A Content-Based Method to Enhance Tag Recommendation[*]

Yu-Ta Lu, Shoou-I Yu, Tsung-Chieh Chang, Jane Yung-jen Hsu
Department of Computer Science and Information Engineering
National Taiwan University
{b94063, b94065, r96008, yjhsu}@csie.ntu.edu.tw

Abstract

Tagging has become a primary tool for users to organize and share digital content on many social media sites. In addition, tag information has been shown to enhance capabilities of existing search engines. However, many resources on the web still lack tag information. This paper proposes a content-based approach to tag recommendation which can be applied to webpages with or without prior tag information. While social bookmarking service such as *Delicious*[1] enables users to share annotated bookmarks, tag recommendation is available only for pages with tags specified by other users. Our proposed approach is motivated by the observation that similar webpages tend to have the same tags. Each webpage can therefore share the tags they own with similar webpages. The propagation of a tag depends on its weight in the originating webpage and the similarity between the sending and receiving webpages. The similarity metric between two webpages is defined as a linear combination of four cosine similarities, taking into account both tag information and page content. Experiments using data crawled from *Delicious* show that the proposed method is effective in populating untagged webpages with the correct tags.

1 Introduction

The phenomenal rise of social media in recent years has enabled an average person from being mere content readers to content publishers. People share a variety of media contents with their friends or the general public on social media sites. Tagging is commonly used on these sites to add comments about the media content, or to help organize and retrieve relevant items. Tagging associates a resource with a set of words, which represent the semantic concepts activated by the resource at the cognitive level. While categorization is a primarily subjective decision process, tagging is a social indexing process.

Tag information is useful in many aspects. One aspect is that tags help describe the content in a page, revealing its semantic meaning. They not only emphasize the key terms of a page but also contain some additional information that is not present in the page text [Bischoff *et al.*, 2008]. Another facet is that tags may be useful for search. This includes personal archive administration, where people use tags to search for documents in their collection, and possibly web search. Even though the issue that whether tags enhances web search has been a subject of debate for some time, tags undeniably provide good information for documents they annotate.

Despite the advantages tags have, tags are not truly helpful in the current web, caused by the fact that most documents, or webpages, contain little or no tag information. Some social bookmarking websites such as *Delicious* provide tag information for pages annotated by users. However, the number of pages being tagged is still too small to have a big impact on current search engines [Heymann *et al.*, 2008a]. According to the estimation of [Heymann *et al.*, 2008a], there are about 30 to 50 million unique URLs posted publicly on *Delicious* and the number of total posts is only a small portion of the web, which has at least billions of webpages. To make matters worse, even if a URL is bookmarked on *Delicious*, the URL may not have enough tag information, because the total number of tags annotating a URL follows the power rule. Figure 1 is drawn using data from a subset of our dataset: 685,418 URLs crawled from *Delicious*. 94% of the URLs have less than 50 total tags, meaning that even if the URL is bookmarked, there is still a high probability that it has very few tags. One solution to deal with the scarcity of tags on the web is to develop an automatic tag annotating mechanism that helps make tag information more available. Unfortunately, on *Delicious*, tag recommendation is available only for pages with tags specified by other users. Therefore, for those webpages with absolutely no tag information, a new tag annotation method must be used.

This paper proposes a method for content-based tag recommendation that can be applied to webpages with or without prior tag information. The recommended tags for a webpage can be used not only as recommendations to users but also to automatically annotate the page. Our method first introduces the idea of *tag/term coverage*, which is an entropy-based metric describing how fully the tags/terms represent the annotated document. Terms here refer to the words in the page

[*] This research was supported by grants from the National Science Council NSC 97-2815-C-002-106/107-E, NSC97-2622-E-002-010-CC2, and Ministry of Education in Taiwan 97R0108.

