Journal of Universal Computer Science, vol. 17, no. 4 (2011), 565-582
submitted: 30/10/10, accepted: 15/2/11, appeared: 28/2/11

On the Construction of Efficiently Navigable Tag Clouds Using Knowledge from Structured Web Content

Christoph Trattner
(KMI and IICM, Graz University of Technology, Graz, Austria
ctrattner@iicm.edu)

Denis Helic
(KMI, Graz University of Technology, Graz, Austria
dhelic@tugraz.at)

Markus Strohmaier
(KMI, Graz University of Technology, Graz, Austria
markus.strohmaier@tugraz.at)

Abstract: In this paper we present an approach to improving navigability of a hierarchically structured Web content. The approach is based on an integration of a tagging module and adoption of tag clouds as a navigational aid for such content. The main idea of this approach is to apply tagging for the purpose of a better highlighting of cross–references between information items across the hierarchy. Although in principle tag clouds have the potential to support efficient navigation in tagging systems, recent research identified a number of limitations. In particular, applying tag clouds within pragmatic limits of a typical user interface leads to poor navigational performance as tag clouds are vulnerable to a so-called pagination effect. In this paper, a solution to the pagination problem is discussed, implemented as a part of an Austrian online encyclopedia called Austria-Forum, and analyzed. In addition, a simulation-based evaluation of the new algorithm has been conducted. The first evaluation results are quite promising, as the efficient navigational properties are restored.

Key Words: Tagging, tags, tag clouds, algorithm, tag cloud algorithm, navigation, navigability, online encyclopedia
Category: H.4

1 Introduction

An example of a semi-structured website is Austria-Forum[1]. Basically, Austria-Forum is a collection of several hierarchically structured Austrian encyclopedias that contain information about biographies, post stamps, coins, or the Austrian Universal Encyclopedia AEIOU[2]. Austria-Forum is a Wiki based system, whose articles within a single encyclopedia are hierarchically structured. Thus, Austria-Forum is also called a structured Wiki [Trattner et al. 2010]. Currently, as of 1^{st} of October 2010 the system provides over 130,000 information items to the user.

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

Due to the hierarchical structure and the rapid growth of the system over the past few months, links between articles in different encyclopedias are sparse even though they might be related to each other. For example, there are several "Mozart" stamps in the Stamps encyclopedia. However, none of these articles has links to the "Mozart" biography, or "Mozart" coins because the articles are created and managed independently.

To tackle the problem of poor connectivity, a simple tagging mechanism was introduced to Austria-Forum [Trattner and Helic 2009]. In tagging systems people use free-form vocabulary [Hammond et al. 2005] to annotate resources with "tags" [Wu et al. 2006, Marlow et al. 2006, Us Saaed 2008]. This is either done for semantic reasons (e.g. to enrich information items with metadata), conversational (e.g. for social signaling) [Ames and Naaman 2007] or for organizational reasons (e.g. to categorize information items) [Körner et al. 2010]. Regardless of "why people tag" [Strohmaier et al. 2010, Nov and Ye 2010, Strohmaier 2008, Körner et al. 2010a], tags can be visualized in so-called "tag clouds". A tag cloud [Ames and Naaman 2007] is a selection of tags related to a particular resource. Upon clicking on a tag, a list of resources tagged with that tag is presented to users leaving them with a possibility to easily navigate to related resources. The main idea of including a tag module into Austria-Forum can best be described via the previously mentioned "Mozart" example. Suppose that users tag "Mozart" stamps, "Mozart" coins, "Mozart" biography, or any other document dealing with "Mozart" with a common tag, e.g. "Amadeus". Whenever users navigate to any of these articles a tag cloud containing all assigned tags is presented by the system. Thus, users can now click on "Amadeus" tag and this presents a list of all other articles tagged by that tag. Consequently, all articles tagged with "Amadeus" are now linked to each other, in fact, they are cross-linked across the hierarchical structure. Due to such indirect linking capabilities, tag clouds are sometimes applied to provide navigational support in tagging systems (cf. systems such as Flickr, Delicious, or BibSonomy).

Recently, in a number of studies tag clouds have been investigated from user interface [Mesnage and Carman 2009, Sinclair and Cardew-Hall 2008] and networktheoretic perspectives [Neubauer and Obermayer 2009]. These studies agree with regard to some interesting findings, such as the observation that current tag cloud calculation algorithms need to be improved. The ability of tag clouds to support "efficient" navigation under the consideration of pragmatic user interface limits, such as tag cloud size and pagination, is very poor [Helic et al. 2010]. In particular, the pagination effect causes the fragmentation of the network destroying the connected component and thus leaving a majority of resources unreachable.

In this paper, we present an approach to constructing tag clouds that support efficient navigation. This new algorithm is based on the idea of hierarchical network models that are known to be efficiently navigable [Kleinberg 2001]. The algorithm has been implemented in Austria-Forum as a general tool for improving connectivity and navigability of the system as a whole.

The paper is structured as follows: Section 2 presents a model for tag cloud based navigation. Section 3 discusses the problems of tag cloud based navigation and current tag cloud construction algorithms. Section 4 presents the idea of a new and optimized tag cloud calculation algorithm based on the ideas of a hierarchical network model within an online encyclopedia system called Austria-Forum. Section 5 provides an analysis of the potentials and limitation of this new approach. Section 7 gives some insights to related work in this field. Finally, Section 7 concludes the paper and provides an outlook for the future work in this area.

2 Model of Tag Cloud Navigation

In this paper, the tagging data is modeled as a pair of the form (r, t), where r is a resource from the set of all resources R, and t is a tag of all tags T. Here, we do not take into account users as we concentrate only on links between resources imposed by tags assigned to those resources. The main navigational aid in a tagging system is a tag cloud and we denote it with TC. Formally, a tag cloud TC is a particular selection of tags from the tag set.

