User-induced Links in Collaborative Tagging Systems

Ching-man Au Yeung
University of Southampton
Southampton, SO17 1BJ,
United Kingdom
cmay06r@ecs.soton.ac.uk

Nicholas Gibbins
University of Southampton
Southampton, SO17 1BJ,
United Kingdom
nmg@ecs.soton.ac.uk

Nigel Shadbolt
University of Southampton
Southampton, SO17 1BJ,
United Kingdom
nrs@ecs.soton.ac.uk

ABSTRACT

Collaborative tagging systems allow users to use tags to describe their favourite online documents. Two documents that are maintained in the collection of the same user and/or assigned similar sets of tags can be considered as related from the perspective of the user, even though they may not be connected by hyperlinks. We call this kind of implicit relations user-induced links between documents. We consider two methods of identifying user-induced links in collaborative tagging, and compare these links with existing hyperlinks on the Web. Our analyses show that user-induced links have great potentials to enrich the existing link structure of the Web. We also propose to use these links as a basis for predicting how documents would be tagged. Our experiments show that they achieve much higher accuracy than existing hyperlinks. This study suggests that by studying the collective behaviour of users we are able to enhance navigation and organisation of Web documents.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Web-based interaction; H.5.4 [**Hypertext/Hypermedia**]: User issues

General Terms

Experimentation, Human Factors

Keywords

Collaborative tagging, Folksonomy, Hyperlink

1. INTRODUCTION

Hyperlinks, or simply links, are probably the most important elements on the Web, as their existence is the reason why the Web is a web: they allow Web users to jump from one hypertext document to another, making navigation through the Web possible. Very often only the author of a hypertext document can decide on which other documents this one can link to. While there are personalised portal sites that general dynamic content based on, for example, the preferences or habits of the users, the majority of hyperlinks are created from the perspective of the authors. While such perspective may be necessary when hyperlinks are created for navigation within a Web site, such author-created hyperlinks can be limited when they come to direct Web users to relevant or potentially interesting documents.

In recent years, there has been a surge in the popularity of user-contributed content on the Web. In particular, tags are extensively used on many Web sites by users to organise and share online resources. In some social bookmarking sites such as Delicious, Web users maintain a collection of documents (Web pages identified by their URLs) that are categorised by their chosen tags. In general, two documents that are in the collection of the same group of users or that are assigned similar sets of tags can be considered as related to each other. From the perspective of hyperlinking, two such documents should be linked to each other such that Web users accessing the first one can be recommended the second one, and vice versa. This leads to an interesting question regarding collaborative tagging: when two documents are tagged by the same group of users or are assigned similar tags, are they linked to each other? If not, how different are these two types of links, namely existing hyperlinks and user/tag similarity, between the documents?

In this paper, we describe two different methods to discover this kind of implicit relations–what we call user-induced links–between documents from a folksonomy. We investigate how these user-induced links between the documents are different from existing hyperlinks on the Web. We show that user-induced links are more likely to link documents that are from different Web sites and are highly related to each other in terms of their content. We also propose to use user-induced links in predicting the tags of a document. Experiments show that user-induced links provide more accurate prediction than existing hyperlinks. Our study suggests that there are differences between the perspectives of authors and readers, and that the link structure on the Web can be greatly enhanced by taking the collective user behaviour in social Web sites into consideration.

We first give a brief description of collaborative tagging in the next section, and describe in detail the notion of user-induced links in Section 3. We then present our analysis of these links in Section 4, and describe our method of predicting tags of documents in Section 5. We discuss the implications of our findings in Section 6, and mention some related studies in Section 7. Finally Section 8 concludes the paper.

CIKM'09, November 2–6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-512-3/09/11 ...$5.00.

2. COLLABORATIVE TAGGING

Collaborative tagging systems such as Delicious[1] and LibraryThing[2] have become very popular among Web users in recent years. They allow users to use freely-chosen keywords–commonly known as tags–to describe and categorise their favourite online resources. For example, a user can post a bookmark of the homepage of BBC (http://www.bbc.co.uk/) to Delicious, and assign to it tags such as `tv`, `media` and `sports`. As tags from different users are aggregated, the tags become an overall description of the document and can be used to facilitate future retrieval.

Collaborative tagging systems are considered to have a number of advantages over traditional methods of organising information [12, 16], as evidently shown by their popularity among general Web users and its application on a wide range of Web resources. In particular, their distinguishable features include the flexibility and freedom offered by these systems to Web users in comparison to traditional systems that usually involve predefined taxonomies or categories. In addition, these systems are quick to adapt to changes in the vocabulary among users, and are therefore particularly favoured by users in the technological domain.

The collaborative tagging activities of participating users result in a user-generated categorisation scheme commonly known as a folksonomy. In general a folksonomy consists of at least three types of elements [13, 21], namely users, tags and Web documents. ***Users*** actively assign tags to Web documents in collaborative tagging systems. ***Tags*** are keywords chosen by users to describe and categorise Web documents. Depending on the design of the systems, tags can be a single word, a phrase or a combination of symbols and alphabets. Finally, ***documents*** refer to the objects tagged by the users in collaborative tagging systems, which can be Web pages, images, videos or even physical objects such as books.

As we are mainly interested in the interrelations between the three types of elements in a folksonomy, additional information such as the time at which a tag is assigned is less relevant. In addition, our primary source of data, Delicious, does not allow any subsumption relations between tags to be defined. Hence, we adopt a basic model of folksonomy which involves only the three basic elements.

DEFINITION 1. *A folksonomy $\mathcal{F}$ is a tuple $\mathcal{F} = (U, T, D, A)$, where U is a set of users, T is a set of tags, D is a set of Web documents, and $A \subseteq U \times T \times D$ is a set of annotations.*

3. USER-INDUCED HYPERLINKS

Hyperlinks in a Web document are generally created by its author. It is conceivable that these hyperlinks may not be adequate from the perspective of the readers of the document. Henzinger [5] mentions two types of hyperlinks, those for navigation and those for recommendation. Recommendation links point users to other documents that contain information related or complementary to the current document. It is possible that an author cannot always ensure that his document has hyperlinks pointing to all of the other relevant documents (or even the most important ones). Useful recommendation links may also be absent because some highly relevant documents are created by rivals of the author and they may be competing to attract more readers. However, from the perspective of the readers, such links can be very valuable.

