LINKING RELATED CONTENT IN WEB ENCYCLOPEDIAS WITH SEARCH QUERY TAG CLOUDS

Christoph Trattner*
Knowledge Management Institute and
Institute for Information Systems and Computer Media
Graz University of Technology
Inffeldgasse 16c, 8010 Graz, Austria

ABSTRACT

In this paper we present a novel tool for exploring related resources in Web encyclopedias called *QueryCloud*. Typically, users come to an encyclopedia from a search engine such as Google, Yahoo! or Bing and upon reading the first page on the site they leave it immediately thereafter. To tackle this problem in other systems such as Web shops, additional browsing tools for easy finding of related content are provided. In order to overcome this issue in the context of Web encyclopedia systems, we introduce a tool called *QueryCloud*. The tool combines two promising approaches – tag clouds and historic search queries – into a new single one, i.e. each document in the system is enriched with a tag cloud containing collections of related concepts populated from historic search queries. To test the feasibility of the approach, we integrated a prototypical implementation of the tool into a large Web encyclopedia called the Austria-Forum and conducted several experiments on a theoretical and empirical level. As our experiments show, *QueryCloud* provides a great alternative to traditional forms of tag cloud creation. With several experiments on a theoretical and empirical level we show that tag clouds generated by our system outperform tag clouds that are based on user-tags in terms of linking content and navigability. This work is relevant for researchers interested in the navigability of emergent hypertext structures and for engineers seeking to improve the navigability of large information systems, such as Web encyclopedias.

KEYWORDS

search query clouds, tag clouds, query tags, tags, linking, web encyclopedia

1. INTRODUCTION

Nowadays, content in Web encyclopedias such as Wikipedia is mainly accessed through search engines (Wikimedia 2010). Typically, users with a certain interest in mind go to a search engine such as Google, Yahoo! or Bing, define a search query there and click on a link from the result list from which they are referred to an article within Wikipedia. Upon reading the document they decide to either go back to the search engine to refine their search, or close their browser if they have already found the information they needed. Such a user behavior on encyclopedia sites is traceable through a typical high bounce rate (Alexa 2010, Gleich et al. 2010). Essentially, users do not "really" browse in online encyclopedia systems such as Wikipedia to find further relevant documents (Gleich et al. 2010) - they rather use search engines such as Google, Yahoo! or Bing for that purpose. It is our opinion that Web encyclopedias simply lack usable tools that support users in explorative browsing or searching. For example, in Web based systems such as Web shops, different approaches have been applied to tackle this situation. Amazon for instance offers the user related information through collaborative filtering techniques for each product. Google or Yahoo! apply a similar approach by offering related content (e.g. sponsored links) to the user by taking the users' search query history into account (Mehta et al. 2007).

On the other hand, social tagging systems have emerged as an interesting alternative to find relevant content on the Web (Heymann et al. 2010). These systems apply the concept of social navigation (Millen and

* Part of this research was conducted while the author was working as a visiting researcher at the University of Pittsburgh, School of Information Science.

Feinberg 2006), i.e. users browse by means of tag clouds, which are collections of keywords assigned to different online resources by different users (Heymann et al. 2010) driven by different motivations (Nov and Ye 2010, Ames and Naaman 2007).

In this paper we introduce a novel tool called *QueryCloud* to offer related content to users of Web encyclopedias. Essentially, the tool is based on the simple idea of integrating a tagging system into a Web encyclopedia and offering related content to users via the so-called resource-specific tag clouds by *automatically* linking related documents over the users search query terms. In this way two promising approaches are successfully combined into a new single one - tag clouds and historic search queries.

To test the approach, we implemented a prototype of the tool and integrated it in a large Web encyclopedia called the Austria-Forum. To evaluate the system by means of link quality, tag network quality and navigability, we conducted several experiments on a theoretical level. Additionally to this, we conducted a user study to investigate whether or not the tags generated by our system are also meaningful for the user by describing the content of a particular Web page. Hence, the overall contribution of this paper can be summarized as follows:

- Introduction of a novel tool called *QueryCloud* that links related content in Web encyclopedia system via so-called resource specific query tag clouds with each other.

- Evaluation of the tool by integrating it into a large Web encyclopedia system called the Austria-Forum and comparing it on a theoretical and empirical level with tags and tag clouds that are based on tags generated by real users of this system.

Essentially, the paper is structured as follows: In Section 2 we present the basic idea of this new approach. Section 3 shortly discusses the implementation of the novel tool. In Section 4 we provide an analysis of the potentials and limitations of our tool. Section 5 discusses related work. The final Section 6 concludes the paper and provides insights in the current progress of the project.

2. APPROACH

The basic idea of *QueryCloud* is to offer related content to users via the so-called resource-specific tag cloud by *automatically* linking related documents via the search query terms of the users. On the one hand, tag clouds represent an interesting alternative navigation tool in modern Web-based systems (Helic et al. 2010, Trattner et al. 2010b). Moreover, they are very close to the idea of explorative browsing (Sinclair and Cardew-Hall 2008), i.e. they capture nicely the intent of users coming to a system from a search engine - users have searched in e.g. Google, Yahoo! or Bing and now they click on a concept in a tag cloud that reflects their original search intent. On the other hand, search query history, i.e. queries that are “referrers” to found documents are an invaluable source of information for refining user search in the system. It is our belief that an integration of such a tool online encyclopedia systems would greatly contribute to leading users to related documents.

In order to make the idea work the users search query history needs to be obtained in a resource specific way. This is achieved by collecting the HTTP-Referrer information from a user coming from a search engine to a particular website (resource) within a Web based encyclopedia system such as Austria-Forum. To retrieve the user’s HTTP-Referrer information, we use in *QueryCloud* the simple approach of a JavaScript code snippet (see Tag Collection Module in Section 3) that has to be included by the owner of the particular Web site.

Now, how does the script function? Let us give an example: Suppose a user goes to Google and searches for “Elfride Jellinek Biographie” and selects the link <http://www.austria-lexikon.at/af/ Wissenssammlungen/Biographien/ Jelinek%2C_Elfriede> from the results list that refers to the Elfriede Jelinek biography in the Austria-Forum. The *QueryCloud* Tag Collection Module then simply parses the HTTP-Referrer <http://www.google.at/search?hl=de&q=elfriede+jellinek+biographien&btnG=Suche&meta =cr%3DcountryAT&aq=f&oq=> information of this site and extracts from it the user’s search query terms “elfride”, “jellinek”, “biographien”. The tags are then stored in the *QueryCloud* Tag Store Module (see Section 3). This procedure is actually performed, whenever a user lands on a Web page from a search engine within the Austria-Forum. The tags are then used to create a resource-specific tag cloud for each site. This is done by the *QueryCloud* Tag Cloud Generation Module (see Section 3). The Tag Cloud Presentation Module (see Section 3) renders the tag cloud then in a visually appealing fashion and presents it to the user. Hence, two pages (or even more) are linked with each other by this approach, if they have the same query tag in common. Upon clicking on a particular query tag in a resource-specific tag cloud the user is provided with a list of links of the resources which have this query tag in common (see Figure 1). By clicking on a particular link in the resource list the user is then forwarded to the resource she was searching for (see Figure 2).

