Navigating One Million Tags

Jacek Gwizdka
Department of Library & Information Science, Rutgers University
4 Huntington Rd., New Brunswick, NJ 08901, USA
assist09@gwizdka.com

Philip Bakelaar
Department of Library & Information Science, Rutgers University
New Brunswick, NJ 08901, USA
bakelaar@rci.rutgers.edu

Introduction
Web 2.0 users engage in collaborative activities. Many 2.0 websites allow their users to associate words (a.k.a. tags) with the users’ own and with other users’ content. The result is an additional layer with descriptions of web resources. The created associations between tag sets and web resources are used to facilitate social navigation, individual resource re-finding, defining communities of interest, ownership identification, identification of resource qualities, and understanding the content of web resources (Golder & Huberman, 2006), to name just a few. In general, tags that were created by others can be viewed as serving two main purposes, 1) promoting understanding of the associated set of web resources, and 2) facilitating navigation of the collection of web resources. At each step of the navigation process, a set of tag serves to describe the set of associated resources, but it also serves as starting points for further navigation.

The availability of tagged web resources to all users makes social bookmarking websites a potentially attractive source of metadata. It is not surprising that such websites were studied as information retrieval devices. For example, Morrison compared folksonomies, internet directories and search engines as information retrieval mechanisms. He found that folksonomies had an overall lowest precision, but in specific cases their precision was comparable to that of search engines. Folksonomies and internet directories had similar levels recall, and both were lower than for search engines. Several researchers suggested combining search engine results with social bookmarking results. Yanbe, Jatowt, Nakamura & Tanaka (2007) discussed enhancing search engine algorithms by adding popularity based on the tagged URL data. Gwizdka & Cole (2007) found that Delicious search results had little overlap with the top pages of Google results, and thus results from both types of web systems complement each other. These studies provide evidence that social bookmarking data can provide information that is useful in Web search. Given the usefulness of the tag-resource associations, can users take advantage of it? We discuss why the current systems do not support users well in navigation and exploration tasks. We then suggest that different ways of presenting tags to users are needed, if the users are to take advantage of the available tag “descriptors”. This poster presents our initial approach to supporting users in navigating tag spaces. Our ideas are embodied in a prototype user interface that we describe below (see figures in Appendix).

Navigation of Tag Spaces
Tag space is a collection of web resources (such as web pages, images, videos) that are “described” by tags, that is, that have tags associated with them. Tags are frequently displayed in a form of a list or a tag cloud (Sinclair et al., 2008). A tag cloud may be shown next to a list of resources to represent all (or most frequent) words associated by a website user with these resources. A tag cloud may be shown next to a user name to represent tags most frequently employed by this user. In both cases, the words in the tag cloud typically serve as links that can be followed to obtain a new view of the associations among the users-resources-tags. For example, a user can follow a tag-link to obtain a list of resources associated with this new tag, or a user can follow a username-link to explore resources tagged by another user. This kind of navigation has been called pivot browsing. Millen et al. (2006) describe it as a lightweight navigation mechanism that enables users to reorient their view by clicking on tags or user names. The tag cloud content is switched at each navigation step to a new one that contains tags associated with the new set of resources. The new set of tags describes these new resources and provides new points for further exploration. However, at the same the display is completely re-oriented around the tag or link clicked. Furthermore, the history of navigating tag-space is not preserved and no information is provided about relationships between tags (and tag clouds). This can be considered potentially confusing and disorienting to the user. Yet the authors have found this approach to be quite common across many sites that employ tags and tag clouds. Navigating the user-resource-tag space by pivot browsing does not preserve the history of tag-space navigation. It currently works upon the assumption that each step in navigation is separate from other steps
(previous and future), and, if there is any dependency or relationship, a user needs to remember the history of her exploration. Yet the search and navigation process is not a series of individual steps but rather an iterative process (Kuhlthau, 1991). In this project, we focus on providing history of navigating the associations between tags and the web resources.

