From del.icio.us to x.qui.site: Recommendations in Social Tagging Sites

Sihem Amer-Yahia[†], Alban Galland[§,†], Julia Stoyanovich[‡], Cong Yu[†]
[†]Yahoo! Research, New York, NY; [§]INRIA Saclay, Paris, France; [‡]Columbia University, New York, NY
[†]{sihem,congyu}@yahoo-inc.com, [§]alban.galland@inria.fr, [‡]jds1@cs.columbia.edu

ABSTRACT

We present X.QUI.SITE, a scalable system for managing recommendations for social tagging sites like del.icio.us. seamlessly incorporates various user behaviors into the recommendations and aims to recommend not only items of interest, but also other relevant information like interesting people and/or topics. Explanations are also provided so that users can obtain a better understanding of the recommendations and decide which recommendations to pursue further. We discuss the technical challenges involved in characterizing different user behaviors and in efficiently computing recommendation explanations.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms

Keywords

Tagging, Collaborative Filtering

1. MOTIVATION

The recent advent of "Web 2.0," i.e., the evolution of the Web from a technology platform to a social milieu, has been accompanied by an explosion in the number and reach of *social content sites* such as *social tagging sites*. A typical example of such sites is Yahoo!'s del.icio.us, which enables users to tag URLs that are of interest, create a network of friends, and subscribe to their friends' feeds to learn about what URLs they have been tagging recently. With the increasing popularity of del.icio.us, we have seen an explosion of URLs in the system. And to help users navigate this large number of URLs, del.icio.us provides hotlists[1]. However, sifting through the large amounts of URLs and *finding the* right *URLs to recommend to the user* is still a challenging task facing del.icio.us, as well as many other similar sites (e.g., Yahoo! Movies, CiteULike, etc.).

URLs, however, are just part of the story. Increasingly, users of del.icio.us want to become connected with people who are not necessarily their friends but are nonetheless interesting to them, or topics that they have never been exposed to but are potentially of interest to them. Combined with URL recommendations, the mission statement has become *finding the right* information *to recommend to the user*, where information can be *URLs*, *people*, or *topics* (as represented by the tags).

Furthermore, users of del.icio.us can have very different behaviors depending on how they are using the system. On one end of the spectrum, there are users who use del.icio.us as a personal bookmarking tool where they tag URLs but don't necessarily share those URLs with the public or their friends (some don't even have any friend). On the other end, there are users who use del.icio.us as a discovery tool, where they establish a large network of friends and frequently explore the URLs that their friends have been tagging; but they themselves only tag very few URLs. Such diverse user behaviors present another challenge: *identifying the right recommendation strategy based on the right information about the user*.

The richness of information within social tagging sites presents a unique opportunity for the design of semantically-enriched recommender systems. We propose to demonstrate X.QUI.SITE, a system which gracefully incorporates user behavior into recommendations. To the best of our knowledge, X.QUI.SITE is the first system which captures users behavior and is able to recommend general information like topics and people, in addition to URLs. X.QUI.SITE is designed as a platform to incorporate multiple recommendation strategies depending on the user behavior and is based on a variety of methods for producing customized hotlists.

Moreover, X.QUI.SITE provides a *summarized trace* of recommendations, thereby *explaining* their provenance. For example, when a new URL is recommended to a user, it can be because many friends of the user are tagging it or it can be because it is related to some tags that the user previously used. At the user's request, such information will be provided to the user so she can make her own judgement of how interesting the recommendation is to her.

X.QUI.SITE is implemented as an external module of del.icio.us and runs on a replica of the production del.icio.us datasets with nightly synchronization. The scalability of the system relies on two efficient algorithms: the network generation algorithm and the explanation computation algorithm.

Sec. 2 contains a technical description of our recommendation methods and algorithms and Sec. 3 provides an overview of the demo.

[1]A list of most popular URLs or tags among a group of users within a given period of time.

SIGMOD'08, June 9–12, 2008, Vancouver, BC, Canada.
ACM 978-1-60558-102-6/08/06.