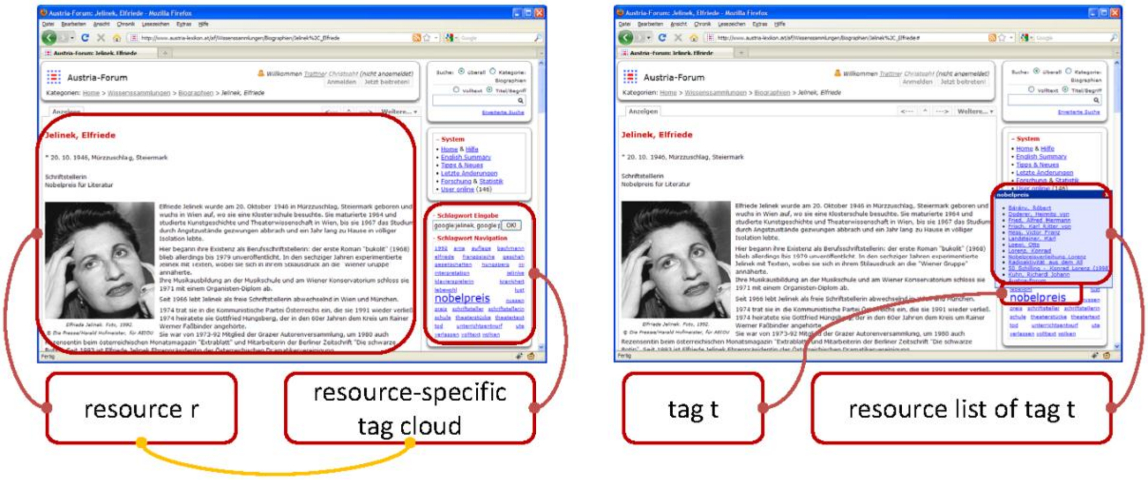


Figure 1. Example of resource-specific query tag cloud and resource list of a tag within Austria-Forum rendered by *QueryCloud*.

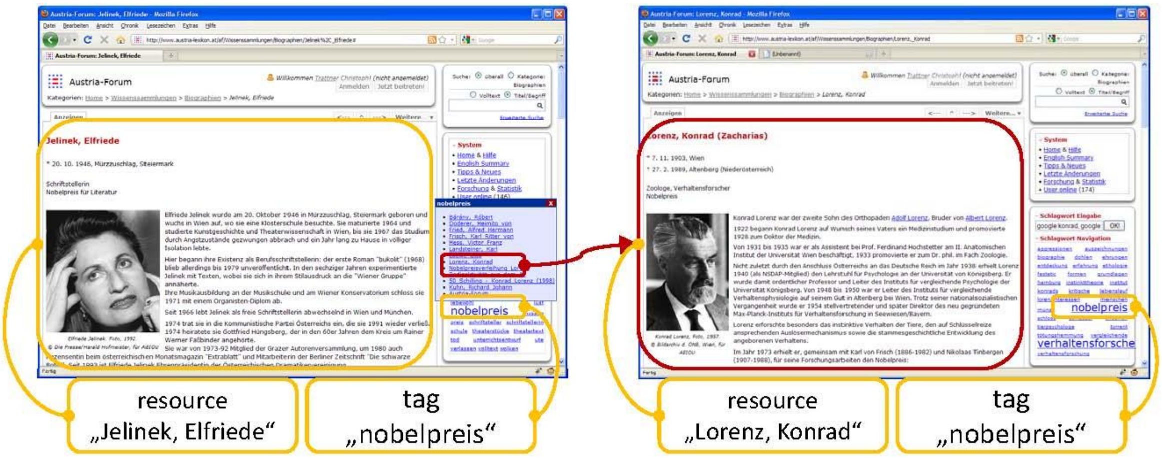


Figure 2. Example of tag cloud driven navigation within Austria-Forum using *QueryCloud*.

3. IMPLEMENTATION

The first prototypical implementation (cf. Trattner and Helic 2010) consists basically of four independent different modules (see Figure 3).

- **Tag Collection Module:** The tag collection module is the first module within *QueryCloud* system. Basically, this module consists of a simple client part module which retrieves HTTP-Referrer information, time information and target page of a user coming from a search engine such as Google, Yahoo! or Bing to a website. The tag collection module is implemented in JavaScript and can easily be integrated into any website with only one line of code

```
<script src=”http://xxx.xxx.xxx.xxx/ref.js” type=”text/javascript”></script>
```

where the *ref.js* file contains the following code (cf. Antonellis 2009)

```
document.write('<scr' + 'ipt src="http://xxx.xxx.xxx.xxx/getReferrer.jsp?referrer=' +
encodeURIComponent(document.referrer) + '&url=' + encodeURIComponent(document.URL) +
'&time=' + (new Date()).getTime() + '" type="text/javascript"></scr' + 'ipt>');
```

- **Tag Storage Module:** This module is the actual heart peace of *QueryCloud* system. It provides a couple of interface routines for storing and deleting data from the database back-end module which is actually implemented with Apache Lucene[1]. Due to reasons of performance two index files are generated: One for expressing the tag resource-relations ($t_n|r_i,\ldots,r_j$) and one for expressing the resource-tag relations ($r_n|t_i,\ldots,t_j$). In this way, tags and resources can be searched independently from each other, to either create a tag cloud for a specific resource or to create a resource list for a specific tag.

