Tag-Based User Profiling for Social Media Recommendation

Chia-Chuan Hung and **Yi-Ching Huang** and **Jane Yung-jen Hsu** and **David Kuan-Chun Wu**
Department of Computer Science and Information Engineering
Graduate Institute of Networking and Multimedia
National Taiwan University
yjhsu@csie.ntu.edu.tw

Abstract

Making recommendations for social media presents special challenges. As tagging becomes common practice at many social media sites, this research proposes a new approach to user profiling based on the tags associated with one's personal collection of contents. To utilize the social interaction implied by tagging, a personal profile can be further extended with the tags specified by one's social contacts. A tag-to-tag matrix is defined to enable collaborative filtering-style recommendations without explicit user ratings. Experiments with collections of bookmarks and the associated tags from 42,463 users are presented and compared using the different views.

Introduction

The phenomenal rise of social media in recent years is transforming the average people from content readers to content publishers. Some popular social media services include del.icio.us[1] (social bookmarking), last.fm[2] (social music), flickr[3] (photo sharing), and YouTube[4] (video sharing), where people share a variety of media contents with their friends or the general public. Tagging is commonly used to add comments or descriptions about the media contents, or to help organize and retrieve relevant items.

Making recommendations for social media presents special challenges. First of all, the colossal set of user-generated content is *open-ended* and rapidly growing, making it difficult to define the vector space for recommenders. Secondly, user feedbacks are mostly *implicit* and *asymmetrical*. For example, adding a bookmark to my personal collection indicates my interests in the topic as well as the intention to access the page in the future. On the other hand, the fact that a bookmark is *not* in my collection does not necessarily represent a lack of interest. Thirdly, most explicit feedbacks are only *binary*.

Recommendations are generally made based on measures of similarity between people, contents and their interactions. A person may be profiled as a vector of attributes of his/her online personal profiles including the name, affiliation, and interests. Such simple factual data provide an inadequate description of the individual, as they are often *incomplete*, mostly *subjective* and cannot reflect dynamic changes. In collaborative filtering (Goldberg *et al.* 1992), a person is profiled by a vector of ratings, one for each media content. Observing that the rich online media collected by an individual provide important insights about the person, can we capitalize on such data in the absence of rating information?

This research explores the role of tagging for social media recommendation. We propose a new approach to user profiling based on the tags associated with the the user's personal collection of social media. In particular, a user is profiled by aggregating the tags specified by the user as well as his/her social contacts.

In what follows, we will start by briefly reviewing related research in both recommender systems and user profiling. The concept of tag-based user profiling is introduced with a set-theoretic definition. The tag-to-tag matrix is defined, followed by the process of making social media recommendation. This paper outlines our experiments with del.icio.us bookmarks and tags, presents the results of tag-based user profiles, and compares the recommendations due to personal view and social view.

Related Work

Recommender Systems Recommender systems has been an active area of research and practical applications (Adomavicius & Tuzhilin 2005). The approaches vary widely in terms the type of information considered in making the recommendations.

A *popularity-based* approach simply recommends the most popular resources, e.g. top music charts. It does not take the attributes from individual user or content into consideration.

A one-dimensional approach makes recommendations based on the attributes of either the people or the media contents independently. For example, *content-based* recommender systems usually analyze the content of items previously rated by a given user to build a model of the user's interests. Relevant items can then be recommended based on the trained user interests model.

A two-dimensional approach makes recommendations

[1] http://del.icio.us

[2] http://www.last.fm/

[3] http://www.flickr.com/

[4] http://www.youtube.com/

based on the *relationships* between the people and items (media contents). For example, *collaborative filtering* recommender systems (Goldberg *et al.* 1992) collect all users' ratings about all items, and make recommendations based on the previous ratings from the group of people who have similar tastes with the given user. Collaborative filtering suffers from start-up problem, so variations to compute item-item, user-user and user-item similarity have been proposed.

Some researches on user modeling also proposed profiling user by tags will benefit inferring knowledge about a user. (Firan, Nejdl, & Paiu 2007) utilize rich genres and user data on the music community site, Last.fm[5], to identify users' preferred music genres. They define several types of tag-based user profiles and their corresponding recommender and search algorithms. A simple tag analyzing method is proposed to consider public tags and tagging frequency on a track owned by a user to determine relevant tags and their associated scores. Follow the idea of collaborative filtering recommender system, they find some similar users from a user-tag matrix, and recommend music pieces containing the similar users' tags. Compare with conventional track-based recommender approach as a baseline, their experiment shows that tag-based user profile significantly improves the quality of results.

User Profiles Research in (Liu & Maes 2005) harvests profiles from social networking websites, such as Friendster[6], MySpace[7], and Orkut[8], to construct the *InterestMap*, a network-style user profile to illustrate the relationship between interests and identities. Unlike traditional recommender systems, the proposed approach recommends by considering the interests of people instead of their historical behavior in a particular application. The InterestMap produces more accurate recommendations, and the preferences and interests of people in real life are modeled in an intuitive and visual fashion.

