SocialTagger - Collaborative Tagging for Blogs in the Long Tail

Shankara B Subramanya and Huan Liu
Department of Computer Science and Engineering
Arizona State University, Tempe AZ-85287 USA
{shankarbs,huan.liu}@asu.edu

ABSTRACT

Social bookmarking is the process through which users share tags for online resources like blogs with others. Such collaborative tags provide valuable metadata for retrieval systems. While the successes of collaborative tagging systems have been demonstrated by popular websites like Del.icio.us, these sites cover only a small fraction of the available blogs on the web. The vast majority of the blogs are not available on any collaborative tagging system and are often tagged only by the authors. This lack of coverage of collaborative tags is a considerable roadblock in using the tag metadata in a web scale information retrieval system. To solve this problem we propose and implement a system to automatically recommend collaborative tags for a blog. The automatically generated tags will help to surface the blogs by making them available on social book marking sites and allow them to be easily discovered and potentially further tagged by a wider population.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Blogs, Collaborative Tagging

1. INTRODUCTION

With the advent of Web 2.0 vast amounts of user generated content have become available on the web. The ever growing number of people contributing on the web has resulted in a virtual explosion of the content. Users contribute not only in posts and articles but also in the form of tags which form the metadata of the content. The metadata of tags provide valuable information for making the content easily accessible and searchable. Tagging brings some of the benefits of the semantic web to the current websites. These tags can either be provided by the author of the content or by other users of the content. The latter process is called as Collaborative Tagging [13, 8] or Folksonomy. In Collaborative(social) Tagging, the users of the content create and share tags to annotate the content which provide valuable indices for navigation and discovery. An example of such a collaborative system on the web which has become widely popular is Del.icio.us or Delicious[1]. Delicious allows users to collaboratively tag webpages, blogs and other resources on the web. Stumbleupon[2] is another of such collaborative tagging systems using which users can tag webpages, images and videos, and also receive recommendations. As opposed to folksonomies some systems allow only the content creator or a designated expert to annotate the content. For example in Youtube.com[3], the videos can be tagged only by the submitter of the video. Flickr.com[4] provides a restricted form of collaborative tagging. While collaborative and individual tagging is applicable to a wide variety of content types, in this work we are concerned with tagging of blogs.

A larger proportion of blogs are individually tagged compared to those that are tagged collaboratively. However tags given by a single individual suffer from the vocabulary problem [6, 4], which is the problem of variability in word usage by two or more people while referring to the same concept. The vocabulary problem has been a well studied problem in the traditional information retrieval. Context, subjective judgment and cultural background play an important role in how language is used by an individual. Different people may read the same blog post and assign different tags to it based on their culture, background, training and experiences. Specifically the following types of problems may occur with individual provided tags [7].

1. Polysemy - Different bloggers can refer to different concepts with the same tag. For example when users in the US use the tag 'football' they probably would probably be referring to 'American Football', but when users in other parts of the world use the tag 'football' they would most likely be referring to 'Soccer'.

2. Synonymy - Different bloggers can refer to the same concept with different tags. Examples of such tags

[1]http://del.icio.us
[2]http://www.stumbleupon.com
[3]http://www.youtube.com
[4]http://www.flickr.com

SSM'08, October 30, 2008, Napa Valley, California, USA.

would be 'blogging' vs. 'blog' and 'Money' vs. 'Finance'.

3. Level variation - Individual bloggers may choose to tag their blogs at different levels. For example while talking about US presidential elections some bloggers may chose to tag it at a high level with the tag 'Election' while some may tag it with the individual party names as 'Democrat' and 'Republican', while some other bloggers may tag it with individual candidate names as 'Clinton', 'McCain' or 'Obama'.

4. Cultural and contextual variations - This is more general and may be one of the reasons for causing the other 3 types of problems.

These variations in the vocabulary of users cause terminological variations in the tags. The problem is exacerbated when only a single user tags the content. In this case an exact match between the tags given by the individual and the search keywords used by other users would be unlikely. Also if the individual who tags the blog is the author then the tags may not be very objective because of the familiarity of the author with the topic. These tags could be idiosyncratic and very specific to the content. For example, 'just another email request for today' or 'tech talk tuesdays'. Such tags may not be of much help for others in discovering and retrieving the content if the choice of words of the given author is different from the ones used by other users.