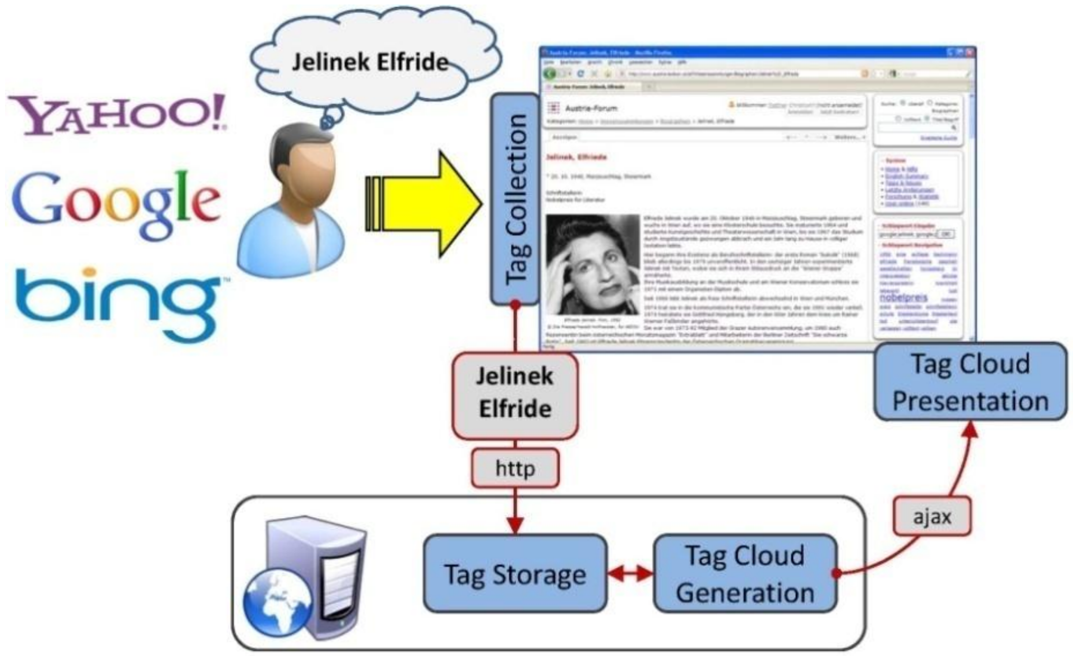


Figure 3. *QueryCloud* - structural diagram.

- **Tag (Cloud) Generation Module:** To provide the access to related documents a resource-specific search query term/tag cloud is calculated by this module. This tag cloud is of the form $TC_r = (t_1,\ldots, t_n, r_1,\ldots, r_m)$, where $r_1,\ldots, r_m$ are the resources which have any of the query tags $t_1,\ldots, t_n$ in common. For retrieving the query tags and the corresponding resources (cf. Figure 1), this module provides a simple HTTP interface using the following two functions:
    - *GetTagCloud(<URL>,<max tag cloud size>)* generates a XML representation of a query tag cloud
    - *GetResources(<URL>,<tag>,<max resource list length>)* generates a XML representation of the resource list for a particular query tag.

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

- **Tag Cloud Presentation Module:** This module is a client-side AJAX module implemented in JavaScript. It retrieves the XML representation of a query term/tag cloud or an XML representation of a resource list of a particular query term/tag from the tag cloud generation module and renders a tag cloud in a visually appealing fashion (cf. Trattner and Helic 2009).

4. EVALUATION

To investigate the feasibility of the tool before actually deploying it, we first decided to only integrate the *QueryCloud* tag collection module into the Austria-Forum life system to collect the search query terms from users coming from search engines such as Google, Bing or Yahoo! to the Austria-Forum for a period of 240 days. On the basis of this dataset, we conducted several experiments on a theoretical and empirical level to evaluate the feasibility of the approach. In the following subsections we will give a short description of our test system, the tag data sets used in the experiments, a detailed description of the experiments and their results.

4.1 Test System

As described before, we used a Web encyclopedia called the Austria-Forum[2] (Trattner et al. 2010a) as our test system to investigate the feasibility of our approach. Basically, Austria-Forum is a wiki-based online encyclopedia containing articles related to Austria. The system comprises a very large repository of articles, where new articles are easily published, edited, checked, assessed, and certified, and where the correctness and quality of each of these articles is assured by a person that is accepted as an expert in a particular field. Currently, as of July 2010, the system contains more than 130,000 information items, attracts more than 4,000 distinctive users (over 80% coming from search engines such as Google, Bing or Yahoo!) each day and is known as the biggest online encyclopedia system on the Web containing content about Austria (Trattner et al. 2010a).

4.2 Baseline

As baseline for our experiments a tag dataset (further referred as the AF dataset) of the Austria-Forum was used. Since, the tags are generated by real users of the system to describe or categorize (Körner et al. 2010a) the content, we used this dataset as our almost golden standard to compare our *QueryCloud* approach with. As of Febuary 2010, the AF tag dataset contains 11,516 tags (#tags), 13,398 resources (#res) and 33,737 tag assignments (#tas) (see Table 1).

Table 1. AF dataset: Number of resources (#res), number of tags (#tags) and number of tag assignments (#tas).

	#res	*#tags*	*#tas*
AF dataset	13,398	11,516	33,737

4.3 Measuring Tag Quantity

Since the success of the whole concept depends on automatically applying tags to the resources of a Web based encyclopedia system we conducted in the first step an experiment that measured tag quantity. In

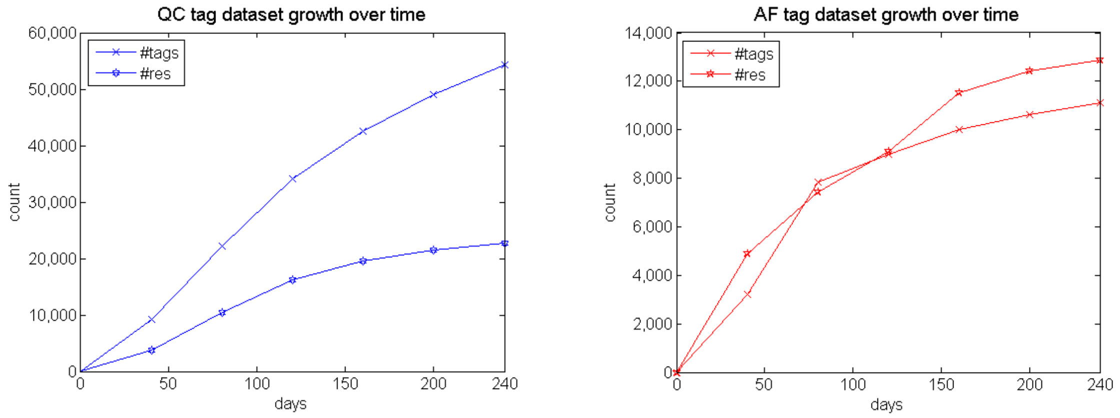
[2] http://www.austria-lexikon.at/

particular, we investigated the growth of the number of resources (#r), the growth of the number of generated tags (#t) and number the tag assignments (#tas) in general for a period of 240 days (cf Trattner 2010c).