[1] delicious.com

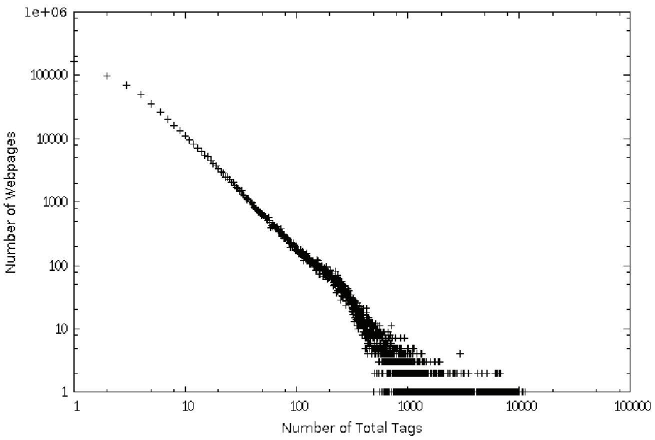

Figure 1: Power Law: The number of total tags versus the number of webpages having that many total tags.

content of the document. The basic idea is that the better the tag/term coverage of a document, the more we can trust its tag/term set. Exploiting the notion of tag/term coverage, we propose a *similarity* metric between two documents based on the tags and terms of both documents. Using the vector space model, we represent each document with two vectors: a tag vector and a term vector. When calculating the similarity score, other than the intuitive method of computing the cosine similarity between the two tag vectors and the two content vectors of the pair of documents, we also take into account the cosine similarity between the tag vector of one document and the term vector of the other document. The similarity metric of the two documents thus consists of a weighted sum of four components, with the weight of each component depending on the tag/term coverage of each document. Finally, using the similarity metric, we allow each document to *propagate*, or to share the tags it owns to other similar documents. After the propagation step, the tags that have a higher weight in a document is viewed as a trustworthy tag and thus may be a good candidate for tag recommendation.

We used tag data crawled from one of the largest social bookmarking sites, *Delicious*. Since users on *Delicious* can add bookmarks along with some descriptive tags into their own collection, this site contains an enormous number of bookmarks and each of which contains a different number of tags. We also crawled the webpages corresponding to each bookmark, which serves as the page content of the webpage. We analyzed the tag information of the dataset and tested our method using cross validation and through a user study. The results show that our proposed method is effective in populating untagged webpages with the correct tags.

2 Related Work

In the era of web 2.0, websites allow users to contribute their contents, and annotate them with a freely chosen set of keywords under the tagging system built by each website. Mika [2005] represented semantic social networks in the form of a tripartite model which is consisted of actors (users), concepts (tags), and instances (resources). Marlow *et al.* [2006] provided a simple taxonomy of tagging systems to analyze and distinguish these tagging systems from different kinds of websites by distinct facets. Golder and Huberman [2006] analyzed the dynamics of collaborative tagging systems, including user activity, tag frequencies, and bursts of popularity in resources.

From the aspect of information retrieval [Salton and McGill, 1986], tags bring new information to items over original contents [Bischoff *et al.*, 2008], and therefore tags can enhance capabilities of existing search engines to find out relevant documents [Heymann *et al.*, 2008a]. Bao *et al.* [2007] proposed iterative algorithms integrating tags into web search for better ranking results. Furthermore, tag types reflect what distinctions are important to taggers. Bischoff *et al.* [2008] refined the scheme presented by Golder and Huberman [2006] by classifying tags into 8 categories and exhibiting tag type distributions across different tagging systems and web anchor texts (or link labels). From comparing categories of tags with query logs and user study, they showed that most of the tags can be used for search, and that tagging behavior represents the same attributes as searching behavior in most cases.

Although tags are helpful to improve search results and divide documents, people on average annotate resources with only a small number of tags [Bischoff *et al.*, 2008]. Tag recommendation, one of the emerging research topics in tagging, can reduce people's tagging effort and encourage them to use more tags to reduce the problem. Xu *et al.* [2006] proposed the criteria for better tag recommendations, including content-based methods and temporal issue, but preliminary results only. Jäschke *et al.* [Jäschke *et al.*, 2007] introduced the FolkRank algorithm, which computes a topic-specific ranking of the elements in a folksonomy, and defeated collaborative filtering algorithms [Adomavicius and Tuzhilin, 2005].