2. COLLABORATIVE TAGS

Table 1 shows the main differences between the individual tags and collaborative tags. Individual tags act as book marking resource for oneself. Collaborative tagging systems extend this by connecting the individual book marking activities of users into a collection of tags for resources shared among multiple users. Collaborative tagging systems transform a solitary browsing experience into a social one [16]. It has the advantage of multiple people looking at the same content. Collaborative tagging systems allow the users to freely choose any free-text tag entry. This freedom plays a critical role in the success of such systems. Without having to rely on a standardized corpus of tags users can quickly adapt and include novel tags. A controlled vocabulary would reduce this dynamism of tagging systems and would also require experts in each domain to define the standardized corpus. With multiple users accessing and freely tagging the same content the difference in the multiple users' culture, background and experiences would ensure the expansion of the vocabulary. This would also ensure that one perspective from a single user does not dominate. It would then be more likely that the keyword used by the searcher will match one of the tags assigned by a user. Also, over time the most dominant concept(s) in the content would emerge and get reinforced and the tags indicating these concepts would become prominent. The tags would thus aggregate the 'collective conscious' of all the users. Thus the above mentioned problems of tags given by a single user would be attenuated in collaborative tagging systems. The tags in such collaborative tagging systems would act as valuable indices in making the content searchable.

3. BLOGS IN THE LONG TAIL

While collaborative tags have significant advantages, only a very small percentage of blogs are actually discovered by other users and are tagged. This lack of coverage of social tags is a significant obstacle in using them for web-scale search applications [10]. The popularity of the blogs in terms of their readership or connection to other blogs has a power law distribution with a characteristic long tail [12, 15]. The blogs in the short head of the distribution are usually discovered and tagged by other users on websites like Delicious while a large number of blogs which are present in the long tail might never be discovered by other users without additional assistance. These blogs in the long tail would be usually tagged only by the author and would not be available on social bookmarking websites. This problem is aggravated by the rich-get-richer effect seen on the web [20].

In the blogosphere the rich-get-richer effect or 'cumulative advantage' is observed because of the social influence condition. The social influence condition in the collaborative tagging systems is that users when searching for blogs would get as a result only those blogs which have already been tagged by other users. So the first few users in the system play a huge role in setting the trends of the system and influencing people who arrive later in the system. The blogs which are tagged by the initial users end up being returned in the searches by later users who themselves tag these blogs. Due to this phenomenon the popular blogs could become even more popular over time and the unpopular blogs become even less popular. In some cases blogs tagged by initial users for some random reason may end up in the short head of the distribution even though they may not be among the best. Many of the blogs that no one has an opportunity to read and then tag can remain in the long tail even though they could be a better result for a given search query.

In order to improve overall blogging experience, it is vital for achieving a critical mass of blogs available for collaborative tagging, which in turn would allow for the web-scale use of metadata information for improved search, or targeted ads. Thus, providing an opportunity for the blogs in the long tail to be accessible by other users necessitates that they can *surface*. One way is to annotating them with tags similar to the collaborative tags. When the collaborative tags for these blogs are available, they can be automatically included in sites like Delicious where they would be easily discovered and further tagged by other users. Surfacing the long-tail blogs will help vastly increase the number of collaboratively tagged blogs.

4. RELATED WORK

There have been several systems that automatically suggest tags for blogs. Brooks et. al. [3] developed a system to automatically suggest tags for blogs. They extracted three terms with the top TF-IDF scores from each post and suggested them as tags. While the tags suggested by this system were useful in obtaining more focused clusters of blog posts, the use of the author-given tags as indices was limited since the tags were restricted to terms which literally appeared in the blog posts.

Xu et. al. [21] developed an algorithm for making tag suggestions based on the criteria of high coverage, least effort and high popularity. The basic idea behind their algorithm was to iteratively select the tags with highest additional contribution given the already selected tags. However to make

Individual given tags	Collaborative tags
Goal is mainly to book mark and to a lesser extent to share	Goal is mainly to share and to a lesser extent to book mark
Suffers from vocabulary problem because of single user	With multiple users providing tags vocabulary problem is reduced
Subjective - Affected by user's knowledge, experience, cultural background etc	More objective - Incorporates perspectives from many users
Narrower coverage of different facets	Wider coverage of different facets

Table 1: Individual and collaborative tags

the tag suggestions their system used tags which were assigned to the same resource but by other users. This approach would not be of much help in the case of long-tail blogs.