Essentially, we could observe that *QueryCloud* generated on average 226 new tags (#t) per day and annotated on average new 95 resources every day (#r) (see Figure 4). The average number of tagged resources was 1290 per day. Compared to this, the Austria-Forum taggers annotated 135 resources on average per day, generated on average 46 new tags (#t) and annotated on average 53 new resources (#r) every day, for the time the tagging functionality was first introduced in Austria-Forum until the time 240days have been passed by (see Figure 4). Hence, the *QueryCloud* tagging approach clearly outperforms the Austria-Forum human taggers by annotating 9 times more resources, and generating 5 times more new tags and annotating 1.7 times more resources in the same period of time. As of July 2010, the QC tag dataset (further referred as QC dataset) contains 54,379 tags (#tags), 22,798 resources (#res) and 309,683 tag assignments (#tas) (see Table 2).

Table 2. QC dataset: Number of resources (#res), number of tags (#tags) and number of tag assignments (#tas).

	#res	*#tags*	*#tas*
QC dataset	22,798	54,379	309,683



(a) (b)

Figure 4. Number of tagged resources (#res) and number of generated tags (#tags) over time for the Austria-Forum: QC dataset (Figure a) on vs. AF dataset (Figure b).

4.4 Measuring Link Quality

After measuring the quantity of the tags produced by *QueryCloud* system we had a closer look at the actual “link quality” of the produced tags by *QueryCloud* system. Since the success of the whole concepts is depended on linking related documents over tags that share more than *one* resource with each other, we conducted an experiment measuring the number orphan tags produced by *QueryCloud* system. Orphan tags (cf. Körner et al. 2010a) are basically tags which are applied to only one resource within a tagging system, i.e. they do not connect any resources with each other. Again, we could observe that *QueryCloud* system performs really well by actually producing 7% less orphan tags than the human taggers (AF) do (see Table 3). Note, for the rest of our experiments we only used the cleaned datasets AF and QC, cleaned from orphan tags and their corresponding resources. In Table 4 the “new” statistics of both datasets are presented.

Table 3. Number of tags (#tags) and number of orphan tags (#orphans): QC dataset vs. AF dataset.

	#tags	*#orphans*
QC dataset	54,379	34,962 (70%)
AF dataset	11,516	8,865 (77%)

Table 4. Orphan cleaned up dataset statistics of QC dataset and AF dataset.

	#res	*#tags*
QC dataset	18,831	11,485
AF dataset	12,103	2,207

4.5 Measuring Tag Cloud Network Quality

Another metric we were interested in was the so-called “network quality” of the *QueryCloud* system. In order to measure this property, we first modeled the *QueryCloud* tag cloud network as a simple tag-resource bipartite graph system of the form $V = R \cup T$, where R is the resource set and T is the query tag set (Helic et al. 2010). Hence, for the following experiments, we assume that neither the tag cloud size nor the resource list is limited, which was been shown in one of our own previous works to influence tag cloud navigability significantly (Helic et al. 2011a). For more details, see (Trattner 2011) for a generic solution for that problem, to create efficiently navigable tag cloud networks with limited resource lists.

Since the “link quality” experiment only showed us how many actual useful tags the systems generates (by means of connecting two or more resources with each other), we investigated in this experiment how many resources are connected with each other. To measure this metric, we calculated the connected components (Levine 1965).

Essentially, we could observe that size of the largest connected component in the QC dataset is 99%, i.e. 99% of all resources within Austria-Forum are connected via tag clouds generated by the *QueryCloud* system (cf. Trattner et al. 2010c). Contrary to this, the AF dataset generates a tag cloud network which is only to 94% connected (cf. Trattner et al. 2010c).

4.6 Measuring Tag Cloud Network Navigability

In the fourth experiment, we examined the property of the tool to navigate to related documents within a Web based encyclopedia system (cf. Helic et al. 2010). In (Helic et al. 2010) we have shown that navigable tag cloud networks have certain properties. According to Kleinberg (Kleinberg 2000a, Kleinberg 200b and Kleinberg 2001) a navigable network can be formally defined as network with a low diameter (Newman 2003) bounded poly-logarithmically, i.e. by a polynomial in log(N), where N is the number of nodes in the network, and an existing giant component.

Thus, as a first step, we examined again the size of the largest connected component. As shown in Section 4.5, *QueryCloud* generates a tag cloud network whose largest connected component contains almost all nodes (99%) of the network. Contrary to this, the AF dataset generates a tag cloud network which is “only” connected to 94%.

Thereafter, we calculated the number of shortest path pairs within *QueryCloud’s* tag cloud network and investigated the effective diameter of the network. As Figure 6 shows, *QueryCloud* generates a tag cloud network, whose effective diameter is around 6.3 hops while the AF dataset generates a tag cloud network with an effective diameter of around 9.9 hops. Putting the results of these two experiments together we can see that *QueryCloud* produces a navigable tag cloud network, $\log_2(18{,}831) = 9.8 > 6.3$ (Helic et al. 2010, Kleinberg 2000a, Kleinberg 2000b, Kleinberg 2001). Contrary to this, the AF tag cloud network is not navigable, since $\log_2(12{,}103) = 9.4 < 9.9$.

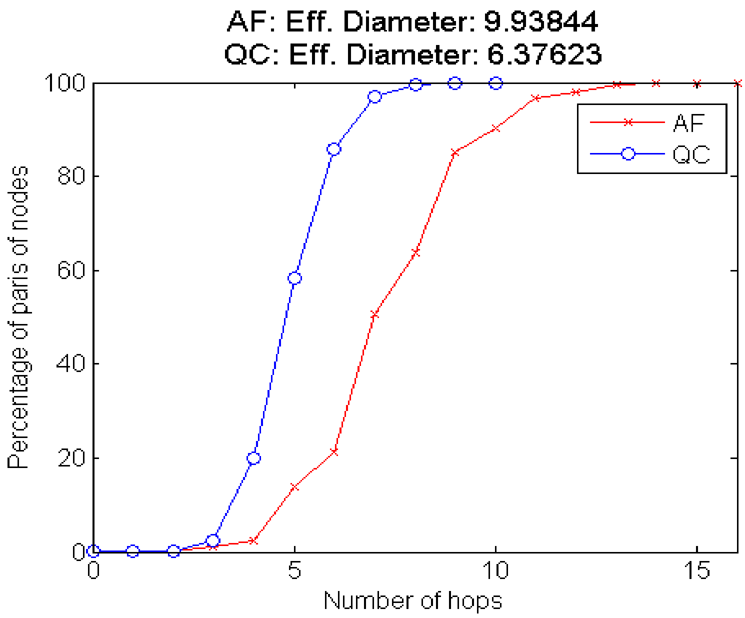


Figure 6. Distribution of shortest path pair lengths for QC tag cloud network (blue line) and AF tag cloud network (red line).

