Improving Navigability of Hierarchically-Structured Encyclopedias through Effective Tag Cloud Construction

Christoph Trattner  
(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 encyclopedia. The approach is based on an integration of a tagging module and adoption of tag clouds as a navigational aid in such a system. 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 possible solution to the pagination problem is discussed. In addition, the paper presents a first implementation prototype developed within an Austrian online encyclopedia called Austria-Forum.

Key Words: Tagging, tags, tag clouds, navigation, navigability, online encyclopedia

Category: H.4

1 Introduction

Austria-Forum\textsuperscript{1} 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 [Trattner et al. 2010]. Currently, the system contains approximately 110,000 information items.

Austria-Forum can be seen as a collection of several hierarchically structured encyclopedias such as Biographies, Post Stamps, Coins, or Austrian Universal Encyclopedia. Articles from a single encyclopedia have a common source and

\footnotetext[1]{http://www.austria-lexikon.at}
are therefore neatly interlinked to each other. Links between articles from two different encyclopedias are sparse even though the articles 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, we introduced a simple built-in tagging system in Austria-Forum. In tagging systems people use free-form vocabulary to annotate resources with “tags” [Hammond et al. 2005, Wu et al. 2006, Marlow et al. 2006, Us Saeed 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]. Independent of “why people tag” [Strohmaier et al. 2010b, Strohmaier 2008], tags can be visualized in so-called “tag clouds” (cf. [Ames and Naaman 2007]). A tag cloud 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 typically applied as a navigational support in tagging systems (cf. systems such as Flickr, Delicious, or BibSonomy) under the assumption that they are useful for navigation. Recently, in a number of studies tag clouds have been investigated from user interface [Mesnage and Carman 2009, Sinclair and Cardew-Hall 2008] and network-theoretic perspective [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. In particular, 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 [Strohmaier et al. 2010a, Trattner et al. 2010].

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 generalized model for tag cloud based navigation. Section 3 discusses the problems of tag cloud based navigation and current tag cloud calculation algorithms. Section 4 presents the idea of a new and optimized tag cloud calculation algorithm based on the ideas of hierarchical network models within an online encyclopedia system called Austria-Forum. Finally, Section 5 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).

3 Problems of Tag Cloud Navigation

Resource-specific tag clouds are a simple way to connect many resources within a tagging system [Strohmaier et al. 2010a], i.e. in a typical tagging system one can
find nearly 99% of the resources interlinked with each other within a tag cloud network. However, this simple approach to building tag clouds has also certain issues. 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 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. For example, whenever the user clicks on a given tag the system randomly selects $k$ resources and presents them to the user. As [Bollobás and Chung 1988] have shown this approach produces a random network that is, even for small values of $k$, completely connected.

However, another important question is: Are such networks also “efficient” navigable? In other words, how useful and usable are tag clouds where the resource selection happens randomly? From the theoretical point of view, Kleinberg [Kleinberg 2000] argued that “efficiently” navigable networks are networks for which efficient decentralized search algorithms exist. Such algorithms can find a short path between a starting and a destination node in a polynomial of $O(\log N)$ ($N$ is the total number of nodes within a network). 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 the destination node. In [Kleinberg 2001], Kleinberg showed that a hierarchical network generation model forms networks which are navigable in
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 hierarchicallyorganized tag clouds. The idea of this component is to provide more links between articles in one and the same category. 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 (hereby, $t_f$ represent a local tag frequency).

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 might 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 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].
Algorithm 1 Tag cloud calculation algorithm

```
getTagCloud: url, n
if (url is category-page) then
    TC^r_n ← select top n tags sorted by t_r where r.url.startsWith(url)
else
    TC^r_n ← select top n tags sorted by t_r
end if
return TC^r_n
```

The main idea of applying such a hierarchical network model is to reuse the hierarchical organization of articles in *Austria-Forum* as the basis for generating the link probability distribution. To put it simply, 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. However, one needs only a few of such “long-range” links to other categories to obtain a connected and efficiently navigable network. Two examples of links generated by such a model are given in Figure 3.

![Figure 3: Hierarchical random selection algorithm](image)

The Algorithm 2 shows our first approximation of such a network generation algorithm. In principle, for each paginated tag a different resource list is presented to users depending on the user current context, i.e. depending on the resource where the user clicked on that given tag. The resources are selected randomly following this simple heuristic: $k - j$ resources are selected from the category of the current resource and $j$ resources are selected from other categories ($j < k/2$).

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 a strong simplification because hierarchies are rarely complete.
Algorithm 2 Resource list calculation algorithm

```plaintext
def getResourceList: url, t, k, j
    category-page ← categoryOf(url)
    R_{category} ← select resources r with t where r.url.startsWith(category-page.url)
    R_{global} ← select resources r with t where NOT r.url.startsWith(category-page.url)
    R_{category} ← select k - j resources at random from R_{category}
    R_{global} ← select j resources at random from R_{global}
    R_k ← R_{category} \cup R_{global}^j
    return R_k
```

or balanced. Algorithms implementing this network model need to work with heuristics and intuitions that approximate the optimal settings. The intuition which we followed with our algorithm was that linkage probability is based on the hierarchy distance between articles. To evaluate such an intuitive assumption as well as to estimate the algorithm parameters, e.g. \( j \) a detailed empirical analysis would be needed. This is, however, out of the scope of this paper and is left for the future work.

5 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 online encyclopedias. Based on a review of tag cloud limitations and an existing hierarchical algorithm for the construction of efficiently navigable networks, we sketch 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 leave the task of empirically testing the navigability of link structures produced by such an algorithm to future work. 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.

References