Due to user interface restrictions the number of tags within a tag cloud is usually limited to an upper bound. To model this situation we additionally introduce a factor n as a maximum number of tags in a tag cloud.

Usually, the most popular tags are assigned to a large number of resources – hundreds or even thousands of resources. When a user clicks on such a tag, tagging systems present a long paginated list of tagged resources. In most cases, 10–100 resources are presented to the users at once (see e.g. Delicious or Bibsonomy). To model these user interface limitation – that we refer to as the pagination from here on – we introduce a factor k that k-limits the resource list of tags within a tag cloud TC.

Finally, let us model the navigation process in a tagging system. Navigation in a tagging system might start from a home page where a system-global tag cloud is presented. Typically, tags with the highest global frequency are selected for inclusion in a tag cloud. Upon clicking on a particular tag a k-limited list of resources is shown. Once the user has selected a specific resource, the system transfers the user to the selected resource and presents a resource-specific tag cloud TC_r. The tags in such a resource-specific tag are selected according to the highest local frequency. In the next step, by selecting a tag from a given resource-specific tag cloud, the system again presents a paginated list of resources and the user might continue the navigation process in the same manner as before (see Figure 1).

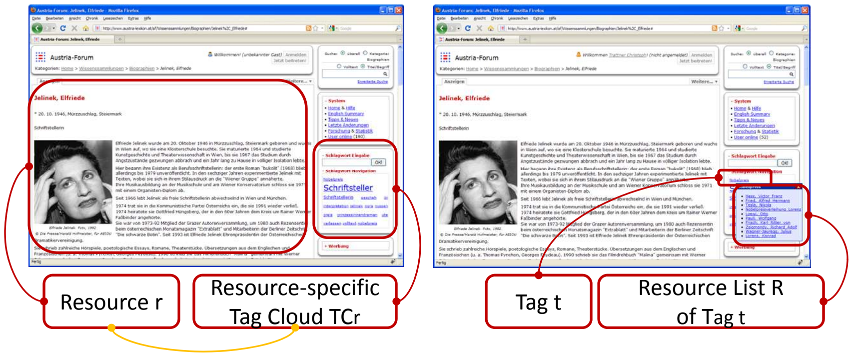


Figure 1: Resource specific tag cloud TC_r and k-limited resource list R of tag t within Austria-Forum

3 Problems of Tag Cloud Navigation

Resource-specific tag clouds are a simple way to connect resources within a tagging system, i.e. in a typical tagging system one can find nearly 99% of the resources interlinked with each other within a tag cloud network [Helic et al. 2010]. However, this simple approach to building tag clouds exhibits certain problems. In particular, resource-specific tag clouds are vulnerable to a so-called pagination effect [Helic et al. 2010]. In other words, by k-limiting the resource list of a given tag (with typical pagination values such as 5, 10, or 20), the connectivity of the tag cloud network collapses drastically. Practically, this leads to a situation where the tag cloud network consists of isolated network clusters (components) that are not linked to each other anymore. In other words, the users cannot reach one network fragment from another network fragment by navigating resource-specific tag clouds. One simple solution to this problem is to select resource for inclusion in a k-limited resource list uniformly at random [Helic et al. 2010]. For example, whenever the user clicks on a given tag in the tag cloud the system randomly selects k resources and presents them to the user. This leads to situation, that not always the same links are selected which leads to the situation that isolated network clusters are created [Helic et al. 2010]. As [Bollobás and Chung 1988, Helic et al. 2010] have shown this approach produces a random network that is, even for small values of k, completely connected.

3.1 Navigable vs. Efficiently Navigable Tag Cloud Networks

Another interesting issue in that context is the question if such randomly generated networks are also navigable. From a network-theoretic point of view

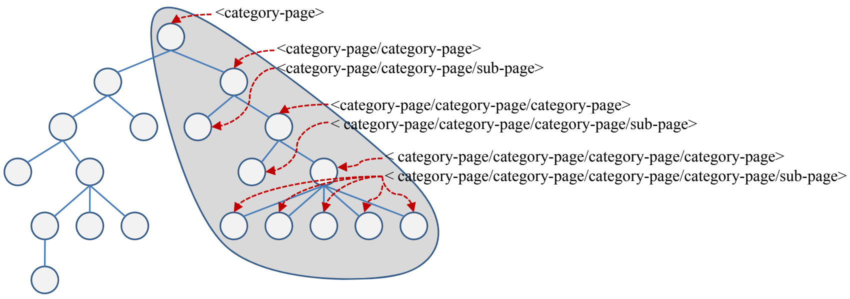


Figure 2: Hierarchical structure and URL addressing schema within Austria-Forum.

Kleinberg [Kleinberg 2000a, Kleinberg 2000b, Kleinberg 2001] showed that a navigable network can be formally defined as a network with a low diameter [Newman 2003] bounded by log(N), where N is the number of nodes in the network, and an existing giant component, i.e. a strongly connected component containing almost all nodes. Additionally, Kleinberg defined an "efficiently" navigable network as a network possessing certain structural properties so that it is possible to design efficient decentralized search algorithms (algorithms that only have local knowledge of the network) [Kleinberg 2000a, Kleinberg 2000b, Kleinberg 2001]. The delivery time (the expected number of steps to reach an arbitrary target node) of such algorithms is polylogarithmic or at most sub-linear in N. Put short, in [Kleinberg 2001] Kleinberg also showed that naive random networks algorithms form network structures which require linear search time ($O(N)$), i.e. in the worst case one has to visit all N nodes within a network to reach a certain destination node, i.e. such networks are not efficient navigable. However, in [Kleinberg 2001] Kleinberg also showed that hierarchical network models generate networks which are navigable in polynomial of $O(logN)$. Thus, we applied a hierarchical network model for tag cloud network generation in Austria-Forum to support efficient navigation.