We argue that collaborative tagging systems such as Delicious offer new opportunities to study how similar or related Web documents are grouped together from the perspective of the users. There are actually two different approaches to discover implicit relations between documents in a folksonomy (as opposed to the explicit hyperlinks between documents). Firstly, given the large number of tags that have been assigned to the documents, implicit links can be found by calculating the similarity between the sets of tags assigned to the documents. For example, [11] describe a system called GiveALink, which involves a global semantic similarity network to capture relationships among resources, and suggest that semantic similarity can be treated as an alternative way of navigating the Web by suggesting users to visit a page similar to the one being visited.

Secondly, a folksonomy is actually quite similar to Web logs and search engine query logs in the sense that it also contains information about the preferences of users under different topics, which are represented by the tags contributed by users. In a Web log or a query log, two documents can be considered as related when users visit both of them in the same context. Similarly, two documents that have both been assigned a particular tag by a large number of users in a folksonomy can be considered as related to each other with respect to the topic represented by the tag. An implicit link can therefore be established between them.

In other words, implicit links between documents in a folksonomy can be discovered by mainly two different approaches: (1) examining the tags that have been assigned to the documents, or (2) analysing the collective behaviour of the users who have tagged the documents. As implicit links in a folksonomy are resulted from the collaborative tagging activities of participating users, we call them ***user-induced links***. In the following sections, we will discuss in detail these two approaches of discovering user-induced links.

3.1 Similarity of Assigned Tags

The first approach of discovering user-induced links in a folksonomy is to calculate the pair-wise similarity between documents based on their tags, and single out those that achieve a certain level of similarity. The similarity between two documents can be measured using many different approaches. Given that documents are characterised by words, similarity is most naturally determined by comparing the set of keywords that are deemed representative of the content of the documents. Such a set of keywords can be extracted by stop-words filtering and weighting schemes such as TF-IDF [17]. A straightforward method of measuring similarity is to use the Jaccard coefficient:

$$Sim(T_a, T_b) = \frac{|T_a \cap T_b|}{|T_a \cup T_b|} \tag{1}$$

where T_a and T_b are the sets of keywords of documents a and b respectively.

However, such simple measure does not take into account the importance of different keywords. It is natural that certain keywords are more central to the content of a document such that they should be given more considerations. In information retrieval, documents are usually characterised by term vectors [10] in a vector space. A term vector is a vector

[1] Delicious: http://delicious.com/

[2] LibraryThing: http://www.librarything.com/"

whose elements indicate the importance of the chosen keywords to the document. Similarity between two documents can be measured by using he cosine similarity calculated on the two respective term vectors:

$$csim(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1|| \times ||\boldsymbol{v}_2||}$$

where $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbf{R}^n$.

Alternatively, one can also consider a document d as characterised by a tuple W_d, which involves a set T_d of tags and a weighting function $w_d(t)$ that maps a tag t to its normalised weight representing its importance to the document:

$$W_d = (T_d, w_d) \tag{2}$$

where

$$T_d = \{t | \exists u, (u, d, t) \in A\} \tag{3}$$

$$w_d(t) = \frac{|\{(u, d, t) | \exists u \text{ s.t. } (u, d, t) \in A\}|}{|\{(u, d, t') | \exists u \text{ s.t. } (u, d, t') \in A\}|} \tag{4}$$

By using this representation of a document, we introduce two different measures for assessing the similarity of two documents. The first similarity measure is a weighted version of the Dice coefficient [9] that is widely used in set comparison:

$$Sim_w(T_a, T_b) = \frac{\sum_{t \in T_a \cap T_b} w_a(t) + w_b(t)}{\sum_{t \in T_a} w_a(t) + \sum_{t \in T_b} w_b(t)} \tag{5}$$

which can be simplified to

$$Sim_w(T_a, T_b) = \frac{\sum_{t \in T_a \cap T_b} w_a(t) + w_b(t)}{2} \tag{6}$$

because $\sum_{t \in T_d} w_d(t) = 1$ as weights of the tags are normalised. This weighted Dice coefficient returns a higher similarity value if the two documents share keywords of higher importance (larger weights).

The second similarity function we introduce here is based on the normalised discounted cumulative gain (NDCG) [7]. NDCG is a performance measure mainly used in information retrieval research to evaluate rankings of documents according to their relevance. It measures how good a ranking algorithm is in assigning suitable ranking to relevant documents. For example, if we have three documents $\{d_1, d_2, d_3\}$ whose relevance scores are $(3, 2, 1)$ respectively (higher score means more relevant), then a ranking (d_1, d_2, d_3) will attain a higher NDCG than another ranking (d_3, d_1, d_2), because the first one assigns higher ranks to documents that are more relevant.

Here, we borrow the idea of NDCG to measure the similarity of two documents based on their tags and the associated weights. Assume that we have two documents d_1 and d_2, and we now want to assess how similar d_2 is to d_1. For d_1, we have a list of tags organised in descending order of their weights, $(t_1, t_2, ..., t_n)$, whose weights are $(w_{d_1}(t_1), w_{d_1}(t_2), ..., w_{d_1}(t_n))$. We treat the tags of d_1 as items to be retrieved and ranked, and treat their weights as their relevance scores. As a result, the list of tags of d_2 can be considered as a ranking result produced by some ranking algorithm to reproduce the list of tags of d_1 as accurately as possible. In this way, two documents with the same set of tags and same ordering according to their weights will achieve an NDCG of 1, two documents that share no tags at all will result in an NDCG of 0. It should be noted that unlike the weighted tag similarity the NDCG similarity measure is asymmetric.