Mishne [14] developed a system called 'AutoTag' which finds similar blog posts to the given blog post and suggests tags from these posts to the user for selection. Sood et. al. [17] developed 'TagAssist' which improved the 'AutoTag' system by using tag compression and case evaluation to filter and rank tag suggestions. The idea behind these systems was to enforce a controlled vocabulary to solve the vocabulary problem. However allowing every blogger to tag independently is invaluable because by doing so a sort of collective wisdom emerges, and it allows capturing individual conceptual associations. If a user is confined to choosing from a predefined set of tags the user would be in effect following the group norm. This would again result in the problem where the first few users giving the tag have disproportionately large influence. The diversity of the tags given by the individual is required to obtain a complete conceptual representation and to influence the group norm [18]. When these freely given tags naturally converge to a shared vocabulary the consensus on the main concepts of the blogs would be captured. This will not be possible when the user's choice is limited.

While we too suggest new tags by finding similar tagged posts, our goal is different from that of the above mentioned systems since we aim to facilitate the generation of collaborative tags for the long-tail blogs. An important difference between our system and the previous systems is in the way similar posts are found. Autotag and TagAssist find similar posts by generating a query from the new post and using it with an information retrieval system. The query is generated by extracting distinctive terms from the new post. Once the similar posts are found are the the tags from these posts suggested to the user. We, however, find the similar posts by taking into account the semantic relationship between blog posts and the user given tags. This will be explained in more detail later. In addition, we not only suggest new tags, but we also keep the existing tags freely chosen by the author. While this may seem to be a small difference, it is critical in growing the set of tags for a blog and achieving the advantages of collaborative tagging.

Finally the evaluation of the suggested tags is an important problem in such systems. Objective evaluation of the suggested tags is a challenging problem. Researchers either compared the tags with actual author-given tags or employed some (usually a small number of) human evaluators to conduct a manual evaluation or just gave examples of generated tags. None of these methods are ideal for the evaluation of collaborative tagging. The first method suffers from the problem that the author-given tags themselves are imperfect as mentioned previously. The human evaluation would also not be a good predictor unless the evaluators are from diverse backgrounds and a large number of evaluators are employed. One of our contributions in this work is that we propose to solve the evaluation problem by considering overlapping data from two independent online sources so as to pave the way for objective and automated evaluation of collaborative tagging. This will be explained in more detail in Experimental Evaluation.

5. THE SOCIAL TAGGING PROBLEM

In this section we formally define the problem of social tagging. Let b be a blog written by a blogger. We consider only individual blogs in this case and not community blogs which are written by multiple users. The blog b would consist of a set of blog posts about the topics of interest of the blogger. The blogger usually assigns a set of tags that represents the blog. So each blog can be represented by two sets of features:

1. The set of tags assigned by the blogger, and
2. The blog posts or the content of the blog

Each blog can be therefore be represented by a tuple

$$b =< tg, p >$$

where tg represents the feature set corresponding to the author-given tags and p represents the features corresponding to the blog posts. The features can be extracted from the tags and blog post text using the familiar bag-of-words approach. Let B represent the set of all blogs under consideration. The set could represent the collection of all blogs in a blog community for which we are interested in finding collaborative tags. Given B, the tags and the blog post features for all the blogs can be represented as two matrices TG and P.

$$B =< TG, P >$$

where TG and P represent blog-term matrices corresponding to the given tags and the blog posts, respectively. The i^{th} row in TG and the i^{th} row in P correspond to tg_i and p_i from the i^{th} blog $b_i =< tg_i, p_i >$, respectively. tg_i can be $\emptyset$ if the i^{th} blogger has not given any tags to the blog.

Given the initial tags tg_i and the blog posts p_i for each blog b_i the problem of Social Tagging is to find the set of tags ts_i such that ts_i satisfy the properties of the collaborative tags as defined in column two of Table 1. If TS represents the blog-term matrix obtained from the collaborative tags ts for all the blogs, then the new representation of the blog can be given as follows.

$$B =< TS, P >$$

6. AN APPROACH TO SOCIAL TAGGING

To solve the above defined problem a tagging system would have to expand the given set of tags for each blog b_i with a new set of collaborative tags. Since one good source for new tags is the pool of tags from other blogs, the collaborative tags can be extracted from other blogs in the set. To extract only relevant tags and to cover as many aspects as possible, for a given blog b_i, a subset of other blogs which are highly similar to it can be chosen and the tags from them can be extracted. To find the similar bloggers, the tagging system can take advantage of all types of available information for each blog (e.g., blog posts, author-given tags and relationships between them). We propose the following key steps in finding collaborative tags based on which we develop a tagging system called SocialTagger.