4 Algorithm

4.1 Tag Clouds Hierarchy

We distinguish between two different types of nodes within Austria-Forum – category-page and sub-page nodes with sub-page nodes being hierarchy leaves (see Figure 2). Information items within Austria-Forum are hierarchically structured and addressable via a hierarchical URL schema.

The first component of the tag cloud generation algorithm in Austria-Forum simply follows the hierarchical data organization and constructs hierarchically organized tag clouds. The idea of this component is to provide more links between articles in one and the same category and to shorten the paths between category-pages and sub-pages. Thus, in order to generate a tag cloud for a particular category-page, the tags of all sub-categories and all sub-pages are aggregated recursively [Trattner and Helic 2009]. On the other hand, in order to generate a tag cloud for a particular sub-page, the resource-specific tag cloud calculation pattern is applied. The hierarchical tag cloud generation algorithm is shown in Algorithm 1 with t_f representing the local tag frequency.

Algorithm 1 Tag Cloud Calculation Algorithm

getTagCloud: url, n
if (url is category-page) **then**
  $TC_r^n \leftarrow$ select top n tags sorted by t_f where r.url.startsWith(url)
else
  $TC_r^n \leftarrow$ select top n tags sorted by t_f
end if
return TC_r^n

4.2 Addressing the Pagination Problem

Hierarchical network models [Kleinberg 2001] are based on the idea that, in many settings, the nodes in a network can be organized in a hierarchy. The hierarchy can be represented as a b-ary tree and network nodes can be attached to the leaves of the tree. For each node v, we can create a link to all other nodes w with the probability p that decreases with $h(v, w)$ where h is the height of the least common ancestor of v and w in the tree. Networks generated by this model are "efficiently" navigable [Kleinberg 2001].

The main idea of applying such a hierarchical network model is to reuse the hierarchical organization schema of articles in Austria-Forum as the basis for generating the link probability distribution p as described before. The hierarchical network model as introduced by Kleinberg takes a complete, balanced tree of nodes to obtain the link distribution. However, such an optimal model is typically not obtainable since real-word networks (cf. Open Directory Project[3], Google Directory[4] or Yahoo! Directory[5]) form hierarchical structures which are

[3] `http://www.dmoz.org/`
[4] `http://directory.google.com/`
[5] `http://dir.yahoo.com/`

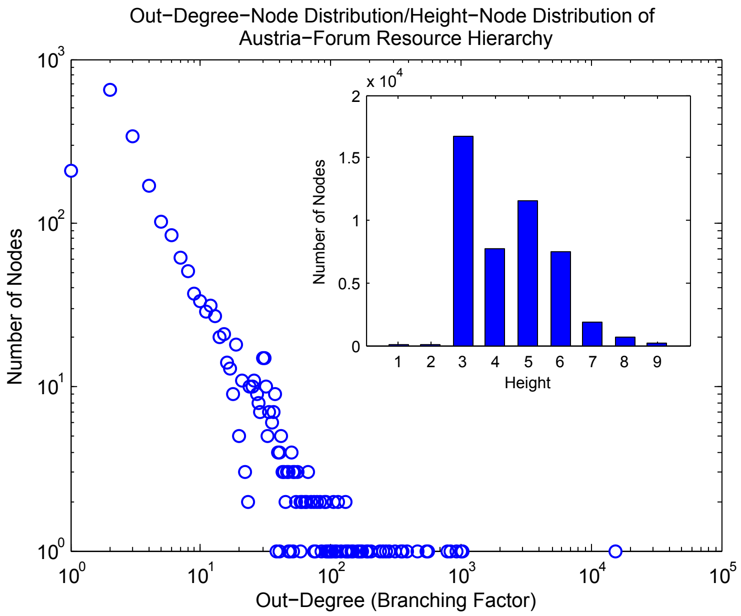


Figure 3: Out-degree distribution and node distribution of Austria-Forum resource hierarchy.

rarely complete nor balanced (cf. [Adamic and Adar]). For instance, in Austria-Forum the average branching factor is around 21 nodes ranging from 1 to over 14,000 nodes per category while the out-degree (branching factor) distribution of the hierarchy follows a power-law distribution (see Figure 3), which do not satisfy Kleinberg criteria such as a constant branching factor b. Thus, an algorithm implementing Kleinberg's model in our setting needs to work with intuitions and approximations.

The intuition which we followed with our algorithm is that the probability that an article is linked with other articles from the same category is higher than the probability that an article is linked with articles from other categories (cf. [Watts et al. 2002, Adamic and Adar, Kleinberg 2001]). Put short, this can be modeled by defining a link selection function that inter-links two nodes (articles) v, w according to a link probability function that is equal to $p = e^{-dist(v,w)}$ (cf. [Watts et al. 2002]) and a distance function that is calculated as $dist(v, w) = h_v + h_w - 2h(v, w) - 1$, where h_v, h_w are the heights of two nodes v, w in the hierarchy and where $h(v, w)$ is the height of the least common ancestor of the nodes v, w in the hierarchy (cf. [Adamic and Adar]). In Algorithm 2 the actual algorithm is presented.

Algorithm 2 Resource List Calculation Algorithm

For any given node $r(t) \in R$ in the resource hierarchy R, where t is the tag applied to this node, we find all other nodes $r_j(t) \in R$ and calculate distance $dist(r(t), r_j(t)) = h(r(t)) + h(r_j(t)) - 2h(r(t), r_j(t)) - 1$. For all found nodes $r_j(t) \in R$ we put $r_j(t)$ according to the distance $dist(r(t), r_j(t))$ into clusters $cl_x = [r_i, ..., r_j]$ and store these clusters into an array $rdist(i)_{r(t)} = [cl_{dist_1}, ..., cl_{dist_{n-1}}]$. Now, to select k links from the resource list, we generate k random numbers $i_k = 1$ ... sizeof($rdist(i)_{r(t)}$) with a probability density function $p = e^{-x}$ with $x = 1$ ... sizeof($rdist(i)_{r(t)}$) and select k clusters $cl_{i_k} \in rdist(i_k)_{r(t)}$ returning for each cluster just one element which is selected uniform at random.

