Collaborative Filtering with del.icio.us

Rick Wash
School of Information
University of Michigan
rwash@umich.edu

Emilee Rader
School of Information
University of Michigan
ejrader@umich.edu

CHI 2006, April 22–27, 2006, Montreal, Canada.
ACM 1-xxxxxxxxxxxxxxxxxx.

Abstract

Popular website del.icio.us was designed with two main goals: organizing a collection of bookmarks stored online, and sharing and discovering new and interesting websites. To achieve this second goal, del.icio.us acts as a collaborative filtering system. It utilizes additional user-submitted keywords called 'tags' as metadata that cannot be easily gathered from other sources. It also promotes information flow through social networks with its inbox mechanism.

Keywords

Collaborative Filtering, Tagging, Web 2.0, Social Bookmarking, Folksonomy

ACM Classification Keywords

H.5.3: [Information Interfaces and Presentation] Group and Organizational Interfaces ---Collaborative Computing, Web-based Interaction; H.3.3: [Information Storage and Retrieval] Information Storage and Retrieval --- Information Filtering; H.3.5: [Information Storage and Retrieval] Online Information Services --- Web-based Services

Introduction

del.icio.us is a website for "social bookmarking" where users can store bookmarks along with descriptive keywords or "tags." When a user of del.icio.us logs in to their account and adds a bookmark, she also tags that bookmark with single words that she feels are somehow related to that web page. The tags are then publicly available – searching by a tag produces all of the sites ever tagged with that word. The six most common tags assigned to the travel website Travelocity.com by users of del.icio.us are: travel, airlines, tickets, shopping, airfare, and search. These tags illustrate various relationships tags can have with the items to which they refer – shopping and search are activities supported by Travelocity, whereas airfare and tickets are what users are shopping or searching for. The primary design goal for del.icio.us was to provide a place for the online storage and organization of bookmarks.

A secondary goal of del.icio.us was to facilitate sharing bookmarks with others and discovering new things. Users can "subscribe" to a tag, which notifies the user whenever someone else submits a bookmark with that tag by placing that bookmark in the user's 'inbox'. They can also subscribe to other users, causing all new bookmarks from that user to appear in the subscriber's inbox.

del.icio.us has drawn a lot of interest from Internet businesses. It is considered one of the premier Web 2.0 applications. Yahoo! recently negotiated a deal to purchase del.icio.us. And, Amazon (amazon.com) has recently introduced tagging into their website in the spirit of del.icio.us. It has also had a lot of attention from intellectuals recently. A number of authors [5,6] have written articles discussing the use of tagging as an organization scheme. The collective tagging of individual items with idiosyncratic terms even has a name: "Folksonomy." Golder and Huberman [3] have a forthcoming paper about the use of del.icio.us to organize information.

Most of the current work on del.icio.us focuses on the use of tags as a method of organizing information. While important, there is this secondary goal of sharing and discovering new bookmarks that has not been adequately studied. This is a form of information filtering, distinct from querying the existing database. [1] We believe that del.icio.us is a novel form of collaborative filter [4] for Internet websites. Users can learn about new websites through the use of the inbox, and can filter out the individually relevant sites by selectively subscribing to relevant tags and similar users.

Dataset

To study these properties, we collected a sample of sites that were bookmarked on del.icio.us. We collected the bookmarks listed on the "popular" and "recent" pages on a number of different days between August and November 2005. We collected the URL pages for each of these sites, which list everyone who bookmarked the site and the tags they applied. In total we had 500 sites in our dataset. We then downloaded all of the user pages and inbox pages for all of the users that had bookmarked any of those sites. These pages contain complete lists of tags the user has used (with frequency counts), and complete lists of subscription information for these users. All of these pages were recollected in late December to have a consistent snapshot in time of the users.