1. Similarity ranking: Rank blogs based on similarity
2. Candidate tag extraction: Extract candidate tags from top ranked blogs
3. Tag selection: Select the collaborative tags from candidate tags

6.1 Similarity ranking

To rank the blogs based on similarity we consider not only the blog posts but also the tags given by the blogger and the relationship between these two sets of features. To do this we use the multivariate statistical technique of canonical correlation analysis. The idea behind this technique is that the blog text and the tags associated with it can be considered as two different views of the same blog. So there would be a semantic relationship between the tag view and the blog text view (intuitively this could be the topic of the blog). Canonical correlation analysis finds a new representation for the original feature vectors such that the correlation between the two sets of features is maximized in the semantic feature space. Since our goal is to find the tags given the blog post this semantic feature space representation would be ideal to match the tags with the blog post.

Canonical Correlation Analysis (CCA)

CCA attempts to find basis vectors for the two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized [11]. The correlation between the two sets of variables may not be visible in their original coordinate system. CCA tries to find a linear transformation for the two sets variables such that in the transformed space they are maximally correlated.

The canonical correlation between any two data sets is defined as

$$\rho = max_{W_x,W_y}\ corr(F_x \cdot W_x, F_y \cdot W_y)$$

where F_x and F_y are the two sets of variables, and W_x and W_y are the basis vectors onto which F_x and F_y are projected, respectively. The equation for ρ can be rewritten as

$$\rho = max_{W_x,W_y}\ \frac{(F_x \cdot W_x, F_y \cdot W_y)}{(||F_x \cdot W_x|| \cdot ||F_y \cdot W_y||)}$$

The problem of finding ρ is therefore an optimization problem with respect to W_x and W_y. This optimization problem can be formulated as a standard Eigen problem [9] which can be easily solved. Since W_x and W_y are always calculated to maximize the correlation of the projections, CCA is independent of the original coordinate system unlike other correlation analysis techniques. There may be more than one canonical correlation, each representing orthogonally separate pattern of relationship between the two sets of variables. The correlations for the successively extracted canonical variates are smaller and smaller. When extracting the canonical correlation the eigen values are calculated. The square root of the eigen values can be interpreted as the canonical coefficients. Corresponding to each canonical correlation the canonical weights for each of the variable in the data set is calculated. The canonical weights represent the unique positive or negative contribution of each variable to the total correlation.

CCA has been used previously by researchers to find the semantic relationship between multimodal inputs. Hardoon et. al. [9] used kernel CCA to find correlation between image and text features obtained from a webpage and used it for content based image retrieval. Vinokourov et. al. [19] used CCA to find the language independent semantic representation of a text by using English text and its French translation as two views. When two multidimensional variables represent the two views of the same object, then the projections found by CCA can be thought of as capturing the underlying semantics of the object. In other words we can say that in the semantic feature space, the different views of the object are highly correlated. So to acquire a missing view of an object we can select the closest match from the observed views of other objects, such that it has maximum correlation with the non-missing views of the current object in the semantic feature space. Now we present the procedure to rank the blogs based on similarity using CCA.

Let $CCA(TG, P)$ denote the canonical correlation analysis of matrices TG and P, which returns the basis vectors and the projections of TG and P on the basis vectors. The basis vectors can be considered as representing the feature space which captures the underlying semantic relationship. Therefore we can rank the blogs which are similar to a blog b_i based on the correlation between p_i and tg_j in the lower dimensional semantic feature space where tg_j is the author-given tag feature vector of blog $b_j < tg_j, p_j > \in B, j \neq i$. The procedure to rank blogs similar to blog b_i is as follows.

1. Perform the canonical correlation analysis between TG and P and find the basis vectors.
   $[A, B, U, V] = CCA(TG, P)$,
   where U and V are the matrices whose columns represent the basis vectors corresponding to TG and P respectively
   A and B are the projection of TG and P onto U and V respectively

2. For each blog $b_j \in B$, where $b_j =< tg_j, p_j >, j \neq i$,
   Project tg_j onto U and p_i onto V
   $st = tg_j * U_E$
   $sp = p_i * V_E$
   where U_E and V_E are obtained by selecting top E basis vectors from U and V, respectively.
   Calculate Pearson $cor_{i,j}$ between st and sp
   $cor_{i,j} = correlation(st, sp)$

3. Rank the blogs $b_j \in B, j \neq i$ based on $cor_{i,j}$ in non increasing order

6.2 Candidate tag extraction

Once for the given blog b_i all the other blogs are ranked based on the similarity, the candidate set of tags is obtained

by taking the union of the tags from top K similar blogs. The author-given tags tg_i are also added into the candidate set, which is a critical in growing the set of tags for a blog to fully achieve the advantages of collaborative tagging. If any of the author-given tag is a phrase, then it is split into its constituent words, all the stop words are removed and the remaining words are added to the candidate set. Let this candidate set of tags be represented by tc_i.