5 Evaluation

To evaluate the presented algorithm, we developed a theoretical framework that integrates the following two modules:

- a **network-theoretic module** based on the Stanford Snap[6] library to calculate and evaluate network properties such as the size of the Largest Strongly Connected Component (LSCC) or the Effective Diameter (ED) [Helic et al. 2010] of the tag cloud network
- and a **searcher module** which implements a hierarchical decentralized searcher to simulate "efficient" tag cloud driven navigation.

5.1 Datasets

In the following section we describe the tag cloud networks which were generated and used for further evaluations. Basically, five different types of tag cloud networks were generated (see Table 1). They all vary in the way how the tag cloud and the resource list is calculated. Since one of our recent studies [Helic et al. 2010] showed that limiting the tag cloud to practically feasible sizes (e.g. 5, 10, or more) does not influence navigability, we set the tag cloud size in our experiments to a fixed value of $n = 30$ which is actually also the size of the tag clouds of Austria-Forum live system. Contrary, we varied the value k, i.e. the maximum number of links in the resource list, to $k = 15, 50, 100$, which is expected to impair navigability [Helic et al. 2010].

Dataset N (=Naive): This tag cloud network simulates the most common and naive tag cloud and resource list calculation approach used these days in tagging systems [Helic et al. 2010]. In other words, the tag cloud calculation algorithm in this model follows a simple TopN approach displaying the most

[6] `http://snap.stanford.edu/`

Name	TC-Algo.	R-Algo.	n	k	Nodes	Links
N_15	TopN	Chron.	30	15	11,716	246,031
N_50	TopN	Chron.	30	50	11,716	637,448
N_100	TopN	Chron.	30	100	11,716	1,039,741
HN_15	TopN-H	Chron.	30	15	12,044	292,692
HN_50	TopN-H	Chron.	30	50	12,044	753,482
HN_100	TopN-H	Chron.	30	100	12,044	1,242,580
R_15	TopN	Rand.	30	15	11,716	254,004
R_50	TopN	Rand.	30	50	11,716	648,937
R_100	TopN	Rand.	30	100	11,716	1,050,708
HR_15	TopN-H	Rand.	30	15	12,044	308,183
HR_50	TopN-H	Rand.	30	50	12,044	777,929
HR_100	TopN-H	Rand.	30	100	12,044	1,265,023
HH_15	TopN-H	Hier.	30	15	12,044	286,513
HH_50	TopN-H	Hier.	30	50	12,044	727,252
HH_100	TopN-H	Hier.	30	100	12,044	1,199,263

TC-Algo. = Tag Cloud Calculation Algorithm, R-Algo. = Resource List Calculation Algorithm, TopN-H = TopN Hierarchically, Chron. = Chronologically Sorted, Rand. = Randomly Sorted, Hier. = Hierarchically Sorted.

Table 1: Tag cloud network statistics: Number of nodes and links.

frequent n tags in the tag cloud while the resource list calculation algorithm sorts the resources descending chronological order and selecting the k most top resources.

Dataset HN (=Hierarchical Naive): This tag cloud network is generated using the hierarchical tag cloud calculation algorithm introduced in Algorithm 1. The resource list is calculated sorting the resources (links) chronologically in descending order and selecting the k most top resources.

Dataset R (=Random): This tag cloud network using a naive TopN algorithm (cf. Dataset G) for tag cloud calculations displaying the most frequent n tags in the tag clouds. The resource list is generated selecting k resources uniform at random.

Dataset HR (=Hierarchical Random): This tag cloud network is generated using the hierarchical tag cloud algorithm introduced in Algorithm 1. The resource list is calculated selecting k resources uniform at random.

Dataset HH (=Hierarchical Hierarchical): This tag cloud network is generated using the hierarchical tag cloud algorithm introduced in Algorithm 1

Name	TC-Algo.	R-Algo.	n	k	LSCC	ED	NAV
N_15	TopN	Chron.	30	15	0.567002	5.99404	unnav.
N_50	TopN	Chron.	30	50	0.761011	5.39847	unnav.
N_100	TopN	Chron.	30	100	0.863008	5.93894	unnav.
HN_15	TopN-H	Chron.	30	15	0.566008	3.47673	unnav.
HN_50	TopN-H	Chron.	30	50	0.755314	2.93258	unnav.
HN_100	TopN-H	Chron.	30	100	0.856941	2.90164	unnav.
R_15	TopN	Rand.	30	15	0.949983	5.93975	nav.
R_50	TopN	Rand.	30	50	0.949983	5.03066	nav.
R_100	TopN	Rand.	30	100	0.949983	5.43866	nav.
HR_15	TopN-H	Rand.	30	15	0.968034	3.73302	nav.
HR_50	TopN-H	Rand.	30	50	0.968034	3.17498	nav.
HR_100	TopN-H	Rand.	30	100	0.968034	2.90565	nav.
HH_15	TopN-H	Hier.	30	15	0.968034	3.46743	nav.
HH_50	TopN-H	Hier.	30	50	0.968034	2.92611	nav.
HH_100	TopN-H	Hier.	30	100	0.968034	2.92633	nav.