The idea of constructing user profiles from tagging data has been proposed in (Michlmayr & Cayzer 2007). They use a profile graph to represent a user, where nodes are tags used by this user and edges are the relations between tags. They design the *Add-A-Tag* algorithm, an adaptive approach that combines co-occurrence and temporal information to determine the edge weight between a pair of tags. They also provide a graph animation to visualize a dynamic user profile. Their user study shows that users still desire to see long-term tag relationships (which are identified by traditional co-occurrence method) in their profile. However they also appreciated that Add-A-Tag adapts better to visualizing recent changes.

Tag-Based User Profiling

Most social media sites support tagging mechanism. For example, bookmarks on del.icio.us may be tagged with the topics of interest to the user; a picture on Flickr may be tagged with its location, the event, people and objects in the picture, color or mood depicted in the picture. *Tagging* associates an object (e.g. a picture, a web page etc.) with a set of words, which represent the *semantic concepts* activated by the object at the cognitive level. Tagging provides a simple yet powerful way for organizing, retrieving and sharing different types of social media.

While categorization is a primarily subjective decision process, tagging is a social indexing process. In (Sinha 2006), Sinha succinctly pointed out that "Tagging captures our individual conceptual associations, but does not force us to categorize. It enables loose coordination, but does not enforce the same interpretation of a concept. We could all tag items as 'art' but mean very different things. That would create chaos in a shared folder scheme, but works well in a social tagging system." In addition, Sinha offered the following insightful oberservations.

- Tagging transforms web browsing from a *solitary* to a *social* experience. Tagging specific resources creates ad-hoc groups, leading to "wisdom of crowds".
- Tagging enables social coordination that is simultaneously more *direct* and *abstract* than collaborative filtering, as tags connects entities directly and enables tranfer of conceptual information.

Social Media Network

To explore the role of tagging for social media, we define a *social media network* to be a heterogeneous network of people and their (common) media collection. Figure 1 depicts a simple network with people and their social connections (denoted by circles and solid lines), as well as bookmark URLs and bookmark actions (denoted by oval-shaped nodes and dotted directed arrows). Note that while users P_x and P_y are not John's social contacts, they are "related to" John via the common bookmark URL_8.

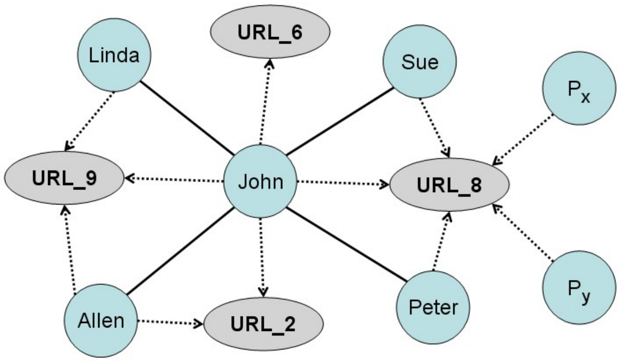


Figure 1: A sample social media network.

Tagging as Profiles

Instead of the vector of item ratings, each user p is profiled as a set of tags and their weights. Weight represents the importance of this tag to p. A tag-based user profile can be

[5] http://www.last.fm/
[6] http://www.friendster.com
[7] http://www.myspace.com
[8] http://www.orkut.com

defined as follows:

$$Profile(p) = \{\langle t_1, w_1\rangle, \langle t_2, w_2\rangle, \cdots\} \quad (1)$$

The profiling tags t_i can be harvested from multiple data sources below.

- All data and descriptions in the registered user profile. This source of information ranges from the bare minimal, e.g. only name and homepage URL for del.icio.us, to rich descriptions as in many social networking sites.
- Tags specified by the user for self description, or tags used explicitly by his/her friends to describe the given user.
- Tags associated with the user's collection of social media, which reflect his/her topics of interest as well as activities.

In this paper, we firstly consider tags on social media content.

The *personal view* of a user profile considers only tags specified by the user, while the *social view* of a user profile includes tags specified by his/her friends on his/her collection of contents. Let **P** be the set of all people, and **C** be the collection of all resources. We can define $P(t_x)$ to be the set of people who have tag t_x in their user profiles, that is, it defines the set of people who share the tag in common. Similarly, we may define the set of people $P(r_y)$ who has a specific resource r_y in their personal collection. Given a person p, his social view is the aggregation of the tag set of his/her personal collection $C(p)$ with the tag sets defined by people $P(r)$ who also collect any resource $r \in C(p)$ and are friends with p.

The two views differ not only in the tag collections, but also in their associated tag weights. The method for analyzing tag weight should be adapted to consider the types of resources, as well as user's tagging behavior and usage pattern. We will describe in more detail our designed method, implemented specifically on bookmarking data, in the section on experiments.

Social Media Recommendation

Recommendation based on tag-based user profile will be made by abstractly considering the relationship between user attributes and content attributes. From a tag-based user profile, we have some tags as his/her user attributes. For example, a user is good at "programming", is fond of "travel", and is a "humorous" guy. Among these users, some user attributes are common, but some may be uniquely owned by an individual. Aggregating the user attributes of *all* individuals, we obtain a *user attributes set*. We can also list the attributes of a piece of content. It is a "wiki" page about "design", and it includes many "cool" ideas. We obtain a *content attribute set* by gathering these content attributes on *all* pieces of content. Note that these content atrributes can be keywords mined by information retrieval techniques or tags that are frequently assigned by many people. (We utilize tags in this paper.) Rather than inferring the relationship between user preference and item attribute from rating data, like many traditional recommender systems do, in our proposed social media recommendation, we obtain the relationship by analyzing these attributes which are represented as tags. Tag-based relationship abstracts the general semantic between users and items, avoiding overly-specific problems in traditional recommender systems.