2. TECHNICAL CONTRIBUTIONS

At its core, a recommendation method is defined by three components: (i) *the recommendation results*, i.e., what information is being recommended; (ii) *the context of the recommendation*, which is characterized by a seed set used to draw recommendations from, and the current user's interest; and finally, (iii) *the strategy*, i.e., how recommendations are computed from the context. We investigate these components and argue for different recommendation strategies and recommendations depending on the context.

2.1 Recommendation Results

The richness of actions in social tagging sites, e.g., forming social ties with other users and choosing tags to describe URLs, provides the ability to think of recommendations along different axes. In other terms, users are no longer limited to getting recommendations of items such as bookmarked URLs. Instead, they can now explore people and topics as well.

Searching for people of interest to a user is a frequently requested feature. Indeed, while users in del.icio.us can define explicit social ties to form their friendship network, they have little help in creating new social ties. In X.QUI.SITE, we propose to derive interest networks (including both topic-based network and URL-based network) and use them to recommend people of potential interest to users.

Similarly, users of del.icio.us are always looking for new and relevant topics to explore. In X.QUI.SITE, we achieve this topic recommendation by recommending interesting tags to the user based on her explicit and derived networks.

These recommendations are in fact intertwined. Recommended people can help a user find out more about the latest URLs that are relevant to her specific topic of interest, while recommended topics (i.e., tags) can help her discover other interesting URLs and people.

2.2 Recommendation Context

Regardless of what information are being recommended, a recommendation is always produced in a certain context. This context may include a variety of information: some that is specific to the user (e.g., the user's network of friends) and some that is global (e.g., the most active group of users for a certain topic). The common piece of context, however, is the *seed set*: the set of users whose behavior is used to generate recommendations for a given user. This seed set of users may or may not depend on the user to whom the recommendation is being served.

The simplest recommendation strategy is based on modeling the seed set as all users in del.icio.us. As a result, the same seed set is used regardless of the user seeking recommendations. Alternatively, a seed set could be composed of a user's explicit social ties (i.e., friendship network), in which case the seed set is often different for different users. In addition to explicit seed sets, there are ones that can be derived based on common behavior between users. There are two basic approaches: using shared topics (i.e., tags) and using shared URLs. In the first approach, the set of tags associated with a given user is used to derive a seed set composed of all users who share a significant number of the tags with the user. For instance, if many of the users' tags include the word *sports*, the user is likely to be interested in sports-related information, and we compute a seed set of people who are interested in sports. This richer computation helps refine the recommendations by identifying information related to the user's interest, thereby customizing recommendations to individual users (or groups of users who share their most popular tags). In the second approach, the system considers overlap in tagged URLs between different users. Each user's seed set is defined as the set of all the other users who bookmarked a large fraction of the user's URLs. In addition to those two basic approaches, the system also combines the topic-based approach and the URL-based approach to define a richer seed set formed by the people who share the user's tags and URLs.

2.3 Recommendation Strategies

When the recommendation results are URLs, a recommendation strategy defines the normalized score of a recommended URL as the percentage of people in the seed set who bookmarked the URL (also referred to as URL popularity). When the recommendation results are people, the score of a recommended user is the strength of connection (as measured by the extent of shared tags or URLs) between her and the user. Finally, when the recommended results are topics, the normalized score of a tag is defined as the percentage of people in the seed set who have used that tag (also referred to as tag popularity).

Consequently, depending on the recommendation results, the recommendation system will apply different strategies to different users. Moreover, while using friends as the seed set to draw recommendations may matter to some users (e.g., those who have friends), using people who have interest overlap may matter to others (e.g., those who tag a lot). Due to lack of space, we only briefly describe URL recommendation strategies here.

Global and Global-Tag: These two strategies adopt the seed set of all del.icio.us users. The `Global` strategy chooses URLs that are globally popular. While these URLs usually represent consensus between most users, we experimentally observed [3] that they only account for a small fraction of any individual user's tagging.

In `Global-Tag`, we model the interests of a user as represented by the vocabulary she uses to tag URLs: if a significant portion of the user's tagging actions includes the tags *sports* and *nutrition*, the user is likely to be interested in sports-related and nutrition-related URLs. This simple observation allows us to refine a single globally popular list and to suggest potentially interesting URLs by drawing from one or more tag-specific popular lists in accordance with the user's interests.