Our Approach

The concept of our interface, called Tag Trails, was influenced by the ideas from web navigation and information retrieval. In web navigation, breadcrumbs are a technique, which aids users in locating their current position and in navigating hierarchical web sites (Aery, 2007). While tag-spaces are not hierarchical, the general idea of leaving a trail of “crumbs” seems applicable. In information retrieval, clickable words can be considered as equivalent to queries that are sent to a system (Golovchinsky, 1997). Information retrieval research has highlighted the usefulness of preserving query history (Kuhlthau, 1991). Hence presenting a history of clicked tags seems applicable. These ideas have shaped the development of our techniques that are aimed at preservation of history and context in browsing tag-described information spaces.

Tag Trails uses three techniques to preserve context and history. First, Tag Trails stores the entire recently seen clouds so that a cloud history could be presented (Figure 1). The history is shown by displaying multiple tag clouds. As a user clicks through various tag clouds, the most recent tag sets are kept and the highest-frequency tags are displayed in alphabetical order. The older the history cloud the darker its background color. We chose to maintain three “history clouds” in addition to the current cloud. With this base display, we are able to apply the two following techniques to show contextual relationships and highlight possible connections between the clouds, of which the user may not have been aware otherwise.

Second, the relationships between the tag clouds are shown by using colors. The background color of the title bar of each tag cloud is used to indicate when that tag appears in another cloud; when it does, its background color is the same as in the title bar. Highlighting all tags using this “color context” technique allows the user to partially determine the level of continuity or relationship between the two tag clouds.

Third, the similarities and differences between the current cloud and the history clouds are displayed. We pair-wise compare tag sets associated with the current cloud and each history cloud. We show the result visually by highlighting the different tags in a distinct color (red). Using a distinct font color helps to keep this aspect from visually dominating the cloud. In the prototype, this feature is user-controlled, and can be set to “Off”, “Highlight similarities”, or “Highlight differences”. Figure 1 shows the “highlight differences” mode, which we believe is the most useful. At one glance, the user can group either the black tags (similar) or the red tags (different) and determine which tags co-occur in the current cloud. We also provide a summary about the number of different tags and total tags in each history cloud in comparison with the current cloud. This information helps the user assess the extent of differences between the clouds.

Summary and Future Work

Social bookmarking systems have been considered as potentially useful in information search. Systems that currently support tag clouds and pivot-browsing often treat tags as discrete, independent entities. However, user-generated tags are part of a social and mental process in which relationships can be present. For instance, multiple tags may be entered at the same time, or common tags may be suggested prior to a user entering their own. We propose that tags should not be treated functionally as independent. The idea behind our prototype system Tag Trails is to incorporate these tag dependencies and relationships and make them explicit to the user, through techniques such as “cloud history”, “color context”, and “similarity/difference”. These techniques reveal the connections between tags as a user navigates through tag-space. We argue that preserving navigation and exploration history in tag-space is useful to a user and can be implemented in a way that preserves the light-weight character of pivot-navigation, and we propose to improve navigation in systems and websites that utilize tag clouds and pivot-browsing by providing a history of tag space navigation and context. Our prototype demonstrates several techniques that have a potential to aid user in inferring this higher-level information. While our particular combination of techniques and visual construction may be applicable to a specific kind of social tagging system, the ideas can be modified to suit similar user tasks in other systems. In the future, we plan to conduct a user study to determine the usefulness and effectiveness of our techniques on the user’s search and navigation process. We also plan to explore alternative visual representations, such as the heat map shown in Figure 2.
References
Appendix. Interface Prototypes.

Figure 1. Tag Trails interface consists of two main areas: the result list on the left and the tag clouds on the right. The interface uses “cloud history”, “color context” and “similarity/difference” to show the history and context of user navigation or search process.

Figure 2. Heatmap that represents top tag co-occurrences for the same tag trail as shown in Figure 1.

Data Set used in the prototypes. In our exploration of preserving and displaying the navigation history, we used data from CiteULike, a social bookmarking web site that is focused on scholarly work. The data was downloaded on July 14th 2008 and loaded to a MySQL database. For the experimental system, only the records from January 1 to July 14 2008 were used. The data was sanitized to reduce the number of tags not useful for navigation (e.g., numbers, words that clearly referred to a particular browser, such as firefox, etc.). Approximately 800,000 records remained. Each record represents a single tagging instance that was submitted by a single user and assigned to a single web resource. For example, if one user tagged a single resource with 10 tags, there would be 10 records entered into the database. The data set contains tags entered by approximately 9,100 unique users and describing 224,000 web resources.