There were a total of 500 sites in our dataset. The average site was bookmarked by 166 users, and had on average 65 distinct tags that had been applied to it. The most popular site was the Google advanced search operators reference page, which was bookmarked by 3773 users. In total there were pages for 39,723 users that were collected and analyzed. The average user had 272 bookmarks, and had used 144 distinct tags on those bookmarks. Additionally, 10966 (27.6%) users had inboxes with one or more subscriptions. The average inbox user had 5.98 subscriptions. In total, there were 508,735 distinct tags in our dataset. The modal number of uses was 1, meaning most tags were only used by one user. The average tag was used by 11.24 users, and the most popular tag ('music') was used by 22130 users.

Collaborative Filtering with the Inbox

The del.icio.us inbox is the primary method for collaborative filtering. This inbox allows users to create subscriptions to other users and tags. Any new bookmarks that match these subscriptions will then be placed on the user's inbox page for viewing. There are three possible types of subscriptions. A user can subscribe to all of the bookmarks from another user (a 'user' subscription). He or she can subscribe to all bookmarks that have a given tag (a 'tag' subscription). Or the subscription can be of the form 'user/tag', which only matched bookmarks by the specified user where that user applied the given tag.

Bookmark Tags in the Inbox

One metric we looked for is the overlap between the set of tags a user has subscribed to, and the set of tags the user has used to bookmark sites. 81% (5839) of users who subscribe to at least one tag have some overlap between their tag subscriptions and their bookmark tags. Over half of these users have used all of the tags to which they subscribe when bookmarking sites. This suggests that users subscribe to topics (tags) that they are interested in. (We judge interest by the fact that they have bookmarked sites on that topic).

Where Do Tags Come From?

We believe that the user's choice of tags provides invaluable metadata for the system's collaborative filtering properties because the tags contain information that cannot be easily obtained elsewhere. 74% of the tags in our sample were used by only one user, indicating that there is a good bit of disagreement about which tags are useful. This indicates the presence of the vocabulary problem. [2] However, this also means that when looking at a site's list of tags, there is a lot of information contained in this list about the contents of the site.

To study this, we compared the tags that were applied to a site with the actual contents of the site webpage. For 335 sites (70%), at least one of the 3 most frequently used words in the webpage appeared as a tag. However, for 414 sites (84%), at least one of those three most frequently used words did not appear as a tag. On average, only 1.253 of the 3 most frequently used words in the source webpage appear as tags for that site. The average site has only 26% of its tags appearing in the webpage at all. We believe that this is evidence that the tags provide useful metadata that is not directly available in the webpage. We also think that this extra metadata is important in improving the quality of del.icio.us as a collaborative filter. This allows users to search and filter using words that do not appear in the target document, something normal Internet search engines are not very good at.

A good example of this is the tag 'productivity,' which is one of the top 100 most frequently used tags in our dataset. This tag is used by a good size community of users on del.icio.us to tag websites that provide useful information for improving productivity. It is particularly common to see it used on sites that advocate a methodology called 'Getting Things Done.' Most of these websites talk about creating to-do lists or techniques for preventing procrastination, but very few of them talk directly about productivity. The del.icio.us users, however, have made this connection, and have linked these sites together by this common theme. 'Productivity' is the seventh most popular tag to subscribe to in our sample.

The Social Network of del.icio.us

Up to this point we have mostly been discussing tags on del.icio.us. But what makes collaborative filters work is that clusters of users can be found with similar interests. del.icio.us takes advantage of similarity between users interests by allowing users to subscribe to some or all of the tags of another specific user with similar interests. In this context, the act of bookmarking a site by that user serves as an endorsement of the site. Subscribing to a user is stating that those endorsements are worth trusting.

This endorsement however comes virtually costlessly to the user. They don't need to actively provide information for the collaborative filter to use. They don't even need to be aware that others are trusting their bookmark decisions as endorsements of websites. All they need to do is set aside a link to the website as something they find important by bookmarking it. Sites that are set aside as important are likely to be interesting to others with similar interests.