In terms of content-based tag recommendation, Heymann *et al.* [2008b] formulated the problem into a supervised learning problem. Using page text, anchor text, surrounding hosts and available tag information as training data, Heymann *et al.* trained a classifier for each tag they wanted to predict. Even though they can achieve high precision using this method, the time required to train the classifiers for each tag becomes substantial when the number of distinct tags increases. In the same paper, Heymann *et al.* achieved good results in expanding the tag set of documents with little tag information using association rules.

3 Methodology

3.1 Notations

Before diving into the notations, we first define the terminology used in this paper. A *URL* is described by the *tags* annotated by the users and the *terms* which are words on the webpage corresponding to the *URL*. The words webpages and documents that appeared in the previous paragraphs are now all referred to as *URL*. The three words, *URL*, *tags* and *terms* will be used throughout the text.

Let U be the set of all URLs. Let T be the set of all tags. For a URL $u \in U$, let $Tag(u)$ be the set of tags that annotate

u and let $Term(u)$ be the set of terms that are in the webpage corresponding to URL u. Something important to point out is that for all terms $y \in Term(u), \forall u \in U$ but $y \notin T$ is removed, or in other words, all terms not in the set T are removed. Therefore the set of all terms is a subset of T. Using the vector space model, where each dimension corresponds to a tag, each URL $u \in U$ can thus be represented by two vectors of equal dimensions, a tag vector $\overrightarrow{v}_{u_t}$ and a term vector $\overrightarrow{v}_{u_m}$. Each dimension holds the number of times a tag/term appears in the URL.

We use $F_t(u, x)$ to denote the number of times tag $x \in T$ appears in the URL $u \in U$, which is stored in $\overrightarrow{v}_{u_t}$, and $F_m(u, y)$ to denote term frequency $y \in T$ in u, which is stored in $\overrightarrow{v}_{u_m}$. We also use $P_t(u, x)$ to denote the normalized $F_t(u, x)$:

$$P_t(u, x) = \frac{F_t(u, x)}{\sum_{x' \in T} F_t(u, x')}, u \in U, x \in T$$

$P_t(u, x)$ can also be viewed as the probability of tag x appearing in URL u. $P_m(u, y)$, the normalized $F_m(u, y)$, is computed in the same way.

3.2 Tag/Term Coverage

Tag/Term Significance Not all tags carry the same amount of information, thus the tag/term significance metric wishes to capture the amount of information a tag/term holds. We explain the concept using tags as an example, and the case is similar for terms. Tags that appear only in a specific group of URLs carry more information than those tags that appear in a wide range of URLs, thus the former tags should have higher tag significance than the latter tags. This is analogous to the IDF in the TF-IDF (Term Frequency-Inverse Document Frequency) metric [Salton and McGill, 1986], in which terms that appear in fewer documents are given more weight. Tags with higher tag significance scores are important because these tags help describe the URLs they annotate in a more accurate way.

We use the entropy of a tag to evaluate its tag significance score. Tags with higher entropy have lower tag significance scores. The entropy $Entropy_t(x)$ and tag significance $TS_t(x)$ of a tag x are defined as follows:

$$Entropy_t(x) = -\sum_{u \in U} P_t(u, x) \times \log(P_t(u, x))$$

$$TS_t(x) = 1 - Normalized(Entropy_t(x)) \ , \ x \in T$$

The values are normalized so that $TS_t(x) \in [0, 1]$. Entropy $Entropy_m(y)$ and term significance $TS_m(y)$ of a term y is calculated in the same way.

Tag/Term Coverage of a URL Tag coverage of a URL defines how well a URL can be represented by the tags annotating it. The tags that have low tag significance cannot as fully represent a URL as the tags having high tag significance do. Moreover, the URL that has few tags certainly cannot be well represented by its tags, for these tags, annotated by only a few users, may contain a high ratio of inappropriate tags that wrongly describe the URL and thus are not trusty. Therefore we include the total number of tags given to a URL in the tag coverage formula. Since the total number of tags given to a URL follows the power law, we take the logarithm of the number of total tags of a URL to shrink the difference between frequently tagged and infrequently tagged URLs. Tag coverage $TC_t(u)$ of a URL $u \in U$ captures the information above and is computed as follows:

$$TC_t(u) = Normalized(\sum_{x \in T} (P_t(u, x) \times TS_t(x) \times \log \sum_{x' \in T} F_t(u, x')))$$

The values are normalized so that $TC_t(u) \in [0, 1]$. The case is the same for the term coverage $TC_m(u)$ for a URL u.