4.7 Measuring Tag Cloud Network Efficiency

Now, since we have shown that tag cloud network of *QueryCloud* system is navigable, we also wanted to know, how "efficiently" navigable the tag cloud network is in general. As shown by Kleinberg, an "efficiently" navigable network is a network possessing certain structural properties so that it is possible to design an decentralized search algorithm, i.e. an algorithms that only has local knowledge of the network, and whose delivery time (the expected number of steps to reach an arbitrary target node) is poly-logarithmic or at most sub-linear in N, where N are the number of nodes in the network (Kleinberg 2000a, Kleinberg 200b and Kleinberg 2001). Essentially, we implemented a decentralized searcher (see Algorithm 1) based on the ideas of (Adamic and Adar 2005) to evaluate the actual efficiency of the system to navigate to related documents using tag clouds for navigation (Helic et al. 2011). In order to support efficient search, we utilized a hierarchical background knowledge base (in our case a tag-taxonomy) to find related content within a tag cloud network (see Algorithm 1). As appropriate tag taxonomy, we us a tag-taxonomy which is based on the so far best known tag-taxonomy induction algorithm (creating best semantic relations and search results) called Deg/Cooc (Strohmaier et al. 2012).

Algorithm 1. Hierarchical Decentralized Searcher (Adamic and Adar 2005, Helic et al. 2011).

Searcher(<tag-resource graph G>, <start node START>,<target node TARGET>)
T ← GetDegCoocTagTaxonomy(G) // returns a tag-taxonomy based on
// degree centrality and tag co-occurance (Helic et al. 2011)
currentNode ← START
while currentNode != TARGET **do**
neighbors ← get all adjacent nodes ∈ G from currentNode
// finds closest node according to dist = min, where dist(A,B) = h(A)+h(B)-2h(A,B)-1
currentNode ← findClosestNode (neighbors, T)
end while

To find a certain resource A (e.g. tagged as “spätromantik”) from a certain node (e.g. tagged as “schloß”) within the network (see Figure 7 and Figure 8), the searcher first selects all adjacent nodes for the start node and then selects the node from the network (“kirche”) which has the shortest distance dist(A,B) = h(A)+h(B)-2h(A,B)-1 to target node B in the tag-taxonomy, with h(A), h(B) being the heights of the two nodes A,B in the hierarchy and with h(A,B) being the height of the least common ancestor of the two nodes A,B in the hierarchy (Adamic and Adar 2005, Helic et al. 2011). This process is continued until node B is reached. In order to get statistically significant results, we simulated 100,000 user-search-requests randomly requesting for a resource B within the system and starting also starting at randomly selected resource A.

As shown in Figure 9, we can observe that with the help of *QueryCloud* system a user is able to find a resource within Austria-Forum in an efficient way, i.e. within only 8 hops almost 98% of all resources in the Austria-Forum can be reached, $\log_2(18{,}831) = 9.8 > 8$. Contrary to this, the AF tag dataset provides successful finding of related resources in the Austria-Forum only in 68% of the cases and this also in significantly more steps (14 hops) than with the QC dataset.

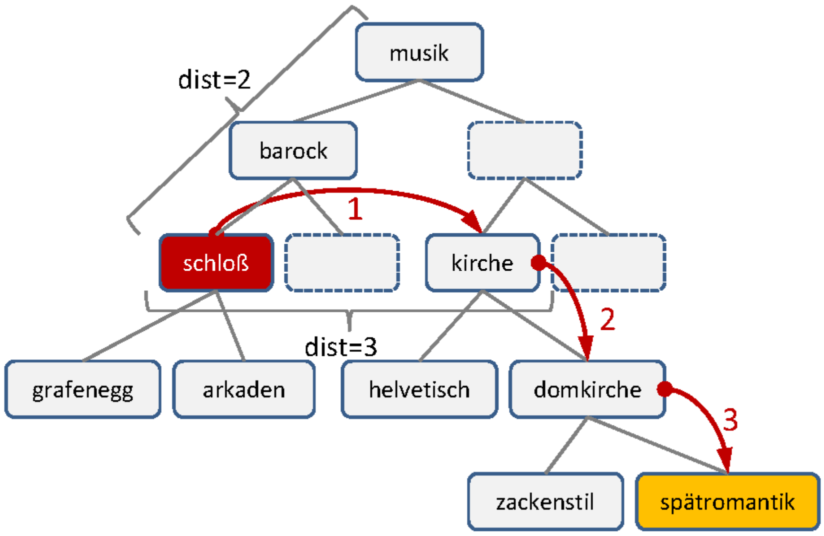


Figure 7. Example of a tag-taxonomy generated for AF dataset.

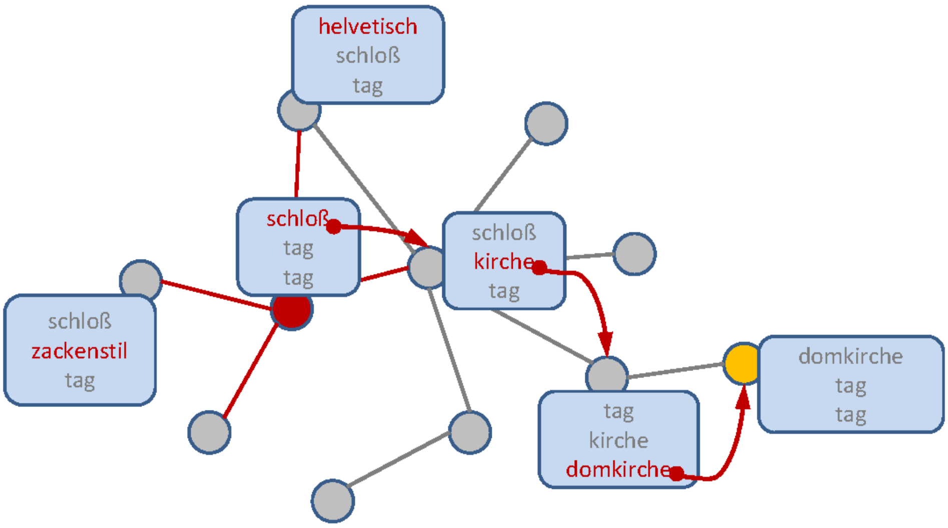


Figure 8. Example of a resource-specific tag cloud network and a search through it utilizing the tag-hierarchy from Figure 7 as back ground knowledge.