Users with similar interests are linked through a user subscription: User A subscribes to User B. However, information flows in the opposite direction: when User B bookmarks a site, that information goes into User A's inbox. We used our data to build a social network of users where users are linked through this 'subscribes to' relation.

Using the inbox, information can spread through this network to groups of people with similar interests. However, some people are more influential than others in this information flow. We classified users into three categories. Users who are in the top 1% in number of subscriptions to other users, but aren't often subscribed to are classified as "Sinks," since they gather lots of information but it does not continue to spread through them. Conversely, users who are in the top 1% in the number of users who subscribe to them, but who don't have many subscriptions themselves, are classified as "Sources." These people are likely put new information into the network. Finally, users who are in the top 1% in both subscriptions and subscribers are classified as "Hubs", since lots of information is likely to flow through these users. Our sample has 116 Sources, 129 Sinks, and 39 Hubs. These three classes of user tend to have a very large number of bookmarks (2239 for Sources, 1810 for Sinks, and an amazing 2855 for Hubs) and distinct tags (600, 591, and 946 respectively) on average. These users seem to be very heavy users of del.icio.us.

Figure 1 shows the social network for one of the larger hubs in our dataset. This hub happens to be Joshua Schachter, the creator of delicious. As can be seen, Joshua is connected to many of the other influential people in our dataset, including 30 of the 38 other hubs.

Discussion and Future Work

We believe that del.icio.us enables an interesting form of collaborative filtering through its inbox mechanism. Our data leads us to conclude that the inbox provides relevant information to users who have chosen to use it, that tags provide interesting and useful metadata that is not easily available elsewhere, and that inboxes work because information flows across subscriptions, often through well-connected 'hubs', and provides 'endorsements' of sites from trusted users. Other collaborative filters could benefit from incorporating tags into the metadata they use to make recommendations, since tags provide unique information. It would not surprise us to learn that this is part of the reason behind Amazon's recent interest in tagging. Collaborative filters would also likely benefit from providing users an explicit subscription system to take advantage of this endorsement effect.

There is much that still needs to be studied about del.icio.us as a collaborative filter. A large portion of del.icio.us users have similar interests (tags like 'AJAX', 'linux', 'javascript', and 'css' are all in the top 20). It is unclear how well del.icio.us can handle a large group of users with very diverse interests. We hope our dataset can lend some insight into this.

One potential problem with del.icio.us's inbox is the free rider problem. A user can reap all of the benefits of the collaborative filter without providing any metadata to the system. This incentive problem can possibly lead to system failure if too many people try to 'free ride' on the efforts of others. However, del.icio.us has two simultaneous goals (organization of bookmarks is the other), and we think that it is the presence of this second goal that gives sufficient incentive for users to provide metadata to the system.

References

[1] Belkin, N. and Croft, W. Information Filtering and Information Retrieval: Two Sides of the Same Coin *Commun. ACM* 35, 12 (1992), 29-38.

[2] Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. The vocabulary problem in human-system communication. *Commun. ACM,* 30, 11 (1987), 964-971.

[3] Golder, S. and Huberman, B. The structure of collaborative tagging systems. To appear in *Journal of Information Science* (2006)

[4] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. GroupLens: An open architecture for collaborative filtering of netnews. *Proc. CSCW 1994*, ACM Press (1994), 175-186.

[5] Shirky, C. (2005). *Ontology is Overrated: Categories, Links and Tags.* http://www.shirky.com/writings/ontology_overrated.html

[6] Udell, J. (2005). *Managing Metadata*. InfoWorld, October 20, 2005. http://www.infoworld.com/article/05/10/20/43FEmetadata_1.htm