Tag-to-Tag Matrix

We design a matrix for capturing public preferences, i.e. how much percentage of people who like topic A will also like topic B. Thus a tag-to-tag (T2T) matrix records the relevance between a pair of user tag (user attributes) and content tag (content attribute). (Note that a user tag may belong to the associated tag-based user profiles of many users, and a content tag may belong to many pieces of content.) In this T2T matrix, each row entry represents a user tag in *person attribute set*, while each column entry represents a content tag in *content attribute set*. A higher relevance value means the corresponding pair of tags is highly relevant. For example, people who like "art" are interested in the content about "design". Thus, tag "art" and tag "design" are relevant. We calculate the value in each field of the T2T matrix using Equation (2)

$$T2T[ut_i][ct_j] = \frac{|P(ut_i) \bigcap P(ct_j)|}{|P(ut_i)|} \quad (2)$$

where ut_i and ct_j represent user tag i and content tag j respectively. $P(t_x)$ is a set of people who have tag x in their tag-based user profiles. Thus the value of $T2T[ut_i][ct_j]$ is the proportion of the people with both tags i and j within the people with tag i. Note that when $ut_i = ct_j$, i.e. two tags are identical, $T2T[ut_i][ct_j]$ is 1.

Making Recommendation

Given a person p and a piece of content c, we will determine a recommendation score by looking up the T2T matrix. Suppose p's tag-based user profile has n tags, $\{ut_1, ut_2, \ldots, ut_n\}$, and corresponding tag weight w_i for $i = 1$ to n. Content c has k frequently assigned tags, $\{ct_1, ct_2, \ldots, ct_k\}$. We define a p's tag ut_i to be a *tag feature vector*, and elements in this vector are floating numbers corresponding to c's tag ct_j. Thus the length of each tag feature vector is k. A tag feature vector represents the meaning that, given a user who has this tag ut_i and its corresponding tag weight reflect the importance, the probability he will interested in this piece of content corresponding to its each tag. The tag feature vector for tag ut_i is defined as:

$$tag\_feature[ut_i] = \langle \phi(ct_1), \phi(ct_2), \ldots, \phi(ct_k)\rangle \quad (3)$$

where $\phi(ct_j) = w_i \times T2T[ut_i][ct_j]$.

Then all tag feature vectors of this user are combined to obtain an overall user feature vector. For a content tag, we choose the maximum feature value among this user's tag feature vectors. Instead of using other operations, e.g. summation, we only consider the most relevant user attribute to avoid summing up many irrelevant attributes as a high feature value.

$$\Theta(ct_j) = \max_{\forall i \in n} tag\_feature[ut_i][ct_j] \quad (4)$$

Thus the overall user feature vector can be defined as:

$$user\_feature = \langle \Theta(ct_1), \Theta(ct_2), \ldots, \Theta(ct_k)\rangle \quad (5)$$

This user feature vector represents, for each idea mentioned in c, the highest probabilties p may have. Thus we simply sum up these probabilties to obtain a recommendation score. If this score is above a threshold T, we decide to recommend the content.

Experiments with Social Bookmarking

The proposed idea can be applied to any type of social media with tags. In our experiment, for simplicity, social bookmarks are used as the data source, and each bookmarked URL is assumed to have multiple tags.

The Data

Del.icio.us is a popular social bookmarking website, which contains rich and public personal bookmark collection. Our analysis is applied on two sets of del.icio.us data. One is "URL" data set, and the other is "USER". To ensure we have enough tagging data, we set some conditions to filter our collected data. We randomly sample 65,131 users who have 300 to 1500 bookmarks. There are 7,258,267 unique URLs among these users' bookmarks. We then choose those URLs which have been bookmarked by 70 to 200 people. Thus 42,844 URLs are left, and we name this set of data as the "URL" set. Finally, we examine each user's bookmarking data, and only someone who has at least 50 bookmarked URLs within our URL set would belong to our "USER" set. There are 42,643 users in the USER set.

For each URL and user in both sets, we log the user's complete tagging history. There are 34,427 distinct tags (that are tagged by at least two people, from the statistic of del.icio.us) in the URL set (we name these tags as content tags in the previous section). For each user, we choose top 30 tags from his/her personal user profile. Thus a total of 28,290 distinct tags (USER tags) are included. Table 1 shows a summary of our experiment data.

Set	Amount	Average	Tags
URL	42,844 URLs	112.556 people bookmarked (per URL)	34,427
USER	42,643 users	94.718 bookmarks (per user)	28,290

Table 1: Summary of two data sets

We also collect each user's personal social networking data for producing social user profile. There are 12,794 users who have at least one friend in the USER set (2.89 friends in average). For these users, we also choose top 30 tags from their social user profiles. A total of 11,600 distinct tags (USER tags) are included.