TC-Algo. = Tag Cloud Calculation Algorithm, R-Algo. = Resource List Calculation Algorithm, Chron. = Chronologically Sorted, Rand. = Randomly Sorted, Hier. = Hierarchically Sorted, LSCC = Largest Strongly Connected Component, ED = Effective Diameter, NAV = Navigability, TopN-H = TopN Hierarchically Calculated, unnav. = unnavigable, nav. = navibale

Table 2: Tag cloud network dataset statistics: Largest Strongly Connected Component, Efficient Diameter and Navigability.

and the hierarchical resource list algorithm introduced in Algorithm 2.

5.2 Evaluating Navigability

In order to evaluate whether the generated tag cloud networks are navigable or not, the size of the largest strongly connected component (LSCC) and the effective diameter (ED) was calculated. As already defined before (see Section 3.1), we consider navigable networks to be networks that have a low diameter bounded logarithmically and a giant component. As shown in Table 2, naive constructed tag cloud networks (N_15 − N_100 and HN_15 − HN_100) are formally seen not navigable. This is the case, since these types of networks do not have a giant component containing nearly almost all nodes of the network. Contrary, all other networks form navigable network structures, i.e. they contain a giant component and an effective diameter that is bounded logarithmically. Note, networks built on such a hierarchical approach generate networks that have a lower diameter

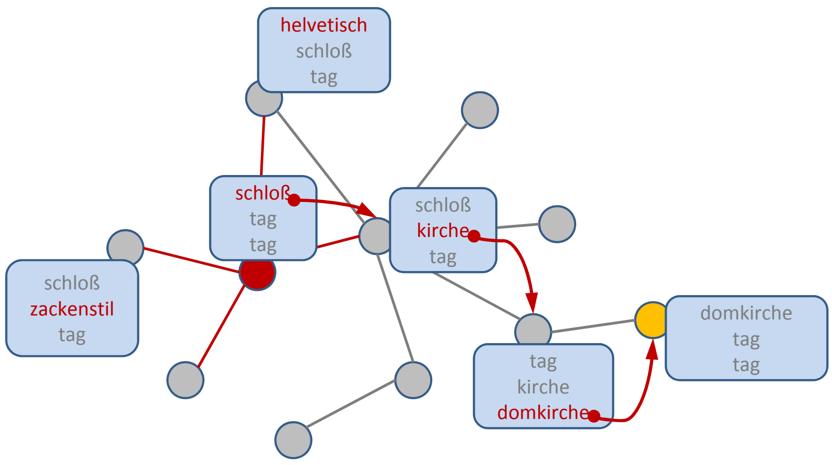


Figure 4: Shows an example of a resource-specific tag cloud network in Austria-Forum and a search through it.

than networks implementing a general TopN tag cloud calculation approach. This is not surprising, since such networks generate more long range links from category-pages to sub-pages, i.e. they shorten the paths to reach the sub-pages in the system.

5.3 Evaluating Efficiency

In order to evaluate the efficiency of our new approach, a hierarchical decentralized searcher was developed to simulate "efficient" tag cloud driven navigation. The searcher is basically an adoption of the work made by [Adamic and Adar] which uses background knowledge from the underlaying resource network structure to navigate the tag cloud network.

To model tag cloud based navigation, we define the tag cloud network as a bipartite hypergraph of the form $V = R \cup T$ [Helic et al. 2010], where R is the set of resources and T the set of tags. Since the resource lists are limited to a certain value k which forces the tag cloud network to collapse into a directed unipartite tag-resource network (with resource specific tags), we developed a searcher that walks along the underlying projected directed resource-resource network.

In Algorithm 3, the actual searcher algorithm is presented. In words, the algorithms works as follows:

To find a certain target resource w (e.g. tagged as "domkirche") from a certain start node v (e.g. tagged as "schloß") within the network (see Figure 4), the searcher first selects all adjacent nodes v_i for the start node and then

Algorithm 3 Hierarchical Decentralized Searcher (cf. [Adamic and Adar])

Searcher: resource-resource graph G, resource-hierarchy T, start node v, target node w
while v != w **do**
  $v_i \leftarrow$ get all adjacent nodes $\in G$ from v
  // finds closest node according to $dist = dist_{min}$
  // where $dist(v_i, w) = h(v_i) + h(w) - 2h(v_i, w) - 1$
  v $\leftarrow$ findClosestNode (v_i, T)
end while

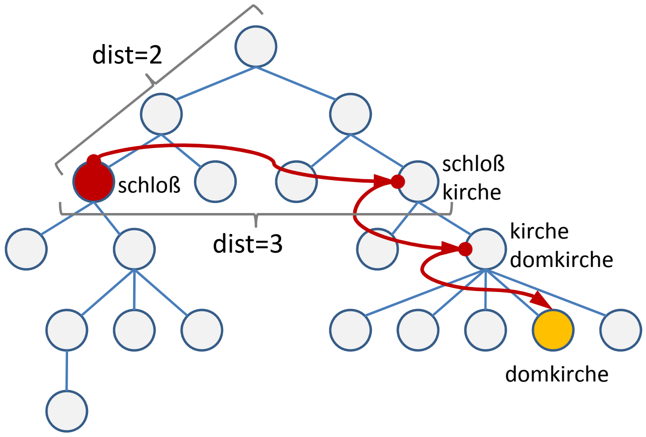


Figure 5: Shows an example of the Austria-Forum resource - taxonomy and a sample of tags they have applied.

selects the node v from the network ("kirche") that has the shortest distance $dist(v_i, w) = h(v_i) + h(w) - 2h(v_i, w) - 1$ to w node in the resource taxonomy T, with $h(v_i), h(w)$ being the heights of the two nodes v_i, w in the hierarchy and with $h(v_i, w)$ being the height of the least common ancestor of the two nodes v_i, w in the hierarchy [Adamic and Adar]. In the next step, the adjacent nodes of v are again selected and the distances $dist(v_i, w)$ are calculated, while the node v with shortest distance is selected in the end. The process is continued until the target node w is reached.