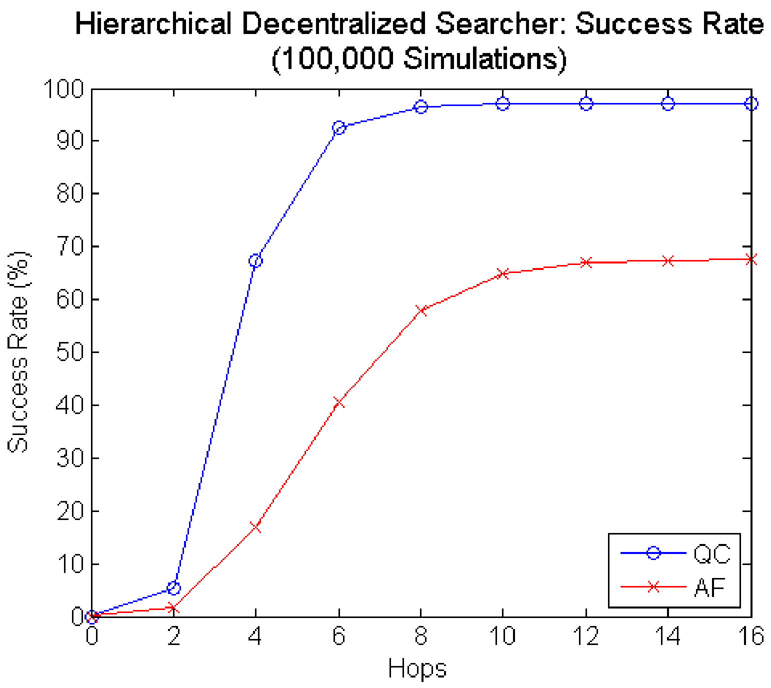


Figure 9. Success Rate plot for the hierarchical decentralized searcher in the QC tag cloud network (blue line) and AF tag cloud network (red line).

4.8 Measuring Tag Relevance

Last but not least, we investigated, in our final experiment, the relevance of the tags, respectively the tag clouds generated by our *QueryCloud* system. Even if related work has shown that query tags can be utilized to generate shorter navigational paths between documents (Antonellis et al. 2010) or that they are very similar to tags generated by the users (Carman et al. 2009), none the previous studies has shown, whether or not query tags provide also a meaningful source for the users to describe Web content. For that reason, we conducted a user study, where we asked our test subjects to find out of a set of query tags the terms which are not relevant for a given Web page in the Austria-Forum system. As baseline for the experiment we again used the user generated tags from the AF tag dataset. The reason why we did not ask our users to explicitly find “the” relevant terms out of the set of query tags, was basically the fact that we observed in a pilot study that users had more problems in highlighting items which are relevant than finding items which are really not relevant. Furthermore, was the number of not relevant items small than the number of relevant items which made it easier for the user to just point out the non-relevant items and to complete the online questionnaire in a faster way. All in all, the setup of the user study was the following:

1. In the first step, we selected uniform at random, 250 resources overlapping resources from both datasets (QC and AF). In order to get meaningful results, only those resources where taken into account for which at least two tags (query tags and user tags) were present per resource. For the samples we took, the size of the tag cloud for the QC dataset was 5.2 terms on average and 2.7 terms on average for the AF dataset.

2. After that, we combined the tags of both datasets. We did that, since we wanted our users not to know, whether they are rating a query tag as relevant for a given resource in Austria-Forum or a user tag. The overlap between the QC dataset and the AF dataset was in general not very high. Contrary to the findings

of (Carman et al. 2009) we can find only 9% of the query tags in the AF dataset and 16% of the user generated tags in the QC dataset.

3. In the third step, we implemented an online study with 50 tasks (=resources) per questionnaire. Overall we created five different questionnaires which covered the sample of 250 resources we have chosen from the Austria-Forum system.

4. Finally, we setup up the user study on one of our servers in the TU-Graz domain and invited collogues and friends to participate in the study. The whole study started at December 1st 2011 and was online for one week. The questionnaire can be still accessed online[3] (see also Figure 10).

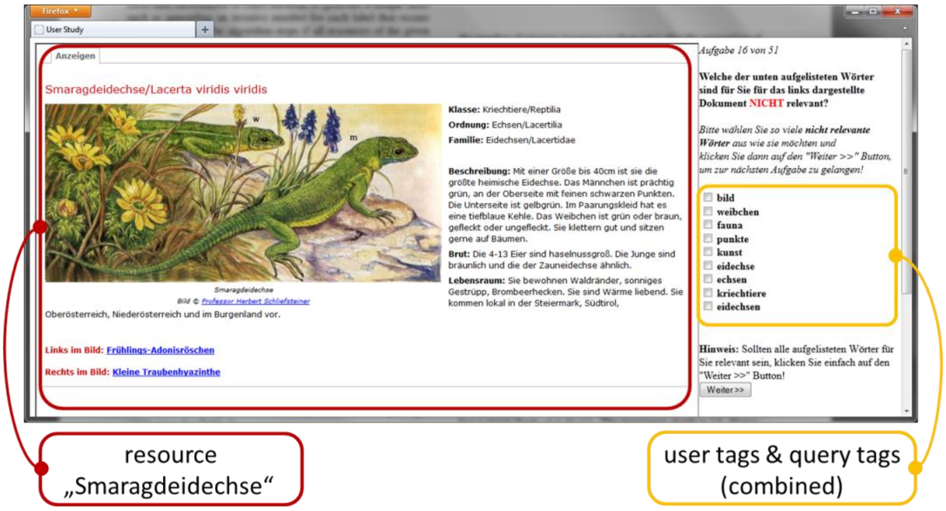


Figure 10. Sample of a resource and the corresponding user tags and query tags (combined in one list) as presented to the users during the user study.

Overall, we had 15 test subjects from three different departments of our university, who participated in the experiment. The test subjects were randomly assigned to each questionnaire. The mean length of time to complete a questionnaire was 35 minutes. For the final evaluation, we took only those tags into account, where at least all three test users had an agreement on. In order to conduct the overall rating quality of our test subjects, we calculated the inter-rater agreement (=k) of the users according to Fleiss'kappa. As shown in Table 5, the mean inter-rater agreement of the users for the query tags was k=0.22 and for the user tags it was k=0.21. According to the inter-rater agreement levels of Landis and Koch (Landis and Koch 1977) this can be interpreted as fair agreement (k=0.21-0.40).