Tag Analysis

Each bookmarked URL is given one or more tags to describe the content of the webpage. We define a value, *capacity*, to represent how much a tag can describe the content referred by a URL. By analyzing the nature and idea of a tag, the popularity of an identical tag to the same content, and the tagging order, we can determine the capacity of a tag.

Tagging Order Research on users' tagging patterns (Golder & Huberman 2006) discovered that the first tag used has the highest median rank (i.e. greatest frequency), and successive tags have a decreasing median rank. Thus they suggest that the early tags in a bookmark represent basic levels, "because they are not only widespread in agreement, but are also the first terms that users think of when tagging the URLs in question." Therefore, we exploit this idea and assume that the first tag is more relevant than the second tag on a bookmark. Let $\mathbf{C}$ denote a bookmark collection and $\mathbf{T}(\mathbf{C})$ denote a set of tags which are assigned on $\mathbf{C}$ by many people. A tuple of bookmarking data is denoted as $b = (p, \mathbf{T}(c, p), c)$ which means a person p who tagged a piece of content $c \in \mathbf{C}$ with a sequence of tags $\mathbf{T}(c, p) = \{t_1, t_2, \cdots, t_n\}$. We firstly define the order-weight of tag $t_i \in \mathbf{T}(c, p)$ as

$$w_{order}(t_i, p) = \frac{1}{|\mathbf{T}(c, p)|} \times \begin{cases} \exp^{-i/10} & \text{if } i \leq 10 \\ \exp^{-1} & \text{if } i > 10 \end{cases} \quad (6)$$

where i is the index of t_i in this ordered tagging sequence, and $|\mathbf{T}(c, p)|$ is a normalized term. Here we assume tags after the 10^{th} tag have equal order weight. Exponential decreasing function is applied because it is more easily implemented, rather than defining a linear decreasing function.

Tagging Popularity We sum up the weights provided by people who have the same tags on the same URL, and normalize it with the number of bookmarks for this URL to determine the importance of a tag t for this URL referring content c. We define $\mathbf{P}(c, t)$ as a set of people who had once assigned this piece of content c with tag t, and $|\mathbf{P}(c)|$ as the amount of people who had once bookmarked c's URL. Thus combining with Equation (6), the *capacity* of t on c is

$$capacity(t, c) = \frac{1}{|\mathbf{P}(c)|} \times \sum_{p \in \mathbf{P}(c,t)} w_{order}(t, p) \quad (7)$$

User Profile

The importance of a tag t_i to an individual, denoted as a weight w_i, is determined by analyzing its capacity, and the volume of covered bookmarks. The results are different from two viewpoints although we are describing the same person. (Figure 2)

Personal View Suppose a person p owns a set of content pieces $\mathbf{C}(p)$ which he/she has bookmarked. From the personal viewpoint, we only consider the tags assigned by p, and denote these tags as a set $\mathbf{T}(\mathbf{C}(p), p)$. For each tag $t_i \in \mathbf{T}(\mathbf{C}(p), p)$, it's weight w_i from p's view can be calculated by

$$TagWeight_{personal}(t, p) = \frac{\sum_{\forall c \in \mathbf{C}(p)} capacity(t, c)}{|\mathbf{C}(p)|} \quad (8)$$

Thus p's personal tag-based user profile can be defined as

$$Profile_{personal}(p) = \{\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \cdots\} \quad (9)$$

where $t_i \in \mathbf{T}(\mathbf{C}(p), p)$ and w_i is calculated by Equation (8).

Social View From the social viewpoint, we consider the tags assigned by p's social contacts, $\mathbf{A}(p)$. Note that we still focus on the pieces of content in $\mathbf{C}(p)$, but strictly consider the content pieces of URLs which are bookmarked both by p and by his/her social contacts, denoted as $\mathbf{C}_{social} = \mathbf{C}(p) \bigcap \mathbf{C}(\mathbf{A}(p))$ Our considered tags are those tags which were assigned by p's social contacts on these content pieces, denoted as $\mathbf{T}(\mathbf{C}_{social}, \mathbf{A}(p))$. For each tag $t_i \in \mathbf{T}(\mathbf{C}_{social}, \mathbf{A}(p))$, its weight w_i from social view can be calculated by

$$TagWeight_{social}(t,p) = \frac{\sum_{\forall c \in \mathbf{C}_{social}} capacity(t,c)}{|\mathbf{C}_{social}|} \tag{10}$$

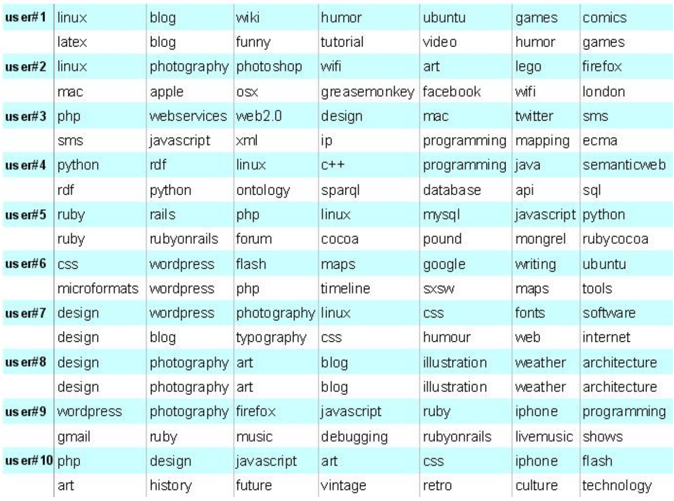
Then we can obtain a user tagging profile from social view.

$$Profile_{social}(p) = \{\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \cdots\} \tag{11}$$

where $t_i \in \mathbf{T}(\mathbf{C}_{social}, \mathbf{A}(p))$ and w_i is calculated by Equation (10).