6.3 Tag selection

Once the candidate tags are extracted, the final set of tags for the given blogs b_i are selected from this candidate tag set tc_i. The selection of the final set of tags ts_i is done by considering two properties of each individual tag in tc_i, namely, Tag re-occurrence and Tag co-occurrence.

Tag re-occurrence: The most important advantage of the collaborative tagging as mentioned before is that since many people tag the same blog, the vocabulary problem is attenuated. Collaborative tagging therefore helps in enforcing consistency among tags with the most representative tags being used by many users. Figure 1 shows the frequency distribution of the tags from the blogs collected in our experiments. Among the tags there were a very large number of niche tags like 'tech talk tuesdays' which are used only by a single user. To avoid selecting these tags we use the property of tag re-occurrence of a tag. This property measures the number of authors who have used the given tag in their blogs. By selecting only those tags which have a minimum tag re-occurrence value of $ReOccur\_Cutoff$, we are able to select a better set of tags. The re-occurrence cutoff also helps in eliminating spam tags since a tag used by many people would not likely be a spam. The value of the $ReOccur\_Cutoff$ parameter is determined experimentally.

Tag co-occurrence: Tag co-occurrence of a pair of tags measures the number of blogs in which a pair of tags occurs simultaneously. A high co-occurrence of two tags would reasonably indicate that the two tags are related or have similar concepts. Tag co-occurrence has been used previously for tag clustering [2]. Sood et. al. [17] showed the effectiveness of tag co-occurrence in validating tags. Tags which are from the same topic would usually have high tag co-occurrence. So to validate the candidate set, we find the co-occurrence of each of the candidate tag with each of the tags given by the author. The candidate tag is selected if any of the co-occurrence value is greater than a predefined threshold value $CoOccur\_Cutoff$.

7. EXPERIMENTAL EVALUATION

As defined in Section 5 our goal is to generate a set of tags which would be similar to the tags generated by collaborative systems and would have the properties of collaborative tags as defined in column two of Table 1. However, verifying whether a set of tags satisfy the properties of collaborative tags has practical challenges since the evaluation of the system-generated tags requires the actual collaborative tags as the ground truth. In other words, if such real-world collaborative tags were available for blogs, it would have been possible to directly compare the generated tags with them. First let us examine what is needed for experiments, and then we will propose a solution to address the ground-truth challenge.

An empirical study requires blog data including (1) blog posts, (2) author given tags, and (3) collaborative tags.

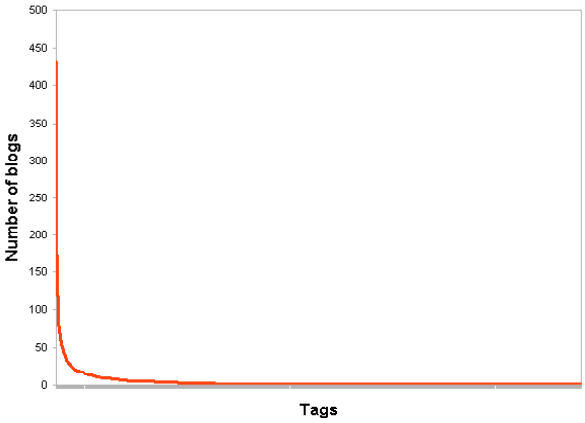

Figure 1: Tag Frequency Distribution

Number of blogs from BlogCatalog	2000
Number of selected blogs in Delicious	100
Number of distinct author given tags	10577

Table 2: Blog data statistics

However, such blog data is not readily available from any single source on the web. While many large blog datasets are available, none of these datasets have the collaborative tags for a blog. Popular blog search engines and blog directories like Technorati[5] and BlogCatalog[6] collect information from a large number of blogs. However these systems aggregate only the author given tags. The collaborative tags are available only from the social bookmarking sites like Delicious. Therefore we need to collect overlapping data from the two different and independent sources for the same set of blogs. Below we first detail the data collection from each of the two sources, and then report the evaluation results.