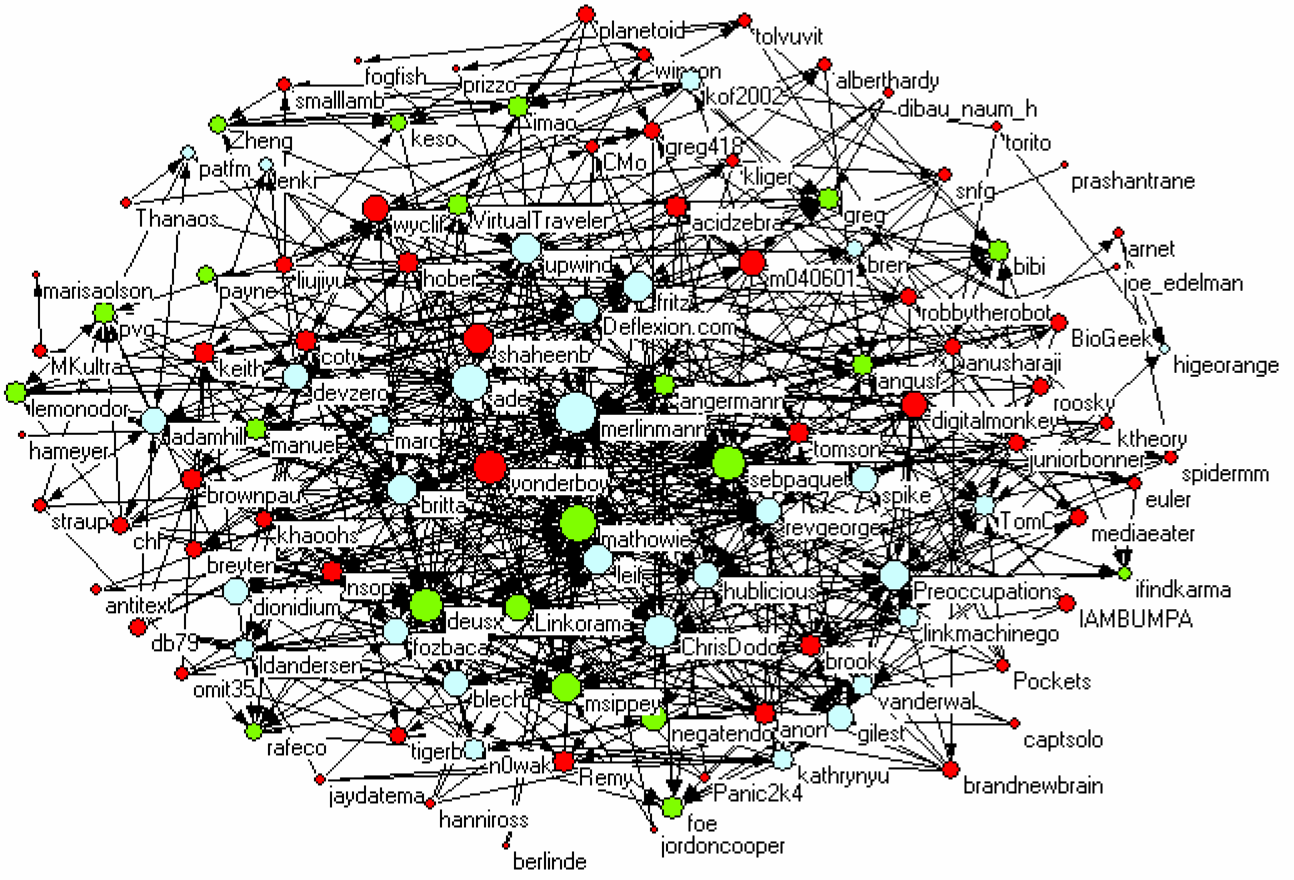

Figure 1: Social Network Diagram for Joshua Schachter, the creator of del.icio.us. This network is composed of users who either subscribe to Joshua, or to whom Joshua is subscribed. This network only includes users who are classified as Sources (green, mathowie), Sinks (red, yonderboy), or Hubs (blue, merlinmann). Arrows indicate a "subscribes to" relation. The size of a node represents the number of connections (degree) of the node. Joshua himself is not in this picture since every node would link to him, which would greatly complicate the diagram. With such a large network, Joshua is obviously a hub