Table 5. Mean Fleiss'kappa (=k) inter-rater agreement for QC dataset and AF dataset over all test users.

	k	*Agreement level*
QC dataset	0.22	Fair agreement
AF dataset	0.21	Fair agreement

In Table 6 the mean non-relevance feedback ratings (=NRR) for the two datasets QC and AF of the user test are presented. As shown, we can observe overall a very small number of tags in both datasets which were rated as non-relevant. For the QC dataset sample the mean non-relevance rating was 4.8% and for the AF dataset sample it was 2.4%. To see whether or not the two values were also statistically significant different,

[3] http://navtag.tugraz.at:8080/UserStudy0.jsp

we performed two-sample t-test. As also shown in Table 6, on a confidence interval of 95%, the differences were not significant, p=0.376.

Even if the experiment did not show, that query tags generated by *QueryCloud* out-perform user generated tags by means of term relevance, the study revealed that (even if query tags are not identical to user tags) they are to a high degree relevant for the user. Overall, we could show that query tags generated by *QueryCloud* are almost to the same degree relevant for the user of a given Web page as user generated tags.

Table 6. Mean non-relevance feedback rating (=NRR) for QC dataset and AF dataset over all test users and p-value for the two sample t-test.

	NRR	*p*
QC dataset	4.8%	0.376
AF dataset	2.4%	

5. RELATED WORK

The almost first notable work in the field of search queries and user tags was a study conducted by (Antonellis et al. 2009). In their work the authors performed a set of experiments to study the information value of search engine queries when treated as "tags" or "labels" for the Stanford domain. In particular the authors tried to answer the question, how much extra information these query tags provide for web pages by comparing them to tags from the Delicious bookmarking site and to the page text. As datasets for their experiments the authors used a self-collected query log dataset retrieved over the users HTTP-Referrer Information (see also Section 3) and a crawled Delicious tag dataset for the stanford.edu domain. The authors conclude their work, that query tags can provide substantially many (on average 250 tags per URL), new tags (on average 125 tags per URL are not present in the page text) for a large fraction of the stanford.edu domain.

In another study which was conducted in the same year (Carman et al. 2009) investigated tags and query logs to see, whether or not the terms people use to annotate websites are also similar to the websites search query terms. Interestingly, the authors found out that the vocabulary used for tagging and search is quite similar, however not identical. Additionally to these findings, the authors tried to answer the question, whether or not search queries are more related to page content than tags generated by users. In a number of experiments, they found out that query tags are more related to page content than user tags. As datasets for their experiments the authors used the AOL search query log and tags from Delicious.

In a very interesting follow-up work (Antonellis et al. 2010) of (Antonellis et al. 2009), the authors describe how they built a navigational tool on the basis of query tags. Additionally to this, they proposed a framework for comparing different tag selection methods. Similar to our own work they investigated user tags and query tags by means of the so-called navigational utility. Put simply, they calculated the navigational utility of a page as the number of resources that could be reached through this page and the closeness to other resources based on the pages tags. The most interesting finding in their work was that based on the measure of the navigational utility of a page, query tags increase the navigational utility of a page more than tags extracted from tags or user tags, or in other words, query tags are a better source for navigational tags than tags extracted from text or tags assigned by users. Their findings actually go along with the results as we also observed them in this paper. However, contrary to our own work, the authors focused in their work only on the tag/link section problem to increase the navigational utility of a page rather than focusing on the general question to what extend query tag clouds outperform user generated tag clouds by means of tag quantity, link/network quality, navigational efficiency and relevance, as we did in this paper.

The last to mention work in this field is a study that was actually conducted by us in 2010. In particular, in (Trattner et al. 2010c) we presented the idea to combine search query tags and user tags to increase the navigability of tagging system. To that end, we introduced measures such as tag cloud coverage and the size of the largest strongly connected component. Overall, we could show that the navigability of tagging systems could be significantly increased, if we enrich an existing tagging system with query log tags.

6. CONCLUSIONS AND CURRENT WORK

In this paper we presented a novel tool called *QueryCloud* for exploring related resources in Web encyclopedias. The tool aims to offer additional navigational paths to related resources for users of such systems in general, and for users who come to these systems from search engines such as Google, Yahoo! or Bing in particular. Furthermore, we showed the potentials and limitations of the tool by integrating it into a large Web encyclopedia system called the Austria-Forum. By comparing *QueryCloud* on a theoretical and empirical level with tag clouds that are based on tags generated by real users of the Austria-Forum we showed that our system out-performs these tag clouds in terms of linking and navigating related content. Or in other words, the results revealed that the system is able to produce a good amount of tags which are of good "link" quality and tag cloud networks with are of good "network" quality and efficiently navigable. Additionally to this, we could show in a user study that query tags generated by *QueryCloud* are almost to the same degree relevant for the user of a given Web page as user generated tags.

Currently, we are evaluating *QueryCloud* with real users, i.e. we deployed the system on the Austria-Forum life server a couple of months ago and track since then the users' click paths. The final goal of this research is to see the differences between the simulations and the user's navigational behavior for both approaches - the Austria-Forum user generated tag clouds and the tag clouds generated by *QueryCloud*.

ACKNOWLEDGEMENT

We would like to thank Prof. Denis Helic, Mr. Christian Körner, Ms. Yiling Lin and Mr. Denis Parra for supporting us with fruitful discussions, valuable inputs, and comments during this work. This work is funded by - BMVIT - the Federal Ministry for Transport, Innovation and Technology, program line Forschung, Innovation und Technologie für Informationstechnologie, project "NAVTAG - Improving the navigability of tagging systems" and the Marshal Plan Funding Agency.

REFERENCES

Adamic, L. and Adar. E. 2005. How to search a social network. Social Networks, 27(3):187-203, July 2005.

Ames, M. and Naaman, M. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 971-980. DOI= http://doi.acm.org/10.1145/1240624.1240772

Antonellis I., Garcia-Molina H. and Karim, J. 2009. Tagging with queries: How and why. In *Second ACM International Conference on Web Search and Data Mining,* Late breaking Results,WSDM 2009, Barcelona, Spain - February 9-12, 2009. URL= http://wsdm2009.org/wsdm2009_antonellis.pdf (last visited 12. July 2010).

Antonellis, I., Karim, J. and Garcia-Molina, H. 2010. Navigating the with Query Tags. Technical Report, Stanford. 2010. URL= http://ilpubs.stanford.edu:8090/943/3/Navigation_tags_TR.pdf (last visited 28. November 2011).