Formally, calculating the NDCG similarity of d_2 to d_1 requires several steps. Firstly, lists of tags of the two documents are prepared, with the tags ordered in descending order of their weights:

$$l_{d_1} = (t_{d_1,1}, t_{d_1,2}, ..., t_{d_1,n}) \tag{7}$$

$$l_{d_2} = (t_{d_2,1}, t_{d_2,2}, ..., t_{d_2,n}) \tag{8}$$

Secondly, the discounted cumulative gain (DCG) at position p is calculated by:

$$DCG_p = w_{d_1}(t_{d_2,1}) + \sum_{i=2}^{p} \frac{w_{d_1}(t_{d_2,i})}{\log_2 i} \tag{9}$$

Thirdly, we need the ideal discounted cumulative gain (iDCG) at position p, which is the DCG at position p when tags are ranked exactly according to their weights (the ideal case). It is used to normalise the DCG obtained using the above equation such that the final NDCG value varies in the range of 0 to 1.

$$iDCG_p = w_{d_1}(t_{d_1,1}) + \sum_{i=2}^{p} \frac{w_{d_1}(t_{d_1,i})}{\log_2 i} \tag{10}$$

Finally, the NDCG value is calculated by simply obtaining the ratio between DCG and iDCG:

$$NDCG_p = \frac{DCG_p}{iDCG_p} \tag{11}$$

Given these similarity measures, it becomes possible to discover user-induced links between pairs of documents that are similar to each other. One important issue in using similarity to discover implicit links is that we are likely to discover a huge number of implicit links. This is because it is very likely for the documents to share one or two very general tags that are, however, not particularly related to each other in terms of their content, and would nevertheless achieve none zero similarity. Hence, a threshold value of similarity should be specified in other to narrow down the results to a reasonable and useful set of implicit links. In summary, the process of discovering user-induced links using one of the similarity measures can be represented by function that takes the set of documents, the chosen similarity measure and the similarity threshold as parameters:

$$G_s(D, Sim, threshold) = \{(d_i, d_j)\} \tag{12}$$

3.2 Association Rule Mining

The second approach of discovering implicit links involves finding out pairs of Web documents that have both been tagged by the same group of users, probably with the restriction on the same tag or same set of tags. In other words, we try to identify these links by studying user preferences. The method for identifying such pairs of Web documents can in fact be readily borrowed from the data mining research area. The task of mining association rules from large databases [1] aims at identifying implicit patterns within a large database of transactions. In traditional association rule mining, a classic example would be that people who buy bread and butter in the supermarket are very likely to buy milk as well. Borrowing such idea to the context of collaborative tagging, the problem becomes one of identifying pairs of Web documents such that when users have tagged one of them they are very likely to tag the other one as well. In other words, we can use the technique of association rule mining to discover these user-induced links.

Formally, let $D = \{d_1, d_2, ..., d_n\}$ be a set of Web documents, and C be a database of document collections. Each $c_u \in C$ represents the set of documents that have been tagged by the user u. In traditional association rule mining, let X and I_k denote sets of items, rules can assume the form of $X \Longrightarrow I_k$, meaning that the presence of X in a certain transaction implies a high probability of the presence of I_k in the transaction. However, in the case of identifying user-induced hyperlinks, it is not very helpful to discover something like 'd_1, d_2 and d_3 should altogether have a link to d_4', as links should be originated from a single document to another single document. Hence, we will focus on discovering association rules in the form of $d_i \Longrightarrow d_j$.

Two major concepts in association rule mining are *support* and *confidence*. In our context, support of a set of documents is defined as the proportion of collections in the database that contain the set of documents:

$$supp(X) = \frac{|\{c_u | X \subseteq c_u, c_u \in C\}|}{|C|} \tag{13}$$

In general, we aim at discovering rules that have large supports. This is because a larger support implies that the rule involves documents that are more popular among the users. Therefore rules of larger supports will find themselves more useful in the future.

Confidence of a rule $d_i \Longrightarrow d_j$, on the other hand, is defined as the proportion of collections in the database in which the rule is correct:

$$conf(d_i \Longrightarrow d_j) = \frac{supp(\{d_i, d_j\})}{supp(\{d_i\})} \tag{14}$$

In general, we also want the confidence of a rule to be as high as possible. The confidence of a rule actually corresponds to the extent to which the rule is a valid one. A rule that has a higher confidence would mean that it would be more likely to obtain a correct result when the rule is applied. In the context of discovering user-induced links in folksonomies, a higher confidence means that the user-induced link is deemed appropriate by more users and therefore it is more likely that such a link would benefit other users as well.

Similar to the case of NDCG similarity, user-induced links discovered by using association rule mining are not symmetric. The existence of the rule $d_i \Longrightarrow d_j$ does not imply the existence of the rule $d_j \Longrightarrow d_i$, because the two rules would have different levels of support and confidence. In summary, the process of discovering user-induced links in a folksonomy using association rule mining can be represented by the following function:

$$G_u(D, C, min\_supp, min\_conf) = \{(d_i, d_j)\} \tag{15}$$

4. ANALYSIS OF USER-INDUCED LINKS

By using the two methods described above, we identify user-induced links in data collected from Delicious and compare them with existing hyperlinks in terms of several different aspects. In performing the analysis and comparison, we focus on whether the links (including existing hyperlinks and user-induced links) can be considered as good recommendation links. While it can be a subjective judgement of whether a link makes good recommendation to a user, we believe there are several aspects of a link that we can study to answer the question. These aspects include whether a link connects two documents from the same domain/Website, the similarity between documents on the two ends of a link, and whether users are equally interested in the linked documents. We will perform our analysis along these dimensions.

4.1 Data Collection

To conduct the experiments, we collect data from Delicious, which is one of the most popular collaborative tagging systems. The documents submitted by the users covers a wide range of topics. Since Delicious contains a huge amount of data, and one can usually only obtain a relatively small subset of it, the collected data will be very sparse if we collect in a random manner. Hence, we collect data on a per-tag basis. We first collect at random 130 tags from Delicious by looking up the popular tags. Then for each of these 'seed tags' we go on to crawl Delicious to obtain a set of documents that have been assigned the tag, along with all the users who have tagged the documents. Altogether we have about 130 thousand unique documents, 1.2 million unique users and 0.8 million unique tags. On average we have about 1,200 documents for each seed tag. This data set is an expanded version of the one used in another study of ours that concerns expertise in folksonomies [15].

To obtain the existing link structure among the collected documents, we download each of them and parse the HTML source code to identify their outgoing links. Since our experiments focus on the characteristics of the documents on the two ends of a link, we do not consider links that point to documents not in our data. At the end of this process we have 56,900 links. The maximum number of outgoing links for a document is 58, and the maximum number of incoming links is 240.