3.3 Similarity between Two URLs

We utilize the two vectors, $\overrightarrow{v}_{u_t}$ and $\overrightarrow{v}_{u_m}$, of the URLs to compute similarity between each pair of URLs. We use the cosine similarity metric to calculate the similarity between two vectors. The cosine similarity of two vectors $\overrightarrow{u}$ and $\overrightarrow{v}$, $CosSim(\overrightarrow{u}, \overrightarrow{v})$, is defined as follows:

$$CosSim(\overrightarrow{u}, \overrightarrow{v}) = \frac{\overrightarrow{u} \cdot \overrightarrow{v}}{\|\overrightarrow{u}\| \|\overrightarrow{v}\|}$$

Since each URL is represented by two vectors, the similarity of two URLs should consist of four cosine similarities. The linear combination of the four cosine similarity between two vectors then becomes the similarity between two URLs. The weight of each cosine similarity is decided according to the tag coverage and term coverage of each URL respectively. Considering tag coverage as a factor is reasonable, because the tag vector of a URL that has high tag coverage is more descriptive of the URL than the tag vector of the URL having poor tag coverage. The same case applies for taking term coverage into consideration. The similarity is calculated as follows:

$$\begin{aligned} Sim(u, w) = \ & TC_t(u) \times TC_t(w) \times CosSim(\overrightarrow{v}_{u_t}, \overrightarrow{v}_{w_t}) \\ & + TC_m(u) \times TC_t(w) \times CosSim(\overrightarrow{v}_{u_m}, \overrightarrow{v}_{w_t}) \\ & + TC_t(u) \times TC_m(w) \times CosSim(\overrightarrow{v}_{u_t}, \overrightarrow{v}_{w_m}) \\ & + TC_m(u) \times TC_m(w) \times CosSim(\overrightarrow{v}_{u_m}, \overrightarrow{v}_{w_m}) \\ & u, w \in U \end{aligned}$$

Due to sparsity issues when calculating $CosSim(\overrightarrow{v}_{u_m}, \overrightarrow{v}_{x_m})$, we used PLSA [Hofmann, 1999], which is a dimension reduction algorithm, to lower the dimensions of the term vectors. Since the term vectors are extremely sparse, the reduction of dimensions will make the vector less sparse, thus making the results of the cosine similarity more accurate.

3.4 Tag Propagation

Now that we have the similarity for any two URLs, we can calculate the propagated weight of a tag for a URL. We use the term propagation because a URL gives another URL a tag if they are similar. In the following text, we will refer to the URL receiving the tag as the receiver, and the URL providing the tag as the provider.

The tag propagation formula consists of three parts. The first part is the similarity between two URLs. It is intuitive that the more similar the two URLs are, the more weight the provider will give to the receiver. The second part is the probability of the tag appearing in the provider. The third part takes into consideration how trustworthy the provider is. We use the total number of tags given to a URL to estimate the extent the provider can be trusted. URLs with more total tags annotated can be trusted more. Therefore, the formula is as follows:

$$Prop(u,x) = \sum_{w \in U} Sim(u,w) \times P_t(w,x) \times \log \sum_{x' \in T} F_t(w,x')$$
$$u \in U\,,\, x \in T$$

After we have the propagated weight of each tag in a URL, we can now recommend tags for this URL. For a URL u, we rank the list of tags to be recommended according to these tags' propagated weight. That is, the tag x with the highest $Prop(u,x)$ is ranked at the first place in the list and is therefore the best candidate for recommendation, and the tag that has second highest propagated weight is ranked second and so on. Note that for a URL already having prior tag information, the recommended tags may include the tags that are not in the original set of tags annotating the URL. These extra tags work as the supplements to the maybe insufficient tag information of the URL.