Results

In this section, we describe some observations from the results of our experiment. We randomly select 10 users and depict their personal and social tag-based user profiles as Figure 2.



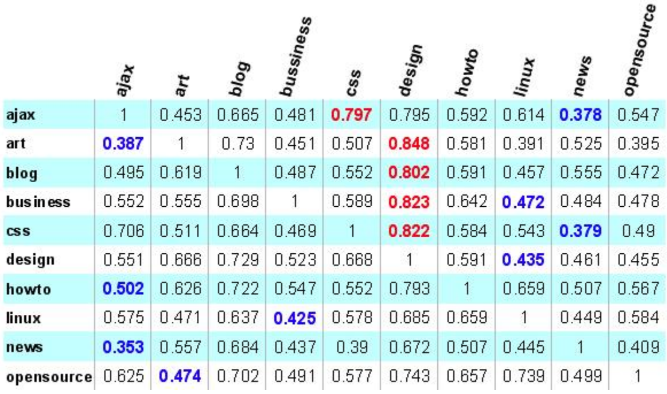
user#1	linux	blog	wiki	humor	ubuntu	games	comics
	latex	blog	funny	tutorial	video	humor	games
user#2	linux	photography	photoshop	wifi	art	lego	firefox
	mac	apple	osx	greasemonkey	facebook	wifi	london
user#3	php	webservices	web2.0	design	mac	twitter	sms
	sms	javascript	xml	ip	programming	mapping	ecma
user#4	python	rdf	linux	c++	programming	java	semanticweb
	rdf	python	ontology	sparql	database	api	sql
user#5	ruby	rails	php	linux	mysql	javascript	python
	ruby	rubyonrails	forum	cocoa	pound	mongrel	rubycocoa
user#6	css	wordpress	flash	maps	google	writing	ubuntu
	microformats	wordpress	php	timeline	sxsw	maps	tools
user#7	design	wordpress	photography	linux	css	fonts	software
	design	blog	typography	css	humour	web	internet
user#8	design	photography	art	blog	illustration	weather	architecture
	design	photography	art	blog	illustration	weather	architecture
user#9	wordpress	photography	firefox	javascript	ruby	iphone	programming
	gmail	ruby	music	debugging	rubyonrails	livemusic	shows
user#10	php	design	javascript	art	css	iphone	flash
	art	history	future	vintage	retro	culture	technology

Figure 2: The top 7 tags of tag-based user profiles. For each user, the first row is from personal viewpoint while the second is from social viewpoint.

The results of tag-based user profiles from two viewpoints are different. Note **user#10** give us very different impressions after looking his/her personal and social profile. We calculate symmetric difference between personal and social profiles of the 12,794 users, and obtain the average differences are 91.2% (only consider top 10 tags) and 95% (consider the entire tag profile). The results are very different from two viewpoints although we are describing the same person.

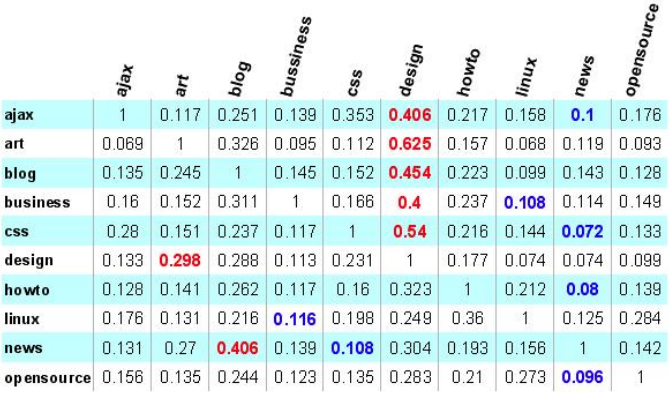
Results of Tag-to-Tag Matrix

After applying Equation (2), we obtain two tag-to-tag matrices from personal and social viewpoints, as Figure 3 and Figure 4 depicted respectively. We only show top 10 common tags from USER and URL tags in these two figures. Table 2 lists, for each tag, how many people own this tag, and how much URLs have been tagged on this tag.