URL-Interest and Tag-URL-Interest: These two strategies adopt social ties to definite an appropriate seed set for the user. The social ties here are implicitly derived based on user's tagging behaviors. In `URL-Interest`, these ties are generated based on overall tagging actions, while in `Tag-URL-Interest`, these ties are generated based on tagging actions within the scope of a few topics that are most relevant to the user.

Collaborative Filtering is a popular method in recommender systems that uses statistical techniques [2] to determine interest overlap between users based on their behavior such as common ratings of movies or common purchasing and browsing patterns. In `URL-Interest`, we adopt a similar approach, and construct the common interest network,

which links two users if the sets of URLs they tagged overlap significantly. However, we observe that using the entire set of URLs tagged by a user as a basis for constructing the social ties, while leading to high-quality overlap in interest, only applies to a small subset of the users in our user base [3].

One factor which limits the effectiveness of deriving interest overlap between users in Collaborative Filtering is *sparsity*: there are often many more items in the system than any one user is able to rate or review. This issue is further aggravated in the context of del.icio.us, where the set of URLs corresponds to a potentially infinite set of websites. Sparsity is one of the main reasons why using overlap in URLs to derive common interest networks is only effective for a subset of del.icio.us users. Another important reason is that people rarely agree on everything: you may agree with your mother on cooking, and with your adviser on research, but your adviser's opinion on food is hardly relevant. The `Tag-URL-Interest` strategy uses this idea to construct social ties that combine tag and URL overlaps. Such networks have wider applicability than URL-only interest networks, and can be used to construct recommendation lists of high quality [3].

Friends: del.icio.us also allows each user to define her friends and family, which can be considered as explicit social ties. The `Friends` strategy directly adopts this explicit social network as the seed set for recommendations. The rest is similar to the previous two strategies.

2.4 Algorithms Challenges

X.QUI.SITE recommendations are computed in three steps. First, different user similarity networks are generated offline based on common user behavior. Second, the recommendation strategies which are applicable to a user are executed and their results are computed and merged. Third, recommended items, i.e., URLs in this case, are explained. We summarize the challenges behind each step.

2.4.1 Network Computation

Unlike social networking sites such as Facebook, where users create large friendship networks, explicit connections are much rarer in collaborative tagging sites where the primary function is to help users identify and organize interesting content. Therefore, we generate implicit networks. By creating a link between two users who have shared common URLs, we are able to generate an implicit similarity network for a larger fraction of users. Another mechanism for generating implicit networks include using profile information (e.g., age and income) about the users. This information is not always available. As a result, URL overlap is more effective.

When the number of users grows, generating an implicit network is non-trivial. As an example, in `URL-Interest`, user-user ties are generated based on overlap in tagging actions. If we consider 400K unique users who have tagged at least one URL, a naive algorithm, which does a comparison between each pair of users, needs 160 billion comparisons to compute the `URL-Interest` networks. At the rate of 10 micro-seconds per comparison, it will take the algorithm 18 days to finish! Most of the comparisons are wasted, however, because an average user shares common URLs with a number of users far less than the total 400K other users. In other terms, the resulting user-user similarity matrix is often very sparse. Based on this observation, we developed an item-based similarity computation algorithm which is based on organizing items based on how many users have tagged/rated them and only does a comparison between two users if that comparison is likely to create a link.

2.4.2 Recommendation Generation

A user who has friends and tags regularly could benefit from more than one recommendation strategy: `Friends`, `URL-Interest` and `Tag-URL-Interest`. The question is then: how to combine recommended lists of items generated by each strategy. We provide a configuration file to specify the weight of each method and use them to combine recommended URLs.

2.4.3 Explanation Generation

The explanation of a recommended item di, for a given user du, is defined as the set of *contributors* who tagged the item.

A naive approach to computing explanations is to obtain the list of items to recommend to a user du, and then retrieve the set of contributors of each item di, by intersecting di's taggers and du's network. The system then collects the results and assembles all the contributors of a single item into a single list. This post- processing approach is not efficient since it aggregates items twice, once in the recommendation generation, another time in the explanation generation.