4 Evaluation

In order to evaluate the proposed content-based tag recommendation method, we performed experiments for the extreme case, where the URLs requiring tag recommendation have no prior tag information.

4.1 The Dataset

Our goal for this step is to construct a ground truth set required for evaluating our method.

First of all, we need to build a set of URLs with tag information. The URL crawling process starts from the *Popular Bookmarks* on the *Delicious* homepage. For each user who has bookmarked one of the popular bookmarks, his/her bookmarks are added to the set of harvested URLs. For each new URL, we look for previously unseen users who have bookmarked this URL, and look into his/her bookmark list for new URLs. Using this process, we crawled a total of 1,049,580 URLs, and collected all the tags given by any user. The user information is ignored for now, but may be valuable for personalized recommendations.

The *annotation frequency* of each tag is computed by tallying the number of users who have annotated any URL in the set with the tag. The top 100,000 tags in terms of annotation frequency are selected as the *tag/term vocabulary set*. We observed that tags with low annotation frequency are often misspelled words (e.g. wednsday) or concatenated words (e.g. onlinedesigntools), which are not the best candidates for tag recommendation. Consequently, any tag with annotation frequency of 15 or less is removed from the dataset.

The number of tags associated with each URL varies widely. In this dataset, documents with a total of 50 tags have an average of 17.96 distinct tags with a standard deviation of 5.18 tags. To ensure that our system is able to recommend up to 10 distinct tags for each webpage based on the ground truth dataset, any URL annotated with fewer than 50 tags by all users were removed from the set. In other words, the threshold is set at 50 so that most documents have more than 10 distinct tags as the ground truth.

However, given that most users tend to specify just a limited number of tags, many URLs have insufficient tag information even with the above filtering in place. In order to reduce the problem, the association rule algorithm [Agrawal *et al.*, 1993] has been adopted to expand the tag set of a URL. Experiments in [Heymann *et al.*, 2008b] showed that association rules may improve recall of single tag queries of URLs with little tag information. By applying all rules of length less than 3 and confidence 0.9, an average of 8.16 tags were added to each URL. One may wonder whether this expansion is reasonable, as adding associated tags may be "corrupting" the ground truth set. We will discuss this problem in the next section.

In the next step, all remaining URLs are filtered according to their *term information*. We crawled the page contents of the remaining URLs, ignoring files that were dead links or were not plain text. As the ability to deal with image content is beyond the scope of this research, URLs with a large number of images but few terms will be ignored for now, even though it may be relatively easy for people to tag these pages. The number of terms on each page is counted after the HTML tags are removed. Any URL with fewer than 50 or more than 100,000 terms in the corresponding content is eliminated from the dataset.

As both the number of tags and the number of terms associated with each URL follow the power law, the majority of the crawled URLs are filtered out in the process. As a result, the filtering process netted 85,230 URLs from the initial set. Each URL in the final da taset has on average 1084.33 total tags, 99.97 distinct tags, 1043 total terms, and 364.19 distinct terms.

4.2 Prediction Accuracy by Cross Validation

The first set of experiments look into the effectiveness of our system in propagating the accurate tags to the untagged URLs. In our 5-fold cross validation process, we removed one-fifth of the URL's tags, and treated the remaining information (tag information for 4 folds and all the term information) as input into our system. After the propagation phase, the URLs that had their tags removed would be repopulated with tags, and we compare the repopulated tags with the ground truth for precision. Since users usually do not have the patience to annotate many tags for a URL, we only calculated up to precision at 10. During the training process, the only parameter we had to select is the similarity threshold ϵ. All pairs of URLs with similarity less than ϵ are removed and not used in propagation. In addition, we compared our results with the results from term PLSA, which calculates the similarity between two URLs by calculating only the cosine similarity between their respective term vectors. Note that the dimensions of the term vectors in our method and term PLSA have already been reduced to 500 dimensions using PLSA. The propagation method used in the term PLSA is the same as the one used in our method. The results of our algorithm

Algorithm	ϵ	P@1	P@3	P@5	P@10
Ours	0.02	0.69	0.63	0.59	0.55
Ours	0.03	0.66	0.59	0.56	0.51
Ours	0.04	0.62	0.55	0.51	0.46
Term PLSA	N/A	0.56	0.50	0.47	0.42

Table 1: Cross Validation Results: Results of term PLSA and our algorithm at different ϵ threshold levels.

are shown in Table 1.