	ajax	art	blog	bussiness	css	design	howto	linux	news	opensource
ajax	1	0.453	0.665	0.481	**0.797**	0.795	0.592	0.614	**0.378**	0.547
art	**0.387**	1	0.73	0.451	0.507	**0.848**	0.581	0.391	0.525	0.395
blog	0.495	0.619	1	0.487	0.552	**0.802**	0.591	0.457	0.555	0.472
business	0.552	0.555	0.698	1	0.589	**0.823**	0.642	**0.472**	0.484	0.478
css	0.706	0.511	0.664	0.469	1	**0.822**	0.584	0.543	**0.379**	0.49
design	0.551	0.666	0.729	0.523	0.668	1	0.591	**0.435**	0.461	0.455
howto	**0.502**	0.626	0.722	0.547	0.552	0.793	1	0.659	0.507	0.567
linux	0.575	0.471	0.637	**0.425**	0.578	0.685	0.659	1	0.449	0.584
news	**0.353**	0.557	0.684	0.437	0.39	0.672	0.507	0.445	1	0.409
opensource	0.625	**0.474**	0.702	0.491	0.577	0.743	0.657	0.739	0.499	1

Figure 3: The result of T2T matrix from personal viewpoint (personal USER tags v.s. URL tags). Here we only show top 10 tags from the common USER and URL tags



	ajax	art	blog	bussiness	css	design	howto	linux	news	opensource
ajax	1	0.117	0.251	0.139	0.353	**0.406**	0.217	0.158	**0.1**	0.176
art	0.069	1	0.326	0.095	0.112	**0.625**	0.157	0.068	0.119	0.093
blog	0.135	0.245	1	0.145	0.152	**0.454**	0.223	0.099	0.143	0.128
business	0.16	0.152	0.311	1	0.166	**0.4**	0.237	**0.108**	0.114	0.149
css	0.28	0.151	0.237	0.117	1	**0.54**	0.216	0.144	**0.072**	0.133
design	0.133	**0.298**	0.288	0.113	0.231	1	0.177	0.074	0.074	0.099
howto	0.128	0.141	0.262	0.117	0.16	0.323	1	0.212	**0.08**	0.139
linux	0.176	0.131	0.216	**0.116**	0.198	0.249	0.36	1	0.125	0.284
news	0.131	0.27	**0.406**	0.139	**0.108**	0.304	0.193	0.156	1	0.142
opensource	0.156	0.135	0.244	0.123	0.135	0.283	0.21	0.273	**0.096**	1

Figure 4: The result of T2T matrix from social viewpoint (social USER tags v.s. URL tags). Here we only show top 10 tags from the common USER and URL tags

We marked lowest (blue) and highest (red) values in each row. Note that some lowest and highest positions are changed in different views. Furthermore, all values are lower in social view, The tags in social profile are much less than personal profile, because we strictly define our method for the friends that owned common URLs with sampled user. However, it is a strong condition, and in fact, only a few cases are satisfied.

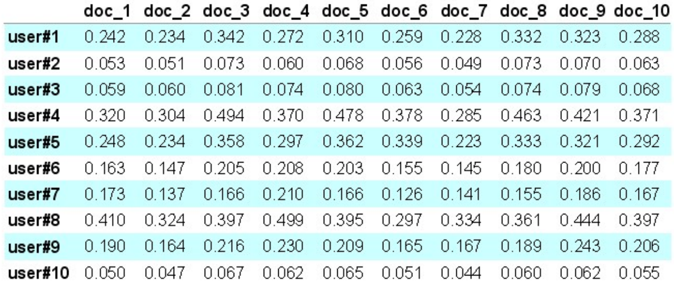
Results of Social Media Recommendation

The scores of Social Media Recommendation are depicted on Figure 5 and Figure 6 from personal view and social view

	#People	#URL
ajax	14344	3664
art	12518	7107
blog	17330	12173
business	7603	5323
css	18086	3896
design	20969	14540
howto	11059	11633
linux	17509	4274
news	8147	4455
opensource	6612	5933

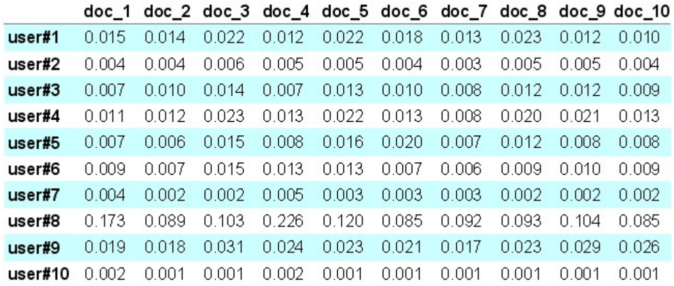
Table 2: A list of the top 10 tags from the common USER and URL tags

respectively. Ideally, the threshold should be obtained by machine learning techniqes, and adjusted from user feedbacks. Here we simplify the process, setting it as the average of the scores in each row, and then determine whether a piece of content is recommended or not. The decisions for recommendation are showed in Figure 7. We marked the inconsistent decisions between different views. Note that **user#10** has many inconsistencies because his profiles from two viewpoints are very different.