4.2 Results

We identify user-induced links between the documents in each of the 130 data sets (corresponding to the 130 seed tags) by the two proposed methods using different parameters. For the similarity approach, we vary the similarity threshold. As this approach tends to return a lot of user-induced links, we only focus on links between documents with similarity of at least 0.5. For the association rule mining approach, we set the minimum support at 100 and vary the minimum confidence level. We find that very few user-induced links achieve a confidence level of 0.5 or above. While we can also vary the minimum support in our experiments, we are more interested in user-induced links that are supported by a relatively larger number of users, thus reflecting the preferences of a large community, hence we fix the minimum support in our experiments.

Table 1 shows the number of user-induced links generated by using different methods and parameters. An obvious difference between the different methods is that the use of tag similarity generates far more user-induced links than the use of association rule mining. This actually reflects a major difference between the two methods. In using similarity, we compare the tags of different documents, since we focus on a group of documents with a particular tag at a time, the documents are already confined to a single (though very general) topic. As a result, the diversity of tags found in this group of documents is far smaller than the diversity of users who are interested in these documents, thus resulting in a much higher 'tag similarity' than 'user similarity' among the documents.

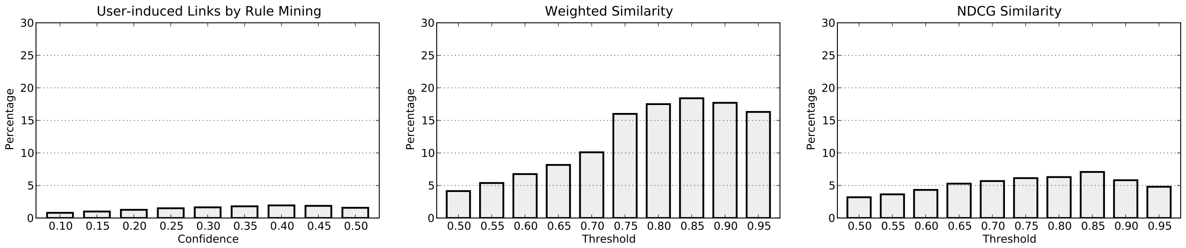


Figure 1: Percentage of user-induced links connecting documents from the same domain.

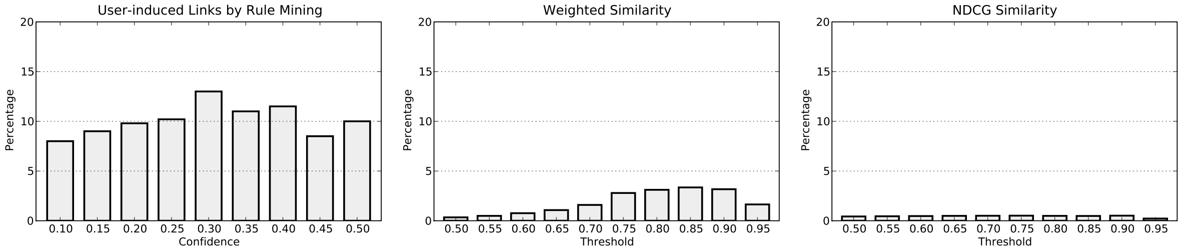


Figure 2: Percentage of user-induced links that are existing hyperlinks.

Association Rule		Weighted Similarity		NDCG Similarity	
Conf.	Links	Thres.	Links	Thres.	Links
> 0.10	75	> 0.50	11,724	> 0.50	15,294
> 0.15	39	> 0.55	7,371	> 0.55	10,897
> 0.20	20	> 0.60	4,279	> 0.60	9,969
> 0.25	13	> 0.65	2,696	> 0.65	8,981
> 0.30	10	> 0.70	1,545	> 0.70	8,050
> 0.35	10	> 0.75	667	> 0.75	7,356
> 0.40	9	> 0.80	516	> 0.80	6,806
> 0.45	7	> 0.85	408	> 0.85	6,466
> 0.50	6	> 0.90	366	> 0.90	6,240
		> 0.95	355	> 0.95	6,125

Table 1: Average number of induced links generated for each tag data set by different methods using different parameters. The figures are averaged over the 130 data sets collected from Delicious.

While we note that weighted similarity and NDCG similarity have a high correlation ($r \approx 0.85$), there are actually some differences between the two. Although both similarity measures consider the weights of the tags, NDCG puts much more emphasis on matching tags that are most important. Consequently, if two documents have their a few most popular tags in the same order, they are very likely to attain a higher value in NDCG similarity than in weighted similarity. As a result, we see that NDCG similarity gives us a lot more induced links than weighted similarity.

4.2.1 *Number of Same-Domain Links*

One important function of hyperlinks is to allow users to navigate from one hypertext document to another, especially those within the same Web site. Arguably, it would be more beneficial to a user if links point to some documents external to the current Web site, which should provide relevant information different from that available in the current one. For example, links from a blog post in one blog to blog posts in another blog would be more informative in general than links to blog posts within the same blog. Hence, it would be interesting to compare this aspect in existing hyperlinks between the documents and links induced from the tagging behaviour of Web users.

For each of the existing hyperlinks and the induced links, we check whether the documents at the two ends of the link are from the same domain. We do this by comparing their URLs and see if they have the same domain name. For example, a test on `http://developer.apple.com/` and `http://support.apple.com/` will be positive as they are both under the domain name of `apple.com`. We note that, however, this may overestimate the number of links connecting documents from the same Web site. This is because two URLs having the same domain name but different sub-domain names may as well be referring to two different Web sites. For example, we may want to consider a blog at `http://userA.blogspot.com/` and another blog at `http://userB.blogspot.com/` as two different Web sites, although they are both under the same domain. In practice, these subtle differences may be difficult to distinguish from one another when automatic processing of the URLs is involved. Nevertheless, since we compare the different types of links on the same basis, this should not be considered as a problem.

Figure 1 shows the percentage of links that connect documents from the same domain for user-induced links generated by using the three different methods. We note that for existing hyperlinks about 33% of them are between documents from the same domain, and the probability of having such a same-domain link in our data sets is about 15%. Firstly, we see that only about 1-4% of user-induced links generated by association rule mining are connecting documents from the same domain. This is much lower than that of existing hyperlinks and by chance, suggesting that users are very unlikely to be interested in multiple documents from the same domain.

