As shown in Table 6 the vocabulary within the AF dataset increased by 394 on average every $10^{th}$ day. 1,624 new bookmarks were added within the same period of time. For the Google dataset (see Table 7) the vocabulary grew the last 60 days by 2,534 on average every $10^{th}$

[2] http://solariz.de/tool-box/deutsche-stopwords.htm
[3] http://www.kde.cs.uni-kassel.de/ws/dc09/dataset

**Table 6: Vocabulary growth of Dataset A as of 01-09-2010**

| day | #$t$ | #$b$ | #$t_{new}$ | #$b_{new}$ |
|---|---|---|---|---|
| -200 | 3,202 | 10,526 | 3,202 | 10,526 |
| -160 | 7,829 | 19,411 | 4,627 | 8,885 |
| -120 | 8,980 | 23,943 | 1,151 | 4,532 |
| -80 | 10,009 | 28,478 | 1,029 | 4,535 |
| -40 | 10,628 | 31,086 | 619 | 2,608 |
| now | 11,097 | 32,480 | 469 | 1,394 |

**Table 7: Vocabulary growth of Dataset G as of 01-09-2010**

| day | #$t$ | #$b$ | #$t_{new}$ | #$b_{new}$ |
|---|---|---|---|---|
| -60 | 3,906 | 6,269 | 3,906 | 6,269 |
| -50 | 7,020 | 12,586 | 3,114 | 6,317 |
| -40 | 10,018 | 19,166 | 2,998 | 6,580 |
| -30 | 12,772 | 25,645 | 2,754 | 6,479 |
| -20 | 15,615 | 32,733 | 2,843 | 7,088 |
| -10 | 17,743 | 38,418 | 2,128 | 5,685 |
| now | 19,867 | 44,365 | 2,124 | 5,947 |

day. 6,337 resource were bookmarked on average every $10^{th}$ day. These results show again, that the Google query tag approach is better than human tagging approach at increasing tag vocabulary in a short period of time. The notions $t_{new}$ and $b_{new}$ denote a new tag and a new bookmark respectively.

But would Google tags also increase the number of tags when combined with the AF Dataset? Table 8 shows the parameters of the combined dataset. Thus, dataset AF+G has 27,824 tags as compared to 11,097 tags in Austria-Forum, which again increases the vocabulary size of dataset AF by nearly 150%. Also, the number of bookmarks is increased by 135%.

Another interesting observation is that the number of new tags in the Google dataset decreases more slowly with time than the number of new tags added by users in Austria-Forum. This means that the Google query tags, at least in the beginning phase, can provide a more steady growth of new tags. As discussed above this proves to be of primary importance for navigability of tagging systems.

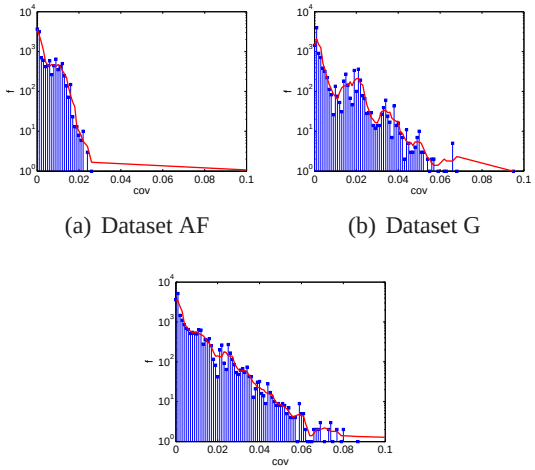
Figure 3 shows a histogram of tag clouds' coverage with a standard tag cloud size of 20, i.e. Top20 algorithm. Obviously, there is an increase in coverage values for the combined

**Table 8: Dataset AF+G as of 01-09-2010**

| day | #$r$ | #$t$ | #$b$ |
|---|---|---|---|
| now | 20,688 | 27,824 | 76,845 |

dataset AF+G. Thus, there is an increased number of tag clouds of the size 20 with higher coverage values. Figure 4 shows average coverage values for dataset AF, G and an increased coverage value of the combined dataset AF+G by approximately 85%. Thus, a number of resources reachable by navigation from a single tag cloud almost doubles on average with the combined dataset.