	doc_1	doc_2	doc_3	doc_4	doc_5	doc_6	doc_7	doc_8	doc_9	doc_10
user#1	0.242	0.234	0.342	0.272	0.310	0.259	0.228	0.332	0.323	0.288
user#2	0.053	0.051	0.073	0.060	0.068	0.056	0.049	0.073	0.070	0.063
user#3	0.059	0.060	0.081	0.074	0.080	0.063	0.054	0.074	0.079	0.068
user#4	0.320	0.304	0.494	0.370	0.478	0.378	0.285	0.463	0.421	0.371
user#5	0.248	0.234	0.358	0.297	0.362	0.339	0.223	0.333	0.321	0.292
user#6	0.163	0.147	0.205	0.208	0.203	0.155	0.145	0.180	0.200	0.177
user#7	0.173	0.137	0.166	0.210	0.166	0.126	0.141	0.155	0.186	0.167
user#8	0.410	0.324	0.397	0.499	0.395	0.297	0.334	0.361	0.444	0.397
user#9	0.190	0.164	0.216	0.230	0.209	0.165	0.167	0.189	0.243	0.206
user#10	0.050	0.047	0.067	0.062	0.065	0.051	0.044	0.060	0.062	0.055

Figure 5: The score of social media recommendation using personal tag-based user profiles.

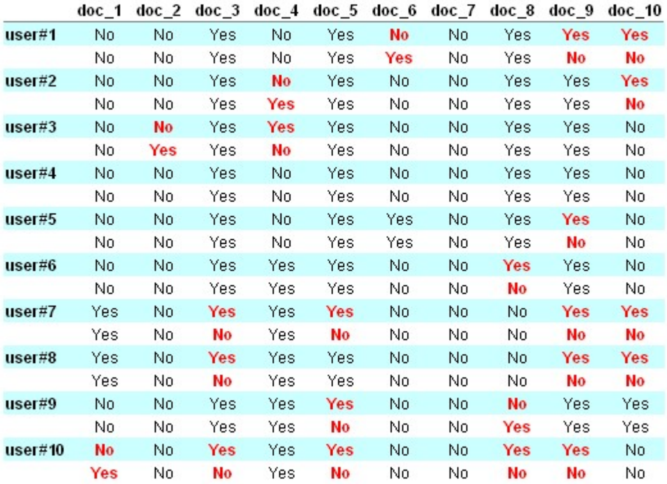


	doc_1	doc_2	doc_3	doc_4	doc_5	doc_6	doc_7	doc_8	doc_9	doc_10
user#1	0.015	0.014	0.022	0.012	0.022	0.018	0.013	0.023	0.012	0.010
user#2	0.004	0.004	0.006	0.005	0.005	0.004	0.003	0.005	0.005	0.004
user#3	0.007	0.010	0.014	0.007	0.013	0.010	0.008	0.012	0.012	0.009
user#4	0.011	0.012	0.023	0.013	0.022	0.013	0.008	0.020	0.021	0.013
user#5	0.007	0.006	0.015	0.008	0.016	0.020	0.007	0.012	0.008	0.008
user#6	0.009	0.007	0.015	0.013	0.013	0.007	0.006	0.009	0.010	0.009
user#7	0.004	0.002	0.002	0.005	0.003	0.003	0.003	0.002	0.002	0.002
user#8	0.173	0.089	0.103	0.226	0.120	0.085	0.092	0.093	0.104	0.085
user#9	0.019	0.018	0.031	0.024	0.023	0.021	0.017	0.023	0.029	0.026
user#10	0.002	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.001

Figure 6: The score of social media recommendation using social tag-based user profiles.

Accuracy

Following traditional accuracy measurment of collaborative filtering recommender system, 5-fold validation is applied on our recommendation results. We randomly re-sample 11,462 users and hide their 20% bookmarks to generate their tag-based user profiles. However, the overall bookmark collection of 11,462 users includes 2,795,303 unique URLs. (A long tail distribution - many URLs are bookmarked by fewer people.) Because of computing complexity, we can not consider all of them. Therefore, from the sampled users' bookmark collection, we select 10,192 pieces of content (URLs) which are bookmarked by at least 72 sampled users as the candidates for recommendation. In fact, the average overlap between a user's bookmarked items and selected candidate pool is only 20.3% (that means the highest precision will not be above 20.3%). Then for each user, according to previously obtained tag-to-tag matrix, these content pieces are ranked by their recommendation scores. The precision curve is depicted as Figure 8. X-axis shows the recommendations numbers (10,192 pieces in total), and y-axis represents the average percentage of matched URLs (i.e. they have been bookmarked by the users) over x recommendations.