The precision (P_i) and recall(R_i) for tagging method M for a given blog b_i are calculated as follows.

$$P_i = \frac{Number\ of\ tags\ from\ M\ matching\ with\ Delicious\ tags\ for\ blog\ b_i}{Total\ number\ of\ tags\ from\ M\ for\ blog\ b_i}$$

$$R_i = \frac{Number\ of\ tags\ from\ M\ matching\ with\ Delicious\ tags\ for\ blog\ b_i}{Total\ number\ of\ Delicious\ tags\ for\ blog\ b_i}$$

The final precision (P) and recall (R) for tagging method M is obtained by taking the average of the precision and recall for M on all the n blogs. The F-measure (F) is obtained by taking the harmonic mean of the precision and recall values.

$$P = \frac{\sum_{i=1}^{n} P_i}{n}, \quad R = \frac{\sum_{i=1}^{n} R_i}{n}, \quad F = \frac{2PR}{P+R}$$

7.1 Data collection

7.1.1 BlogCatalog

The best way to collect blog data from individual authors is using blog directories. BlogCatalog is one such social blog directory in which blogs are listed in multiple categories. It is an open directory where bloggers themselves can list their blogs and promote it. Each blog in the blog directory is

[5]http://technorati.com
[6]http://www.blogcatalog.com

Tagging Method	Precision (P)	Recall (R)	F-Measure (F)
SocialTagger	30.57	34.12	32.25
Latent Semantic Analysis	24.84	33.81	28.64
Author-Given Tags	14.74	33.78	20.52

Table 3: Precision, Recall and F-Measure values for tags from different methods

associated with one blogger who is the author of the blog. When bloggers list their blogs, they list it under predefined categories and they can also assign tags describing the blogs. Once the blogger lists his/her blog on the site, snippets from the blog posts are automatically extracted and are displayed on the BlogCatalog website. Each blog post in a blog may also have tags assigned by the blogger. The union of tags assigned to the blog post and to the blog itself form the single set of tags. These tags along with the blog post text give the representation of a blog as defined in Section 5. BlogCatalog provides APIs to collect the blog post and tag data along with the blog name, blogger name and blog URL. Each blogger in BlogCatalog also specifies a list of friend bloggers. This network of friends forms a connected graph. We traversed this graph in a breadth-first manner to get a list of the bloggers and their blogs. To have a diverse collection of blogs from multiple categories, we chose seed blogs from six different categories, resulting in 2000 individual blogs.

7.1.2 Delicious

To collect the collaborative tags for the blogs from BlogCatalog, we searched for these blogs on the social book marking site Delicious. Delicious is one of the popular social book marking sites on the web with more than a million users. However as mentioned before only a small percentage of blogs which are likely in the short head of the distribution will be actually discovered and tagged by users on the web. This was confirmed by our search for BlogCatalog blogs on Delicious. Among the blogs collected from BlogCatalog we found that only 5% of the blogs were actually tagged and hence available on Delicious. These tags were downloaded and were used as the ground truth for our evaluation. Table 2 gives the statistics of the collected data.

7.2 Experiments and Results

Experiments were conducted using the blogs collected from BlogCatalog. SocialTagger is implemented in Matlab and can be made available upon request. To evaluate the collaborative tags by SocialTagger, we compared them with the collaborative tags obtained from Delicious. We used exact string matching with stemming in comparison. We also compared the original author-given tags with the collaborative tags from Delicious. Since the standard 'Latent Semantic Analysis' (LSA) [5] can handle the synonymy and polysemy problems and perform better than alternative measures such as cosine similarity, we employ LSA as the baseline algorithm in comparative study. LSA finds document similarity by performing the Singular Valued Decomposition on the document-term matrices. So for finding similar blogs LSA uses only the blog post text and not the author given tags. The blogs are ranked by similarities based on the correlation in the latent semantic space. Except for using LSA for calculating similarity, all the other steps of the tagging process are similar to those of the SocialTagger method.