In order to get statistically significant results, we simulated 100,000 search-requests starting randomly selected at a certain resource v_i and targeting at certain randomly selected resource w_i in the tag cloud network. Note, only search pairs v_i, w_i were considered for the simulations for which a path (v_i, w_i) was present in the network. The upper limit for a search was set to a value of max-

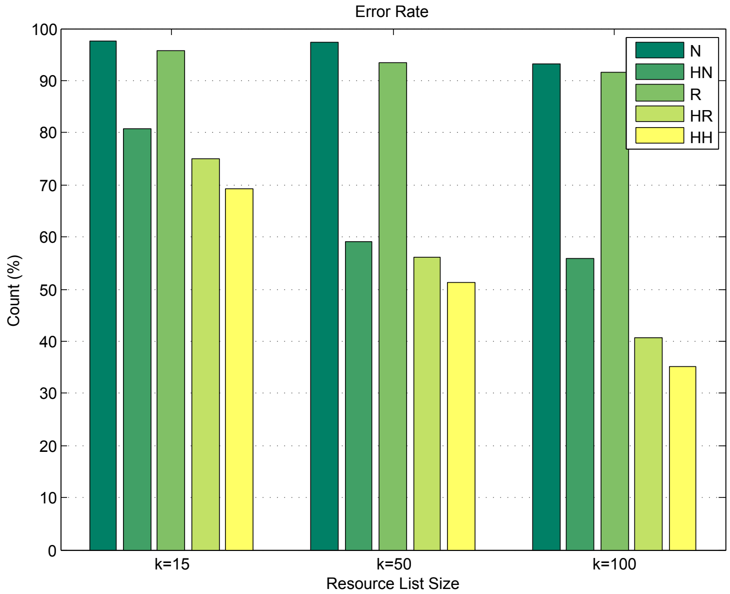
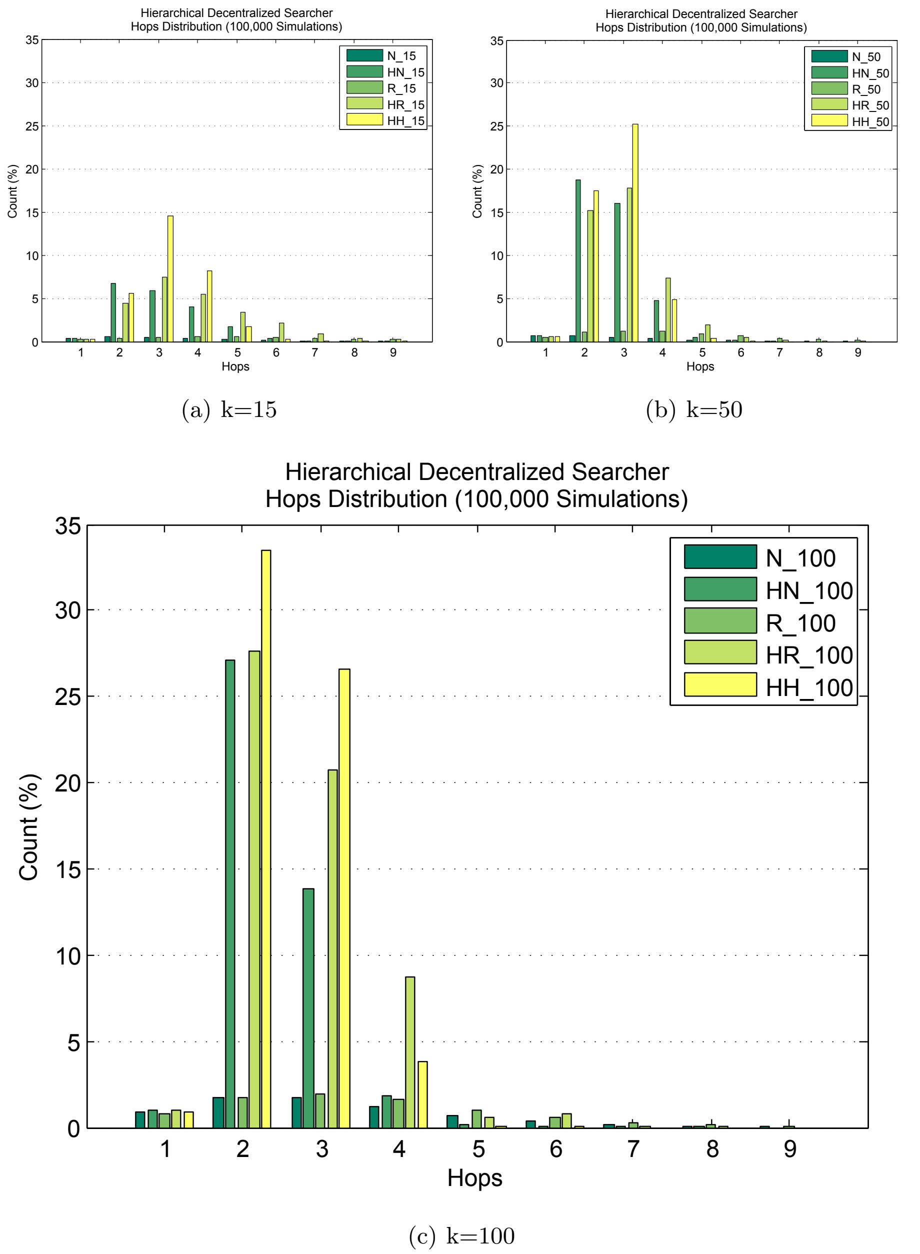


Figure 6: Error Rate for different types of networks. As expected, hierarchically generated networks (see network HG, HR and HH) perform significantly better than naive generated tag cloud networks (see network G and R).

imum 100 hops in the simulations, i.e. we canceled searches which took more than 100 hops to find a target node w_i. If the searcher was not able to find a path further in the tag cloud network, we canceled the search task as well. If a search task was being canceled, we did not reset the searcher to find a new path for the same search pair v_i, w_i.