Alexa 2010. Wikipedia.org – Bounce %, URL=http://www.alexa.com/siteinfo/wikipedia.org (last visited 12. July 2010).

Carman, M. J., Baillie, M., Gwadera, R., and Crestani, F. 2009. A statistical comparison of tag and query logs. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in information Retrieval* (Boston, MA, USA, July 19 - 23, 2009). SIGIR '09. ACM, New York, NY, 123-130. DOI= http://doi.acm.org/10.1145/1571941.1571965

Gleich, D. F., Constantine, P. G., Flaxman, A. D., and Gunawardana, A. 2010. Tracking the random surfer: empirically measured teleportation parameters in PageRank. In *Proceedings of the 19th international Conference on World Wide Web* (Raleigh, North Carolina, USA, April 26 - 30, 2010). WWW '10. ACM, New York, NY, 381-390. DOI= http://doi.acm.org/10.1145/1772690.1772730

Helic, D. Strohmaier, M. Trattner, C., Muhr, M. and Lerman, K. 2011. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th international conference on World wide web* (WWW '11). ACM, New York, NY, USA, 417-426. DOI=10.1145/1963405.1963465 http://doi.acm.org/10.1145/1963405.1963465

Helic, D., Trattner, C., Strohmaier, M., Andrews, K. 2011a. Are Tag Clouds Useful for Navigation? A Network-Theoretic Analysis, *Journal of Social Computing and Cyber-Physical Systems*, 1(1): 33-55, 2011.

Helic, D., Trattner, C., Strohmaier, M., Andrews, K. 2010. On the Navigability of Social Tagging Systems, In *Proceedings of the Second IEEE International Conference on Social Computing*, SocialCom2010, Minneapolis, Minnesota, USA, 2010.

Heymann, P., Paepcke, A., and Garcia-Molina, H. 2010. Tagging human knowledge. In *Proceedings of the Third ACM international Conference on Web Search and Data Mining* (New York, New York, USA, February 04 - 06, 2010). WSDM '10. ACM, New York, NY, 51-60. DOI= http://doi.acm.org/10.1145/1718487.1718495

Körner, C., Benz, D., Hotho, A., Strohmaier, M., and Stumme, G. 2010. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th international Conference on World Wide Web* (Raleigh, North Carolina, USA, April 26 - 30, 2010). WWW '10. ACM, New York, NY, 521-530. DOI= http://doi.acm.org/10.1145/1772690.1772744

Körner, C., Kern, R, Grahsl, H.P. and Strohmaier, M. 2010. Of categorizers and describers: an evaluation of quantitative measures for tagging motivation. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (HT '10). ACM, New York, NY, USA, 157-166. DOI= http://doi.acm.org/10.1145/1810617.1810645

Kleinberg, J. M. 2000a. "Navigation in a small world," *Nature*, vol. 406, no. 6798, August 2000.

Kleinberg, J. 2000b. The small-world phenomenon: an algorithm perspective. In *Proceedings of the Thirty-Second Annual ACM Symposium on theory of Computing* (Portland, Oregon, United States, May 21 - 23, 2000). STOC '00. ACM, New York, NY, 163-170. DOI= http://doi.acm.org/10.1145/335305.335325

Kleinberg, J. M. 2001. Small-World Phenomena and the Dynamics of Information. In *Advances in Neural Information Processing Systems* (NIPS) 14, 2001.

Landis, J.R. and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159-174, 1977.

Levine, N. 1965. Strongly connected sets in topology. In *The American Mathematical Monthly*, Vol. 72, No. 10 (Dec., 1965), pages 1098–1101.

Mehta, A., Saberi, A., Vazirani, U., and Vazirani, V. 2007. AdWords and generalized online matching. *J. ACM* 54, 5 (Oct. 2007), 22. DOI= http://doi.acm.org/10.1145/1284320.1284321

Millen, D. and Feinberg, J. 2006. Using social tagging to improve social navigation. In *Workshop on the Social Navigation and Community Based Adaptation Technologies*, Dublin, Ireland, 2006.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Nov, O. and Ye, C. 2010. Why do people tag?: motivations for photo tagging. *Commun. ACM* 53, 7 (Jul. 2010), 128-131. DOI= http://doi.acm.org/10.1145/1785414.1785450

Sinclair, J. and Cardew-Hall, M. 2008. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34–15, 2008.

Strohmaier, M., Helic, D., Benz, D., Körner, C. and Kern, R. Evaluation of Folksonomy Induction Algorithms. *ACM Transactions on Intelligent Systems and Technology*, accepted for publication in 2012. URL= http://kmi.tugraz.at/staff/markus/documents/2012_ACM-TIST_Folksonomy_Evaluation.pdf (last visited: 9. December 2011).

Trattner, C. 2011. Improving the Navigability of Tagging Systems with Hierarchically Constructed Resource Lists and Tag Trails, *Journal of Computing and Information Technology*, 2011 (to be published).

Trattner, C., Hasani-Mavriqi, I., Helic, D., and Leitner, H. 2010a. The Austrian way of Wiki(pedia)!: development of a structured Wiki-based encyclopedia within a local Austrian context. In *Proceedings of the 6th international Symposium on Wikis and Open Collaboration* (Gdansk, Poland, July 07 - 09, 2010). WikiSym '10. ACM, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/1832772.1832785

Trattner, C., Helic, D. and Strohmaier, M. 2010b. Improving Navigability of Hierarchically-Structured Encyclopedias through Effective Tag Cloud Construction. In *Proceedings of 10th International Conference on Knowledge Management and Knowledge Technologies*, I-KNOW 2010, Graz, Austria, 1-3 September, 2010.

Trattner, C., Helic, D. and Maglajlic, S. 2010c. Enriching Tagging Systems with Google Query Tags, In *Proceedings of 32nd International Conference on Information Technology Interfaces* - ITI2010, IEEE, Cavtat / Dubrovnik, Croatia, 205 - 210, 2010.

Trattner, C. and Helic, D. 2010. Linking Related Documents: Combining Tag Clouds and Search Queries, In *Proceedings of 10th International Conference on Web Engineering* - ICWE 2010, Vienna, Austria, 486 - 489, 2010.

Trattner, C. and Helic, D. 2009. In *Proceedings of IADIS International Conference WWW/Internet 2009* (2009), IADIS International Conference on WWW/Internet, Rom, 76 - 83, 2009.

Wikimedia 2010. Wikimedia Visitor Log Analysis Report - Requests by origin, URL= http://stats.wikimedia.org/wikimedia/squids/SquidReportOrigins.htm (last visited: 12. July 2010).