Table 3 gives the precision, recall and F-measure values

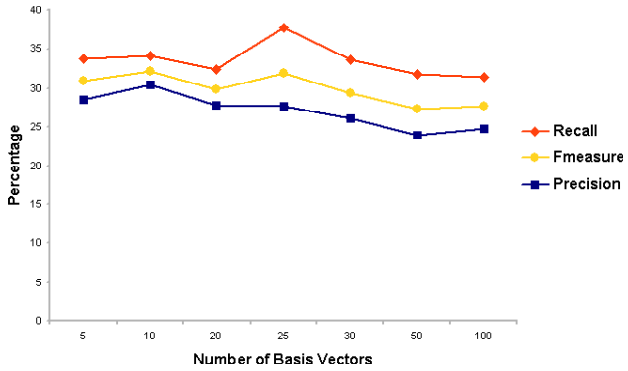

Figure 2: Effect of number of basis vectors on generated tags

for each of these cases. Because of the nature of the problem itself and since absolute string comparison is used which does not consider synonymy, the absolute values of precision and recall are not very high. In the tagging problem high agreement of assigned tags even among humans is rarely possible. But comparing the results from the table we can see that the SocialTagger method achieves significant improvements in both precision and recall values over author-given tags. This clearly shows the advantage of the SocialTagger system in predicting collaborative tags. The advantage of using CCA for measuring similarity between blogs is also seen with SocialTagger showing better precision and recall values compared to latent semantic analysis.

Table 4 shows two example blogs along with the author-given tags, suggested tags and the delicious tags. From these examples we can see that the author-given tags contain many highly specific words and phrases like 'watch mail' and 'redhat'. SocialTagger however is able to find more general tags like 'tutorial', 'blog', and 'marketing' which are similar to the tags obtained from Delicious. Tags found by LSA are worse than those found by SocialTagger.

7.3 Effects of parameters

The SocialTagger algorithm has four tunable parameters, namely, the number of basis vectors E, number of similar blogs selected to get the candidate tags k, $ReOccur\_Cutoff$ and $CoOccur\_Cutoff$. The values for these parameters were experimentally determined.

E is the number of basis vectors used for constructing the semantic feature space as indicated in step 2 of the procedure to rank blogs. Figure 3 shows the effect of basis vectors on the precision, recall and f-measure values. The best results were obtained when E was set to a value of 10.

Figure 3 shows the effect of the number of similar blogs selected to obtain the candidate tags, on the accuracy of the resulting tags. As the number of selected blogs increases the recall value also increases while the precision value decreases. This is because as more blogs are included, the size of the

Blogger	delicious tags	Author-given tags	SocialTagger tags	LSA tags
Marier	tutorial, blog, commands, sysadmin, linux, tips, network, howto	fedora, linux, watch mail, redhat, proactive monitoring, sysad, delete horizontal, server, watch diskspace, google talk with psi, linux internet messaging, watch files using ls	tutorial, consulting, marketing, blog, linux	server
lyndoman	social, networking, seo, forum, marketing, blog	stumbleupon, linkbait, smo, seo, cornwall seo, digg social media	blog, internet, news, general, tips, networking marketing, stats, business, branding, reviews, rating services, smo, stumbleupon, seo	freebies, video, money, internet, online, smo stumbleupon, seo

Table 4: Example tags

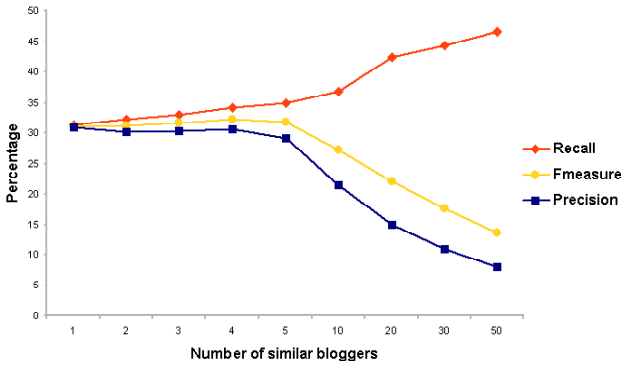

Figure 3: Effect of number of similar blogs on generated tags

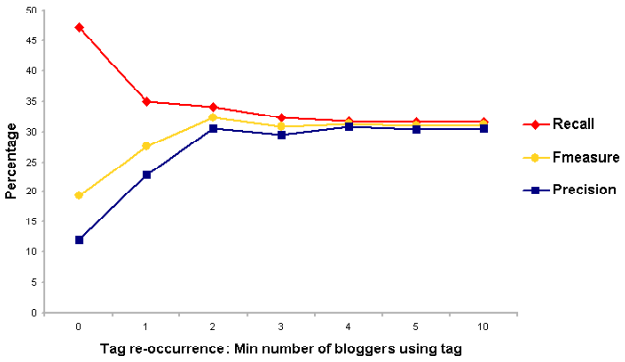

Figure 4: Effect of Reoccur_count on generated tags

candidate set increases and the probability that the actual collaborative tags are also included in the set increases. This results in an increased recall. However as the number of blogs increases, the noise also increases since many of the blogs would not be exactly similar to the blog in question. Hence the number of irrelevant tags in the candidate set also increases causing a decrease in the precision value. We observed the best balance of precision and recall values was obtained when the number of similar bloggers was set to 4.