A more efficient approach avoids the double aggregations by maintaining a view where all the tagging actions belonging to the same user are stored together in a single list. Specifically, consider generating candidate recommendations for a user du. We retrieve the user lists associated with each user in du's network and apply the No Random Access Algorithm (NRA) to compute the candidate recommended items efficiently [1].

3. DEMONSTRATION OVERVIEW

We will demonstrate the following features in X.QUI.SITE:

URL Recommendation: Users can visualize different URL recommendations as shown in Figure 1. The list is a consolidation of recommendations produced from multiple recommendation strategies applicable to a given user (joshua in an example user). The interface allows users to subset the list of URLs to only those derived by a specific recommendation strategy, or those derived from a specific topic or a specific set of friends.

People Recommendation: Similarly, a user can also be recommended with a list of people that are of potential interest to him. Again, the list is combined from different lists that are generated based on different recommendation strategies, including *global-expert*, which identifies people who are experts (e.g., "frequent tagger" would be a simple example of the definition of expert) on the topics that the target user is interested in; *network-expert*, which identifies people who share the same interest with the friends of a given user, and various other strategies.

Topic Recommendation: Users can also browse the list of recommended topics (i.e., tags) in a similar way. Topic recommendation is generated based on the following strategies: *global-top-tags*, which identifies the globally popular hot topics; *friend-network-top-tags*, which identifies the most pop-

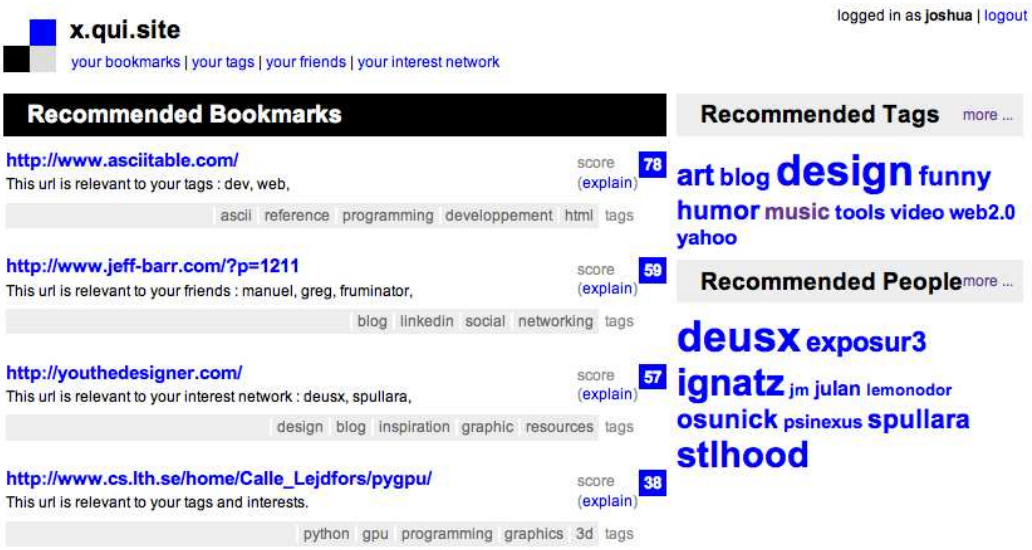

Figure 1: URL Recommendation Interface

ular topics among the user's friends; *interest-network-top-tags*, which identifies the most popular topics among the set of people who share the user's URLs.

Recommendation Explanation: The generation of a recommendation is explained along with the recommended result. We show both the basis of the recommendation (i.e., which URLs, topics, or people are used to make the recommendation) as well as the recommendation process (i.e., which strategy is used).

4. REFERENCES

[1] R. Fagin and et. al. Optimal Aggregation Algorithms for Middleware. In *PODS*, 2001.

[2] J. A. Konstan. Introduction to recommender systems. In *SIGIR07: Proceedings of the 30th Annual International ACM SIGIR Conference*, 2007.

[3] J. Stoyanovich, S. Amer-Yahia, C. Marlow, and C. Yu. A study of the benefit of leveraging tagging behavior to model users' interests in del.icio.us. In *AAAI-SIP*, 2008.