The graphs of the links generated by similarity of assigned tags seem to suggest that there is a difference between weighted similarity and NDCG similarity. However, taking the different number of links generated in the two cases into consideration, this difference is only due to the different distribution of links among the similarity level. The number of links generated by NDCG similarity that attain a similarity level of 0.95 is greater than that generated by weighted similarity that attain a similarity level of 0.60. This shows that NDCG is less fine-grained than weighted similarity, and it is relatively easier to achieve high similarity in NDCG. The graph for weighted similarity suggests that links in which documents are more similar are more likely to be from the same domain. When we pick some of these links for further investigation, we see that many of these links are between a series of documents addressing the same topic in a blog, or tutorials of highly related applications. Nevertheless, compared to existing hyperlinks, there are much fewer user-induced links that connect documents from the same domain.

4.2.2 *Coincidence between Existing Hyperlinks and User-induced Links*

In addition to examining the domains of linked documents, another way to study the usefulness of user-induced links is to see whether such links already exist between the documents. If user-induced links coincide with existing hyperlinks, it suggests that users are satisfied with the existing hyperlinks and do not pay much attention to other documents linked. On the other hand, if user-induced links are mostly new, it means that there are user interests and perspectives that existing hyperlinks have not captured.

Figure 2 shows the percentage of user-induced links that coincide with existing hyperlinks. The graphs seem to show that user-induced hyperlinks generated by association rule mining are more likely to be existing hyperlinks, and that those generated by NDCG similarity are less likely to be so. However, we again have to take into account the different numbers of links generated in different cases. Since there are a lot more user-induced links based on similarity than those based on association rule mining, it is understandable that the formers coincide much fewer existing hyperlinks.

The result for induced links generated by association rule mining is particularly interesting. This is because given the relatively few user-induced links in this case, the overlap between these and existing hyperlinks is at most about 13%. This shows that a hyperlink does not necessarily connect documents both of which users find interesting or useful. In other words, users tend to find out related documents by other means because there are no hyperlinks between them. It is possible that two documents are not directly linked but can be reached by two or more hops on the Web graph. However, as shown in Table 2, only a very tiny portion of documents that are not directly linked can be reached by more hops.[3] In addition, one may suggest that users do not tag both documents connected by a link simply because of the existence of the link: it is sufficient to save one of them which will lead the user to the other. However, given that all these documents have been tagged by some users, it suggests that all these documents deserve to be bookmarked for future retrieval.

Path Length	Frequency	Percentage
∞	6,442	89.26%
1	439	6.08%
2	132	1.83%
3	57	0.79%
4	42	0.58%
5	29	0.40%
> 5	76	1.05%

Table 2: User-induced links and the lengths of the shortest paths between the documents concerned.

To get a better understanding of the user-induced links, we look into documents that are connected by these links but not by existing hyperlinks. We find that many of the user-induced links are (1) between blog posts of highly related topics, (2) news articles on the same topics, (3) Websites offering applications of similar functionalities, and (4) Q&A pages of some portal sites. In all these cases, there are some reasons that hyperlinks do not exist. For example, the author of a document may not be aware of other related documents (as in 1 and 2), or two Websites are competing for readership because they offer similar content (as in 3), or the system is not designed to be aware of the similarity of its content (as in 4).[4]

The results of similarity-based user-induced links are less surprising given the very large number of links generated. However, they do show that existing hyperlinks are very inadequate when they come to recommend related documents to the users. There are just much more related documents out there than those to which hyperlinks within a document point to. Of course, it would not be practical for a document to be linked to all of, for example, the 10,000 documents that contain related materials. Nevertheless, the results suggest that there are clearly room for improvement for existing hyperlinks.

4.2.3 *Similarity and User Preferences*

The two approaches for generated user-induced links are completely different from each other. Association rule mining concerns with the preferences of the users, and a link is generated if enough users are interested in two documents, regardless of the similarity between them. On the other hand, the similarity-based approach generates links based on the tags assigned to the documents, regardless of whether there are many users interested in the documents on the two ends of the links. In this section, we investigate whether the links generated by one method satisfy the requirement of the other method.

Figure 3 graphs the similarity between documents connected by user-induced links generated by association rule mining. We can see that all the pairs of documents attain high similarity in the two similarity measures. It shows that pairs of documents that are interested by many users are actually very similar to each other with respect to the tags assigned to them, which are indicative of their topics. We also calculate the similarity between documents connected by existing hyperlinks for reference. As shown in Figure 3, existing hyperlinks achieve about 0.62 in both similarity measures, which is much lower than those achieved

[3] It is possible that a path exists by traversing documents that are not in our data sets. However this is beyond our scope as that requires the knowledge of the global Web graph.

[4] Case 4 is likely to be found on FAQ documents provided by authors. In many user-contributed Q&A sites like Yahoo! Answers, similar questions and answers are usually recommended to the users by the systems.

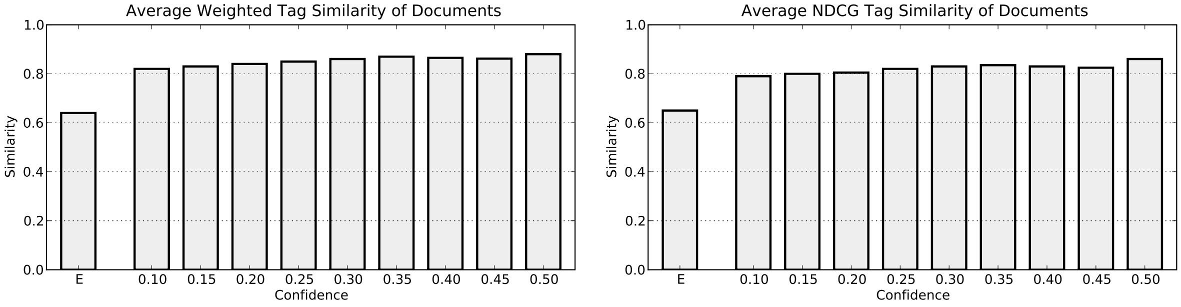


Figure 3: Average similarity of pairs of documents on the two ends of user-induced links generated by association rule mining. 'E' refers to the results obtained for existing hyperlinks.