The precision is highest when $\epsilon = 0.02$, with precision at 1 being 69% and precision at 5 being 59%. The precision at 1 for term PLSA is 56% and precision at 5 being 47%. We also compared precision at 5 of our algorithm with that of term PLSA in each fold. In all folds, our algorithm, using ϵ ranging from 0.2 to 0.4, consistently achieved higher precision than term PLSA. Our results showed that the proposed algorithm outperforms term PLSA. This is because our algorithm, in addition to calculating the term vector cosine similarity used in term PLSA, also takes into account the similarity between one URL's tag vector and the other URL's term vector. In addition, it is worth noting that the lower the ϵ similarity threshold, the better precision we are able to obtain.

Associated tags were added to the dataset due to the observation that the user-specified tag set is often incomplete. However, this step could potentially be "corrupting" our dataset. Our experiments showed that 95.52% of the correctly guessed tags are original, and only 4.48% matches the associated tags. Therefore, we can deduce that the effect of "corruption", if any, would be minimal.

For the URLs that had unsatisfactory tagging results, we found that they can generally be attributed into three general causes. The first cause happens when the URL is a frequently updated webpage, e.g. daily news, so the tags for that page would only be relevant to the contents of the page at the time the page content was crawled. Another situation is when the actual page contents of the webpage is relatively small compared to the advertisements on the webpage. Advertisements are irrelevant to the contents of the page, yet they would severely distort the term vector on short webpages. Parsing HTML tags to remove sidebar information may be a solution to this problem, but a myriad of webpage and sidebar formats adds to the difficulty of implementing this solution. Finally, webpages talking about a relatively rare topic (for example, spinning pens) would have few similar webpages, thus resulting in decreased relevance of recommended tags.

4.3 Results of User Study

In the second set of experiments, we gathered data on the precision of system-recommended tags through an user study. The user study was conducted as follows. A participant was first shown a webpage and asked to view the contents of the page. Next, the participant was asked to key in some tags that he/she will use to annotate the page. Once the participant confirms that he/she has completed the above step, the system will show the participant the five best recommended tags provided by our algorithm and ask the participant to mark the tags they deem irrelevant to the website. We allowed the participant to apply tags to the webpage first before viewing the recommended tags because recommended tags usually affect the user's original intent on which tags to use [Suchanek *et al.*, 2008].

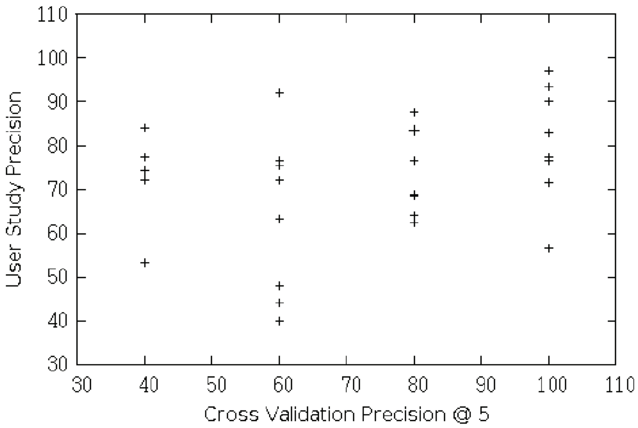

Figure 2: Cross Validation and User Study Precision: Precision from the two evaluation methods are not strongly correlated.

A total of 19 people, all with computer science background, participated in the experiment outlined above on 31 randomly picked URLs from the testing set of the first fold of cross validation, and we collected at least five responses for every URL. The system recommended tags for the 31 URLs were obtained from the first fold of cross validation for the $\epsilon = 0.02$ experiment. Results from our user study show that 32% of the tags the participants provided matches the top five tags the system recommends. 69.45% of the tags the system recommended were also marked as relevant by the participant.