$ReOccur\_Cutoff$ indicates the cut-off value for the tag re-occurrence property of a tag. A candidate tag is included in the final set only if its re-occurrence value is more than or equal to $ReOccur\_Cutoff$. So every tag in the final set will be used in atleast $ReOccur\_Cutoff$ number of blogs. Figure 4 shows the effect of this parameter on the precision, recall and f-measure values. Predictably as $ReOccur\_Cutoff$ increases the precision also increases but recall decreases. However if it is increased beyond 10, the precision and recall values converge. The best f-measure value was obtained when $ReOccur\_Cutoff$ was set to 2. Finally $CoOccur\_Cutoff$ which indicates the minimum co-occurrence value for a new tag with a given tag for the same blog, was set to 1. When the value of this parameter was set to 0 there was no validation of the tags resulting in a significant amount of noise. Any value greater than 1 saw no significant increase in precision but resulted in a reduction of the recall value. So the value for $CoOccur\_Cutoff$ was set to 1.

8. CONCLUSIONS AND FUTURE WORK

The popularity of the blogs on the blogosphere follows a power law distribution with a large number of blogs being present in the long tail. These blogs are not discovered by other users and hence are not available on collaborative tagging systems. The resulting unavailability of social annotations for these blogs in the long tail is a major hindrance for the use of annotations in web-scale information retrieval systems in improving blogging and searching experience. Tags given by individual authors would not be of much help because of the drawbacks like the vocabulary problem. In this work, we present the SocialTagger system toward solving this problem by automatically generating collaborative tags for blogs. In addition, we propose to use collaborative tags obtained from a web-scale social bookmarking system to evaluate tags generated by individual authors, a baseline method, and the SocialTagger system. With the assistance from systems like SocialTagger, existing collaborative tagging systems can tap on a large number of hidden blogs in the long tail, which would pave the way for use of annotations in web-scale systems.

For our future work, we are considering the following extensions to the SocialTagger system. First, we aim to extend the collaborative tag suggestion system to include additional types of contents. While extending it to suggest tags for webpages with text would be fairly straightforward, other types of content like images would present new challenges. Second, for finding similarity between blogs we are consid-

ering only the content of the blogs in terms of blog text and tags. However, other sources of data like comments, in-links and out-links can be helpful by providing additional information to find similar blogs. We plan to evaluate and include these sources in the future. Third, we aim to investigate the scalability issue in blog similarity ranking. When a large number of blogs are available, one can intuitively take advantage of the structural information that hierarchically organize blogs. In other words, blog clustering can be exploited. By first clustering blogs based on categories [1] and then extracting the similar bloggers only from the same cluster it would be possible for the SocialTagger algorithm to handle a very large number of blogs. We believe this is a promising direction for further study. As social media become more accessible and informative to allow wider participation, even more business and research opportunities will present to IT industry and researchers.

9. REFERENCES

[1] N. Agarwal, M. Galan, H. Liu, and S. Subramanya. Clustering blogs with collective wisdom. *8th International Conference on Web Engineering (ICWE08)*, 2008.

[2] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.

[3] Christopher H. Brooks and Nancy Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM.

[4] Hsinchun Chen. Collaborative systems: solving the vocabulary problem. *Computer*, 27:58–66, 1994.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

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

[7] Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[8] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2007. ACM.

[9] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[10] 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, New York, NY, USA, 2008. ACM.

[11] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(312-377), 1936.

[12] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM.

[13] G. Macgregor and E. McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55:291–300, 2006.

[14] Gilad Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 953–954, New York, NY, USA, 2006. ACM.

[15] C. Shirky. Power laws, weblogs, and inequality. 2003.

[16] Rashmi Sinha. A social analysis of tagging. *http://www.rashmisinha.com/2006/01/a-social-analysis-of-tagging*, 2006.

[17] S. Sood, K. Hammond, S. Owsley, and L. Birnbaum. Tagassist: Automatic tag suggestion for blog posts. *International Conference on Weblogs and Social Media*, 2007.

[18] J. Udell. Collaborative knowledge gardening. *InfoWorld. August*, 20, 2004.

[19] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Inferring a semantic representation of text via cross-language correlation analysis, 2002.

[20] Duncan J. Watts. Is justin timberlake a product of cumulative advantage? *The New York Times*, April 2007.

[21] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May*, 2006.