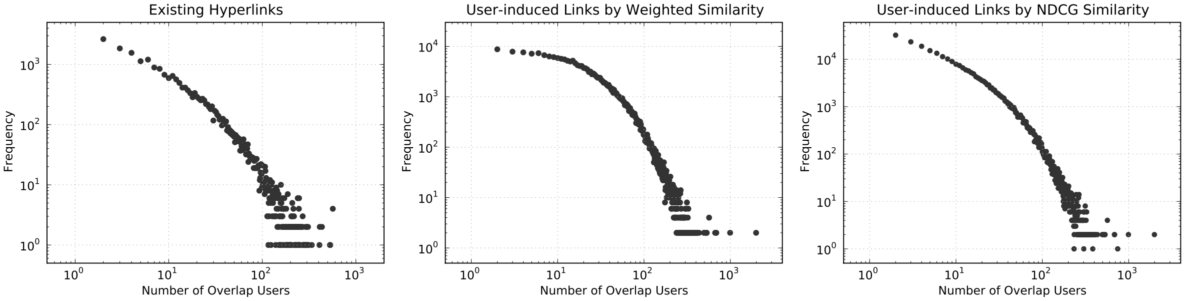


Figure 4: Number of users that have tagged both documents on the two ends of a link. It can be seen that relatively very few users express explicit interests in documents linked by existing hyperlinks or user-induced links generated by tag similarity.

by user-induced links. This result for existing hyperlinks is expected because many of them serve navigational purposes and therefore it is not uncommon for their sources and destinations to involve content of different topics (e.g. a link back to the front page of a Web site or a link from a blog post to the profile of the blogger).

Next, we investigate whether similarity-based user-induced links connect pairs of documents that are interested by many users. Figure 4 graphs the number of users that have tagged both documents on the two ends of a link. For both existing hyperlinks and similarity-generated user-induced links, we see a power law-like distribution of the number of overlap users. In other words, there are only a very small number of links that connect documents both of which are interested by a large number of users, and there are a large number of links for the opposite case. Hence, contrary to the findings in user-induced links generated by association rule mining, high similarity does not guarantee high user preferences. It may suggest that users are much more selective and do not only consider the similarity between two documents. It may also suggest that given the large amount of information users choose to focus on a small number of documents that are related and are useful from their own perspectives.

Putting the findings described in this section together, we can see that explicit user preferences (association rule mining of user collections) represent better filters of useful relations between documents that similarity measures. The former satisfies both user preferences as well as topical similarity between the documents, whereas the latter does not necessarily produce links that are interested by many users. Nevertheless, both approaches can be considered as useful means for identifying implicit relations between documents that are not captured by existing hyperlinks, as our experiments show that user-induced links offer much new information that cannot be found in existing hyperlinks.

5. TAG PREDICTION

The analysis of user-induced links shows that links generated by association rule mining of user collections usually connect documents that are highly related to each other, as judged by the similarity between their tags. This result inspires us to use this particular kind of user-induced links to predict tags of a document. Given that documents connected by user-induced links have highly similar sets of tags, aggregating the tags of documents linking to a chosen document is probably a good way of predicting the tags of this document. It is suggested that tag prediction has several useful applications in collaborative tagging, such as enhancing annotation and retrieval of resources [6].

5.1 Proposed Method

To predict the tags of a certain document, we first need to identify the other documents that have a link to this document. Let $G = (D_G, L_G)$ be a graph with a set D of vertices representing documents and a set L of arcs representing links between the documents. We consider both a graph G_w of existing hyperlinks and a graph G_u of user-induced links generated by using association rule mining of user collections. For a document d_x in the graph, the set of documents that have a link to d_x is given by:

$$P_G(d_x) = \{d | (d, d_x) \in L_G\} \tag{16}$$

Our hypothesis is that documents in $P_G(d_x)$ contain information related to the content of d_x, and therefore the tags of the documents in $P_G(d_x)$ should also be applicable to d_x. We can aggregate the tags of these documents and use them to predict the tags of d_x. We consider two different methods of aggregating the tags of documents in $P_G(d_x)$. Firstly, we consider a simple averaging method: we come up with a set of tags with their weights equal to the average of their weights in documents in $P_G(d_x)$. Let $W^a_{d_x} = (T^a_{d_x}, w^a_{d_x})$ represents the prediction (superscript a means average aggregation), where $T^s_{d_x}$ is the set of tags and $w^s_{d_x}$ is a function that returns the weight of the tags. Our first method of aggregation is given by:

$$\begin{aligned} T^a_{d_x} &= \bigcup_{d \in P_G(d_x)} T_d & (17) \\ w^a_{d_x}(t) &= \frac{1}{|P_G(d_x)|} \sum_{d \in P_G(d_x)} w_d(t) & (18) \end{aligned}$$

In addition, by assuming that an induced link of higher confidence will connect a more related document to d_x, we also consider a slightly more sophisticated method of aggregation by taking the confidence of the link into account. Let $conf(d_1 \Longrightarrow d_2)$ be the confidence of the user-induced link from d_1 to d_2. Our second method of aggregation is given by $W^w_{d_x} = (T^w_{d_x}, w^w_{d_x})$ (superscript w means weighted aggregation), where

$$\begin{aligned} T^w_{d_x} &= \bigcup_{d \in P_G(d_x)} T_d & (19) \\ w^w_{d_x}(t) &= \frac{\sum_{d \in P_G(d_x)} w_d(t) \times conf(d \Longrightarrow d_x)}{\sum_{d \in P_G(d_x)} conf(d \Longrightarrow d_x)} & (20) \end{aligned}$$

Note that our proposed method of predicting the tags of a document is similar to the k-nearest-neighbour algorithm, in which the class label of an item is determined by those that are closest to it, except that in our case the number of neighbours of a document is not fixed and depends on the number of user-induced links that have this document as their common target. In other words, tag prediction can also be considered as a classification problem. Performance of user-induced links in tag prediction is therefore indicative of their usefulness in Web document classification.