As shown in Figure 6, flat and paginated tag cloud networks (labeled as network G and R in Figure 6) produce poor results for a naive hierarchical search algorithm in such networks. The reason for this behavior is the fact that the searcher frequently lands on a sink in the tag cloud network. This is the case since the resource has already been visited before or there is no link offered by the resource the searcher can follow anymore due to the low number of links (see Table 1) because of the pagination effect. "Expanded" networks implementing a hierarchical tag cloud algorithm (see Algorithm 1) perform even better in finding paths from resources v_i to a resources w_i in the network. For instance, for paginated resource lists and hierarchically calculated tag clouds (cf. network N and NH in Figure 6), the searcher fails only in 27% of all cases, while without hierarchically calculated tag clouds the error rate of the searcher in more than



(a) k=15 (b) k=50

(c) k=100

Figure 7: Hierarchical Decentralized Searcher hop-distributions for different values of $k = 15, 50, 100$ (size of the resource list).

89%. Furthermore, we can observe that hierarchically randomly generated networks are better navigable than all other investigated approaches (see network HR and HH in Figure 6). This is the case, since such networks provide more links between the resources of a tagging system (see Table 1) than "flat" general networks. Finally, we can investigate that networks adopting a hierarchical resource list calculation algorithm (see network HH in Figure 6) perform best by means of navigation. In case of Austria-Forum, this type of tag cloud network generates the lowest error rate and the fastest searchable network (see Figure 7) among all others.

6 Related Work

In related research on tagging systems, tag clouds have been characterized as a way to translate the emergent vocabulary of a folksonomy into social navigation tools [Sinclair and Cardew-Hall 2008, Dieberger 1997]. Social navigation itself represents a multi-dimensional concept, covering a range of different issues and ideas. A distinction between direct and indirect social navigation, for example, highlights whether navigational clues are provided by direct communication among users (e.g. via chat), or whether navigational clues are indirectly inferred from historical traces left by others [Millen and Feinberg 2006]. Based on this distinction, our work only focuses on indirect social navigation in the sense that it studies the effectiveness of traces ("tags") left by users in tagging systems. Other types of social navigation emphasise the need to show the presence of others users, to build trust among groups of users, or to encourage certain behaviour [Millen and Feinberg 2006].

Researchers have discussed the advantages and drawbacks of tag clouds, suggesting that tag clouds are a useful mechanism when users' search tasks are general and explorative (for example, learn about Web 2.0), while tag clouds provide little value for specific information-seeking tasks (for example, navigate to www.cnn.com) [Sinclair and Cardew-Hall 2008]. While the paper at hand focuses on network-theoretic aspects, cognitive aspects of navigation have been studied previously using, for example, SNIF-ACT [Fu and Pirolli] and social information foraging theory [Pirolli 2009]. Other work has studied the motivations of users for tagging [Körner et al. 2010], and how they influence emergent semantic (as opposed to navigational) structures. The navigational utility of single tags has been investigated [Chi and Mytkowicz] with somewhat disappointing results. With time the tags become harder and harder to use as they lose specificity and reference too many resources. Such tags are exactly those paginated tags where new pagination algorithms are needed.

Navigation models for tagging systems have been also discussed recently. In [Ramezani et al. 2009] authors describe a navigation framework for tagging systems. The authors apply the framework to analyze possible attacks on tagging systems. In principle, the framework identifies a navigation channels as any combination of the basic elements of a tagging system (users, tags, and resources). Thus, the specific combination which we investigated in this paper can be summarized as the resource-tag or tag-resource navigation channel.

Recent literature also discusses further algorithms for the construction of tag clouds. The ELSABer algorithm [Li et al. 2007] represents an example of such an effort aimed towards identifying hierarchical relationships between annotations to facilitate browsing. The work by [Aouiche et al. 2008] is another example, introducing entropy-based algorithms for the construction of interesting tag clouds. However, these algorithms have not found wide-spread adoption in current social tagging systems, and their usefulness to support navigation is largely unknown. In future work, it would be interesting to compare additional tag cloud construction algorithms with our approach. In addition, empirical studies of tagging systems have for example focused on comparing navigational characteristics of tag distributions to similar distributions produced by library terms [Heymann et al. 2010].

7 Conclusions and Future Work

The main contribution of this paper is the introduction of a novel, tag-based algorithm for interlinking resources in hierarchically-structured Web content. Based on a review of tag cloud limitations and an existing hierarchical algorithm for the construction of efficiently navigable networks, we discussed, implemented, and evaluated by simulation a new approach to tag cloud construction that improves the overall navigability of social tagging systems. While the arguments laid out in this paper are of a theoretical nature, we empirically tested the navigability of link structures produced by such an algorithm and confirmed the theoretical expectations by simulation. Finally, evaluating the usability and usefulness of the proposed algorithm with end users in an experimental setting would bring new insights into the potentials and limitations of the proposed approach.

Acknowledgments

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.

References

[Adamic and Adar] Adamic, L. and Adar, E.: How to search a social network, Social Networks, Volume 27, Issue 3, 187-203, 2005.

[Ames and Naaman 2007] Ames, M. and Naaman., M.: Why we tag: motivations for annotation in mobile and online media. In CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, New York, 2007.

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

[Bollobás and Chung 1988] Bollobás, B. and Chung, F. R. K.: The diameter of a cycle plus a random matching. In SIAM J. Discret. Math. 1(3), pp 328–333, 1988.