	doc_1	doc_2	doc_3	doc_4	doc_5	doc_6	doc_7	doc_8	doc_9	doc_10
user#1	No	No	Yes	No	Yes	**No**	No	Yes	**Yes**	**Yes**
	No	No	Yes	No	Yes	**Yes**	No	Yes	**No**	**No**
user#2	No	No	Yes	**No**	Yes	No	No	Yes	Yes	**Yes**
	No	No	Yes	**Yes**	Yes	No	No	Yes	Yes	**No**
user#3	No	**No**	Yes	**Yes**	Yes	No	No	Yes	Yes	No
	No	**Yes**	Yes	**No**	Yes	No	No	Yes	Yes	No
user#4	No	No	Yes	No	Yes	No	No	Yes	Yes	No
	No	No	Yes	No	Yes	No	No	Yes	Yes	No
user#5	No	No	Yes	No	Yes	Yes	No	Yes	**Yes**	No
	No	No	Yes	No	Yes	Yes	No	Yes	**No**	No
user#6	No	No	Yes	Yes	Yes	No	No	**Yes**	Yes	No
	No	No	Yes	Yes	Yes	No	No	**No**	Yes	No
user#7	Yes	No	**Yes**	Yes	**Yes**	No	No	No	**Yes**	**Yes**
	Yes	No	**No**	Yes	**No**	No	No	No	**No**	**No**
user#8	Yes	No	**Yes**	Yes	Yes	No	No	No	**Yes**	**Yes**
	Yes	No	**No**	Yes	Yes	No	No	No	**No**	**No**
user#9	No	No	Yes	Yes	**Yes**	No	No	**No**	Yes	Yes
	No	No	Yes	Yes	**No**	No	No	**Yes**	Yes	Yes
user#10	**No**	No	**Yes**	Yes	**Yes**	No	No	**Yes**	**Yes**	No
	Yes	No	**No**	Yes	**No**	No	No	**No**	**No**	No

Figure 7: The decisions of social media recommendation. We marked differences between personal and social views.

As our expected, the precision is very low. The average precision is 2.77% when only recommend top 100 URLs. 41% accuracy is our best result, but interestingly, this user has some special (unusual) but important tags such as "illustration" and "belgique" in his/her tag-based user profile. Because these tags refer to much less content pieces (URLs) (compare with some tags such as "design" or "art" refer to thousands of content pieces), it is more easily hitting the ground truth.

The colossal set of user-generated content is open-ended and rapidly growing, and a *long tail* exists in the social media data. Compare with traditional data set (e.g. movie data) that has some central topics and can be formal defined, the topics of social media data are very diverse. Furthermore, bookmarking activity may depend on the order of information receiving. For example, if a person see an incomplete guiding web page at first, he may bookmark it. However, if he/she firstly find a complete tutorial, he may not be in-

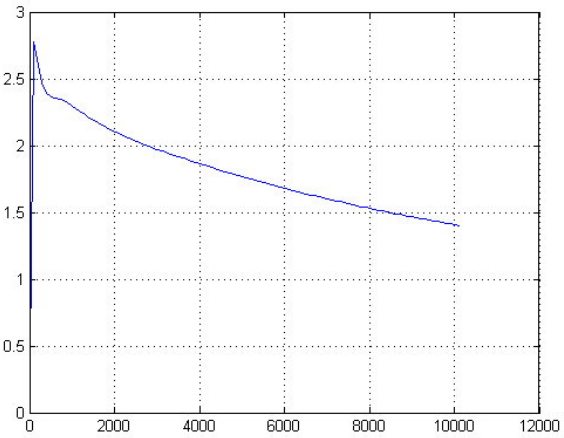


Figure 8: The precision curve of social media recommendation using tag-based user profiles.

terested in the imcomplete guideline. Therefore, we think calculating accuracy is not adequate to represent the performance, and we will further evaluate our work by user study.

Conclusion and Future Work

This paper presented our research on *tag-based user profiling* for social media recommendation. A user is profiled based on the tags associated with his/her social media, as well as the tags on the collection specified by his/her social contacts. We introduced the concept of tag-based profiling with a set-theoretic definition. The tag-to-tag matrix is defined, followed by the process of making social media recommendation. This paper presented our experiments with del.icio.us bookmarks and tags from 42,643 users, filtered with selection criteria to remove outliers. We compared the results of recommendations due to the personal view vs. the social view.

In addition to utilizing a list of weighted tags as profiles to recommend interesting content to users, we believe a person's profile should involve not only tag weights but also semantic relationship between tags. While weighted tags can represent a user's topics of interests with their corresponding preference degrees, tag relationship can represent the relationship between topics. It is expected that tag relationship will improve the results of social media recommendation. A small-scale user study including 15 to 20 testers will be launched. Each tester is presented with several recommend web pages based on his/her tag-based user profile to collect their feedbacks and comments.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and Tsung-Chieh Chang for his help in the data collection for the experiments reported in this paper. This work was supported in part by the Excellent Research Projects of National Taiwan University, under grant 95R0062-AE00-05, and by the National Science Council of Taiwan under grant NSC-95-2622-E-002-018.

References

Adomavicius, G., and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17(6):734–749.

Firan, C.; Nejdl, W.; and Paiu, R. 2007. The benefit of using tag-based profiles. In *Proceedings of the 2007 Latin American Web Conference (LA-WEB 2007)*, 32–41. Washington, DC, USA: IEEE Computer Society.

Goldberg, D.; Nichols, D.; Oki, B. M.; and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM* 35(12):61–70.

Golder, S., and Huberman, B. A. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32(2):198–208.

Liu, H., and Maes, P. 2005. InterestMap: Harvesting social network profiles for recommendations. In *Proceedings of the Beyond Personalization 2005 Workshop*.

Michlmayr, E., and Cayzer, S. 2007. Learning user profiles from tagging data and leveraging them for personal(ized) information access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007)*.

Sinha, R. 2006. A social analysis of tagging. World Wide Web electronic publication.