Similarly to coverage the size of the LSCC could be also increased (see Figure 5). The graph shows that without taking parameter $k$ into account the size of this component for the combined dataset is almost 19,000 resources as compared to approximately 11,000 without Google tags. There is an increase of approximately 70% resulting in almost 50% of all Austria-Forum resources that can be reached in the combined dataset by navigating only in tag clouds. Taking the parameter $k$ into equation shows an increase of more than 100% for $k$ values of 10 and 20.




(a) Dataset AF (b) Dataset G

(c) Dataset AF+G

**Figure 3: Histogram of tag cloud coverage values for dataset AF,G and AF+G with N-TC=20 (p=0.001).**

## 5. Conclusion

As the data analysis shows enriching the

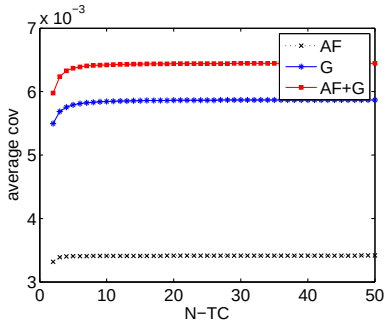
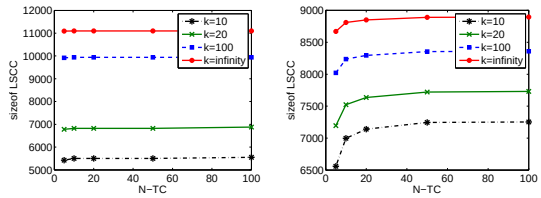
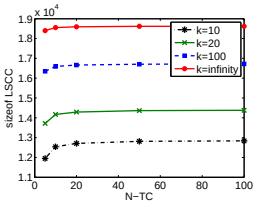



**Figure 4: Average coverage values for all three dataset AF,G and AF+G over tag cloud size N.**




(a) Dataset AF (b) Dataset G




(c) Dataset AF+G

**Figure 5: LSCC over *n*=5,10,20,50,100 and *k*=10,20,100,∞.**

tagging data with automatically collected tags from Google queries increase not only the number of tagged resources but more importantly also the number of tags. As previous research has shown having a steady growth of tags is a prerequisite for efficient navigation in tagging systems. Moreover, the analysis shows that navigation properties of tag clouds, such as tag cloud coverage, or the LSCC are also improved with the discussed approach. Such results are of a primary importance particularly for new tagging systems where the number of bookmarks and tags is relatively small.

Since the work presented in this paper was focusing on improving connectivity of a weakly growing and weakly connected tagging system in the startup phase an interesting focus for future work is to measure average path length for datasets G and AF+G. Average path length is one of the most important factors concerning navigation because it is a theoretical measure of how many clicks on average users need to navigate to a single resource in a tagging system. Thus, we will investigate how effective Google query enriched tag clouds are at improving navigation paths within an existing tag cloud network.

Furthermore, we will also investigate tag cloud entropy values (entropy measures the information value of a tag cloud) in the future. Since tag clouds with low entropy values are found to be better navigable and more usable [2] than tag clouds with high entropy values, we will investigate the effect of Google enriched tag clouds on tag cloud entropy.

## References

[1] Antonellis I., Garcia-Molina H., Karim, J.: Tagging with queries: How and why. ACM WSDM 2009 ; 2009.

[2] Aouiche K., Lemire D., Godin R.: Web 2.0 OLAP: From Data Cubes to Tag Clouds. 4th International Conference, WEBIST 2008, Lecture Notes in Business Information Processing, 18. Berlin, Heidelberg: Springer Berlin Heidelberg ; 2008.

[3] Chi, E.H. & Mytkowicz, T.: Understanding the efficiency of social tagging systems using information theory. Proceedings of the 19th ACM HT conference - HT '08 ; 2008.

[4] Helic D., Trattner C., Strohmaier M.: Measuring Navigability of Socially-Constructed Links in Tagging Systems. Submitted to Hypertext 2010 ; 2010.

[5] Li R., Bao S., Yu Y., Fei B., Su Z.: Towards effective browsing of large scale social annotations. Proceedings of the 16th Web Conference ; 2007.

[6] Trattner C. & Helic D.: Extending the Basic Tagging Model: Context-Aware Tagging, IADIS International Conference on WWW/Internet ; 2009.