5.2 Experiments

From our data sets, we select a set of testing documents that have at least 5 incoming user-induced links and at least 5 incoming hyperlinks from other documents to ensure that we have enough data for the prediction process. After the filtering process we obtain a total of 1,241 documents satisfying the above conditions. On average a document in the set has 9 incoming user-induced links and 14 incoming hyperlinks. We use the average aggregation method to generate predictions from hyperlinks (as they do not have any confidence values), and use both average and weighted aggregation method to generate predictions from user-induced hyperlinks.

We measure the performance of the predictions by using NDCG as well as precision at the nth item. Precision at the nth item is calculated by measuring the precision of the first n tags, i.e. the top n tags with largest weights, in the prediction. On the other hand, NDCG as a performance measure works effectively in the same way as described in Section 3.1. We use NDCG mainly to investigate whether the predictions are accurate in terms of the ordering of the tags. In our experiments, we use the tags assigned to the documents by the users in Delicious as the ground truth.

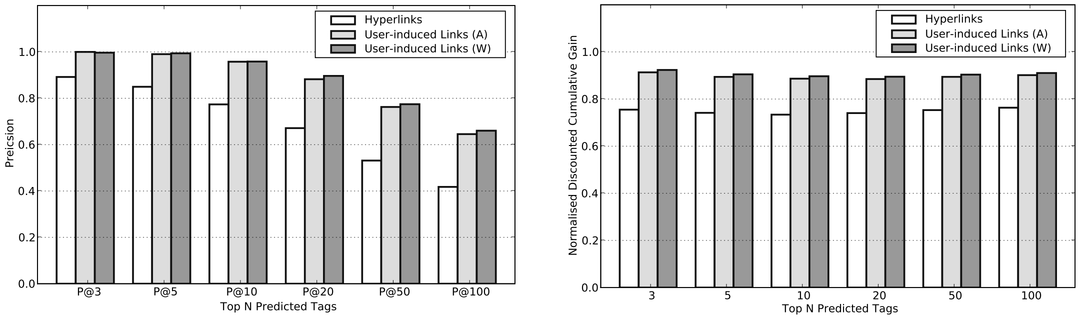
Figure 5(a) shows the precision levels of the predictions for different values of n. We can see that predictions based on user-induced links are significantly more accurate that those based on existing hyperlinks, with precision of 90% or higher for the first 20 tags. The performance of using weighted aggregation gives slightly better results than using average aggregation. Note that the number of predicted tags is always larger than the actual number of unique tags assigned to the document in Delicious since we do not impose any threshold on the weight of the tags. Given the fact that precision decreases as we consider more and more tags in the prediction, it can be concluded that correct tags are usually given higher weights in the prediction than wrong tags. This is confirmed by the results given by the NDCG measure.

Figure 5(b) shows the NDCG values of the predictions when different number of top predicted tags are considered. Again, we see that predictions based on user-induced links attain significantly higher values than those based on existing hyperlinks, and that weighted aggregation gives slightly better results than average aggregation. Judging from the fact that the NDCG values of the predictions are always higher than 0.9, the user-induced links represent a good basis for predicting even the relative importance of different tags to a document. An interesting result is that the values of NDCG do not change much at different positions. They are more or less constant even we consider more tags in the predictions. This is in fact related to the popularity of the tags. We observe that the number of times the tags are used on a document usually follows the power law, with a few tags very popular among the users and a large number of tags that are only favoured by a small number of users. Hence, once the first few tags are correctly predicted a high NDCG value will be obtained, and subsequent correct or incorrect predictions will not change the value significantly.

6. DISCUSSION

Our study reveals that implicit relations between Web documents can be discovered by examining user preferences and document similarity embedded in a folksonomy. We also show that user-induced links are very different from existing hyperlinks in several different aspects, including the proportion of links between documents from the same domain, the number of users interested in the documents and the similarity between the documents.

An important aspect of the Web revealed by this study is that, at least within a collaborative tagging environment, there is a big difference between the perspective of Web authors and that of Web readers [2]. This can also be framed as a difference between the expectation of Web designers and the behaviour of Web surfers, or even a difference between Web 1.0 and Web 2.0. Hyperlinks are supposed to provide users with recommendations of related documents, but it turns out that users find out interesting documents very often without the help of hyperlinks. This suggests that it is very desirable to complement the existing link structure on the Web with information of user preferences.



(a)

(b)

Figure 5: (a) Precision levels of predictions at different number of tags. (b) Normalised discounted cumulative gain (NDCG) of the predictions. (A) means average aggregation of tags, and (W) means weighted aggregation of tags.

However, providing suggestions of hyperlinks to authors by using information of user preferences may only be a limited solution to the problem. In the course of our study, as we have mentioned in Section 4.2.2, we discover many induced links between Web sites that can be considered as rivals or competing for readership. Hence, it is not realistic to expect that authors of these Web sites would create such hyperlinks. In addition, it may as well be the authors' intention to limit the number of hyperlinks due to various reasons.

In the end, it may be worthwhile to consider something like an open hypermedia structure [4] backed by a collaborative tagging system. Links between different Web documents are induced from the collective behaviour of the users, and are maintained externally with respect to the documents involved. These links represent the perspective of the users on how documents on the Web should be linked to each other. There are also possibilities of working towards semantic links since documents have been assigned tags by users. For example, when we generate user-induced links between documents that are all assigned the tag `cooking`, induced links between these URLs can be described by the tag, giving the users an idea of why these URLs are linked. In other words, user-induced links have great potentials to be further studied.

7. RELATED WORK

As long as browsing behaviour is concerned, it is obvious that collaborative tagging systems represent only one of the data sources (albeit a new and popular one from which data can be easily collected) from which relations between Web documents can be induced. Xue et al. [22], while working on improving search within a single Web site, apply association rule mining to Web logs to find out pairs of documents that are frequently visited by users of the Web site. The authors call this kind of implicit relation between documents implicit links. Similar to the results we have reported in Section 4.2.1, they find that the overlap between the set of implicit links and explicit links is small (11%). It is also reported that running the PageRank algorithm on the implicit links gives better performance in retrieval than on the explicit links alone. Along the same line of thought, Kazienko and Pilarczyk [8] propose to use this kind of implicit links to assess the quality of hyperlinks.