[Chi and Mytkowicz] Chi, E. H. and Mytkowicz, T.: Understanding the efficiency of social tagging systems using information theory,H T '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, ACM, NY, 81-88,2008.

[Dieberger 1997] Dieberger, A.: Supporting social navigation on the World Wide Web, Academic Press, Inc., Volume 46 (6), 805-825, Duluth, MN, USA, 1997.

[Fu and Pirolli] Fu, W.T. and Pirolli, P.: SNIF-ACT: a cognitive model of user navigation on the world wide web, Hum.-Comput. Interact., Volume 22 (4), 355-412, Hillsdale, NJ, USA, 2007.

[Hammond et al. 2005] Hammond, T., Hannay, T., Lund, B. and Scott, J.: Automatic construction and management of large open webs. Social Bookmarking Tools (I): A General Review, D-Lib Magazine, 11(4), 2005.

[Helic et al. 2010] Helic, D., Trattner, Ch., Strohmaier, M. and Andrews, K.: On the Navigability of Social Tagging Systems, The 2nd IEEE Conference on Social Computing, SocialCom2010, Minneapolis, Minnesota, USA, 2010.

[Heymann et al. 2010] Heymann, P., Paepcke, A. and Garcia-Molina, H.: Tagging Human Knowledge, Proceedings of the Third ACM International Conference on Web Search and Data Mining, ACM, NY, 51-61, 2010.

[Kleinberg 2000b] Kleinberg, J. M.: 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, 2000.

[Kleinberg 2000a] Kleinberg, J. M.: Navigation in a small world, Nature, vol. 406, no. 6798, August 2000.

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

[Körner et al. 2010] Körner, C., Benz, D., Hotho, A., Strohmaier, M. and Stumme, G.: Stop Thinking, Start Tagging: Tag Semantics Emerge From Collaborative Verbosity, 19th International World Wide Web Conference (WWW2010), ACM, Raleigh, NC, USA, April 26-30, 2010.

[Körner et al. 2010a] Körner, C., Kern, R., Grahsl, H.P, and Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation, Proceedings of the 21st ACM conference on Hypertext and Hypermedia, Toronto, Canada, 2010.

[Li et al. 2007] Li, R., Bao, S., Yu, Y., Fei, B. and Su, Z.: Towards effective browsing of large scale social annotations, Proceedings of the 16th international conference on World Wide Web, ACM, NY, 943-952, 2007.

[Marlow et al. 2006] Marlow, C., Naaman, M., Boyd, D. and Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read, In Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (Odense, Denmark, August 22 - 25, 2006), HYPERTEXT '06, ACM, New York, 2006.

[Mesnage and Carman 2009] Mesnage, C. S. and Carman., M. J.: Tag navigation. In SoSEA '09: Proceedings of the 2nd international workshop on Social software engineering and applications, ACM, New York, 29 - 32, 2009.

[Millen and Feinberg 2006] Millen, D.R. and Feinberg, J.:Using social tagging to improve social navigation, Workshop on the Social Navigation and Community Based Adaptation Technologies, Citeseer, Dublin, Ireland, 2006.

[Neubauer and Obermayer 2009] Neubauer, N. and Obermayer, K.: Hyperincident connected components of tagging networks, In HT'09: Proceedings of the 20th ACM conference on Hypertext and hypermedia, ACM, New York, 229 - 238, 2009.

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

[Nov and Ye 2010] Nov, O. and Ye, C.: Why do people tag?: motivations for photo tagging, Commun. ACM 53, 7 (Jul. 2010), 128-131, 2010.

[Pirolli 2009] Pirolli, P.: An elementary social information foraging mode, Proceedings of the 27th international conference on Human factors in computing systems, 605-614, ACM, NY, 2009.

[Ramezani et al. 2009] Ramezani, M., Sandvig, J.J., Schimoler, T., Gemmell, J., Mobasher, B. and Burke, R.: Evaluating the Impact of Attacks in Collaborative Tagging Environments, International Conference on Computational Science and Engineering 2009, CSE '09, 136-143, 2009.

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

[Strohmaier et al. 2010] Strohmaier, M., Körner, C., and Kern, R.: Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems, 4th International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington, DC, USA, May 23-26, 2010.

[Strohmaier 2008] Strohmaier, M.: Purpose Tagging - Capturing User Intent to Assist Goal-Oriented Social Search, SSM'08 Workshop on Search in Social Media, in conjunction with CIKM'08, Napa Valley, USA, 2008.

[Trattner and Helic 2009] Trattner, C. and Helic, D.: Extending The Basic Tagging Model: Context Aware Tagging, In Proceedings of IADIS International Conference WWW/Internet 2009 (2009), IADIS International Conference on WWW/Internet, Rom, 76 - 83, 2009.

[Trattner et al. 2010] Trattner, C., Hasani, I., Helic, D. and Leitner, H.: The Austrian way of Wiki(pedia)! - Development of a Structured Wiki-based Encyclopedia within a Local Austrian Context, WikiSym 2010 - The 6th International Symposium on Wikis and Open Collaboration, ACM, Gdansk, Poland, 1-10, 2010.

[Us Saaed 2008] Us Saaed, A., Afzal, M.T., Latif, A., Stocker, A. and Tochtermann, K.: Does Tagging Indicate Knowledge Diffusion? An Exploratory Case Study, In Proc. of the ICCIT 08 - International Conference on Convergence and hybrid Information Technology, Busan, Korea, 2008.

[Watts et al. 2002] Watts, D.J., Dodds, P.S. and Newman, M.E.J.: Identity and search in social networks. Science, Volume 296, 1302-1305, 2002.

[Wu et al. 2006] Wu, H., Zubair, M., and Maly, K.: Harvesting social knowledge from folksonomies. In Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (Odense, Denmark, August 22 - 25, 2006), HYPERTEXT '06. ACM, New York, 111 - 114, 2006.