After analyzing the results of the user study, we discovered that some webpages have more tags marked as irrelevant than other webpages, which may be because the tags recommended by the system were of poor quality. We would also like to know whether the webpages that did poorly in the user study also achieved a low accuracy in the cross validation experiment. Therefore, using the results from the cross validation experiment, we calculated the precision at 5 for each of the 31 URLs, and the average precision is 70.97%, which is remarkably similar to the results obtained from the user study (69.45%). Since the obtained average precisions were very similar, we would like to know whether there is any correlation between the precision of the URLs obtained from the two methods of evaluation.

Figure 2 is the scatter plot showing the precision obtained from cross validation versus the precision obtained from the user study. The correlation is found to be 0.47. A low correlation means that the same URL achieves different accuracies in different methods of evaluation. For those cases where a URL achieves better in the user study than the cross validation, we can conclude that the ground truth set is incomplete, because there are still some relevant tags missing in the ground truth set, causing the precision to drop. For the other case, where a URL achieves better in cross validation than the user study, this means that our ground truth set has some noise. These noise may come from personal preferences of users: a relevant tag for one user may not be a relevant tag for another. Another reason may be irrelevant tags expanded by the association rules.

5 Conclusions and Future Work

In this paper, we proposed a method for content-based tag recommendation which can be applied to webpages with or without tags. Since there are many webpages on the web without any tag information, we focused on trying to populate tags for webpages with absolutely no tag information. Using data crawled from *Delicious*, the precision @ 5 of our method is 59% when evaluated with 5 fold cross validation. We also picked 31 webpages for user study, and 32% of the tags given by the participants matches the top 5 recommendations our method recommends. 69.45% of the tags recommended by the our method is marked as relevant by the participants. The above results show that our method is acceptable in terms of precision, but there is still some room for improvement. Some pages are inherently not suitable for content-based recommendation, for example news webpages which are updated everyday. Also, if we could find the more important terms of the webpage, and not be affected by the links or advertisements on the sides of the webpage, our precision can improve.

Other than improving our precision, there is also a lot of room for improvement in trying to recommend useful tags. The tags our method recommends are usually those popular tags, which sometimes are quite vague and not as informative as specific tags. Therefore one of the things we can do in the future is to recommend tags of different popularity to the user or webpage. Popular tags tend to be more vague and can be given to a variety of webpages, and not as popular tags tend to be more descriptive and maybe more difficult to recommend since it is only relevant to a smaller amount of webpages. Another interesting future work is to recommend tags with different types defined in [Bischoff *et al.*, 2008]. Recommending tags with a variety of types: topic, time, location, usage/context and so on may help in providing more useful information for webpages. And finally, since the users are the ones who are using the tags, therefore personalized content-based tag recommendation will be another interesting future work.

References

[Adomavicius and Tuzhilin, 2005] Gediminas Adomavicius and Alexander Tuzhilin. 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, 2005.

[Agrawal *et al.*, 1993] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[Bao *et al.*, 2007] Shenghua Bao, Xiaoyuan Wu, Ben Fei, Guirong Xue, Zhong Su, and Yong Yu. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World wide web*, pages 501–510, 2007.

[Bischoff *et al.*, 2008] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 193–202, 2008.

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

[Heymann *et al.*, 2008a] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, 2008.

[Heymann *et al.*, 2008b] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538, 2008.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[Jäschke *et al.*, 2007] Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In *PKDD 2007: Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, pages 506–514, 2007.

[Marlow *et al.*, 2006] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, 2006.

[Mika, 2005] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *ISWC '05: Proceedings of the 4th International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536, 2005.

[Salton and McGill, 1986] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[Suchanek *et al.*, 2008] Fabian M. Suchanek, Milan Vojnovic, and Dinan Gunawardena. Social tags: meaning and suggestions. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 223–232, 2008.

[Xu *et al.*, 2006] Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the Collaborative web tagging workshop at WWW 2006*, Edinburgh, Scotland, 2006.