Shen et al. [18] present a method to generate implicit links from search engine query logs for the purpose of classifying Web documents. They propose that two documents are linked by an implicit link if they are both chosen (clicked) under the same query submitted by the same user. The authors find that making use of these implicit links improve results of Web document classification, due to the fact that implicit links tend to connect similar documents. However, we note that Web logs and query logs do not necessarily show the positive preferences of the users, as we cannot be sure that a user is interested in a document simply because he has visited it or because he has clicked on it after submitting a query. On the contrary, users tag a document usually because they are interested in it. Hence, we believe folksonomies provide more reliable data for studying implicit links between documents.

The act of identifying similar documents within a folksonomy can be considered as inducing links from the user behaviour. For example, Markines et al. [11] describe a system, GiveALink, which involves a global semantic similarity network to capture relationships among resources. The authors suggest that semantic similarity can be treated as an alternative way of navigating the Web by suggesting users to visit a page similar to the one being visited.

Considering that two documents are similar and related to each other based on the tags assigned to them is a very common idea among studies of folksonomies. However, such idea has only been seen in implementing recommendation systems using collaborative tagging (e.g. [14, 19, 20]). More works can be found on establishing relations between users (community discovery) or relations between tags (ontology or semantic network generation). For example, Mika [13] shows that explicitly considering the behaviour of the users can lead to better networks of tags that represent the relations between the tags from the perspective of the user community.

Tag prediction is also studied in some previous works. Heymann et al. [6] apply association rule mining to discover relations – e.g. documents with the tag `w3c` are likely to be assigned `web` as well – between tags, and use these relations to predict tag assignments. Budura et al. [3] propose to use neighbourhood information to predict whether a tag should be assigned to a document. Their work is similar to our study of tag prediction in that they determine the suitability of a tag by examining its occurrence in the neighbourhood of a document defined by the hyperlink structure. However, these previous studies of tag prediction focus on predicting additional tags to a document, rather than suggesting tags to a document that has no tags as in our case.

In summary, while finding related documents within a folksonomy is very commonly mentioned in the literature, the idea that these relations can be realised as hyperlinks between documents, and the idea of comparing these implicit links with existing hyperlinks have eluded the research community so far. Implicit links described in our work have also not been considered for tag prediction or classification in the literature.

8. CONCLUSION

We study user-induced links, a form of implicit relations, between documents as discovered in collaborative tagging. We show that both user preferences and tag similarity can be used to generate many user-induced links, and approach of using association rule mining generates very high quality user-induced links because they are both highly preferred by the users and connect documents that contain highly related content. We also show that user-induced links can be used to predict tags of documents with a very high accuracy. Our study reveals the difference between the perspectives of authors and that of readers on the Web.

As we have discussed in the previous section, user-induced links have great potentials, and therefore we wish to extend our study in several different directions. In this study we study these links by grouping documents under the same tag. It would be worthwhile to relax this restriction and study whether there exist cross-topic user-induced links. In addition, our current study is mainly quantitative. We hope to study the utilities of user-induced links by conducting user studies in the future, so as to confirm whether these links are useful from the perspective of Web users. Finally, we would also like to extend our work to other collaborative tagging systems to investigate any domain specific characteristics or user behaviour that may result in a different kind of user-induced links.

9. ACKNOWLEDGMENTS

Ching-man Au Yeung is supported by the R C Lee Centenary Scholarship (2006-2009) generously offered by the Drs Richard Charles and Esther Yewpick Lee Charitable Foundation, Hong Kong.

10. REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD 1993*, pages 207–216, New York, NY, USA, 1993. ACM.

[2] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW '07*, pages 501–510, New York, NY, USA, 2007. ACM.

[3] A. Budura, S. Michel, P. Cudre-Mauroux, and K. Aberer. Neighborhood-based tag prediction. In *ESWC 2009*, pageds 608–622, 2009.

[4] A. M. Fountain, W. Hall, I. Heath, and H. C. Davis. MICROCOSM: an open model for hypermedia with dynamic linking. In *Hypertext: concepts, systems and applications*, pages 298–311. Cambridge University Press, New York, NY, USA, 1992.

[5] M. Henzinger. Hyperlink analysis on the world wide web. In *HYPERTEXT '05*, pages 1–3, New York, NY, USA, 2005. ACM.

[6] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *SIGIR '08*, pages 531–538, New York, NY, USA, 2008. ACM.

[7] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[8] P. Kazienko and M. Pilarczyk. Hyperlink assessment based on web usage mining. In *HYPERTEXT '06*, pages 85–88, New York, NY, USA, 2006. ACM.

[9] S. J. Ker and J. S. Chang. A class-based approach to word alignment. *Comput. Linguist.*, 23(2):313–343, 1997.

[10] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[11] B. Markines, H. Roinestad, and F. Menczer. Efficient assembly of social semantic networks. In *HYPERTEXT '08*, pages 149–156, New York, NY, USA, 2008. ACM.

[12] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. December 2004.

[13] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics*, 5(1):5–15, 2007.

[14] S. Niwa, T. Doi, and S. Honiden. Web page recommender system based on folksonomy mining for itng'06 submissions. In *ITNG'06*, pages 388–393, 2006.

[15] M. G. Noll, C. M. Au Yeung, N. Gibbins, C. Meinel, and N. Shadbolt. Telling experts from spammers: Expertise ranking in folksonomies. In *SIGIR '09.* ACM, 2009.

[16] E. Quintarelli. Folksonomies: power to the people. ISKO Italy-UniMIB meeting, 2005.

[17] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, Cornell University, Ithaca, NY, USA, 1987.

[18] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. A comparison of implicit and explicit links for web page classification. In *WWW '06*, pages 643–650, New York, NY, USA, 2006. ACM.

[19] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08*, pages 259–266, New York, NY, USA, 2008. ACM.

[20] K. Shiratsuchi, S. Yoshii, and M. Furukawa. Finding unknown interests utilizing the wisdom of crowds in a social bookmark service. In *WI '06*, pages 421–424, Washington, DC, USA, 2006. IEEE Computer Society.

[21] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06*, pages 417–426, New York, NY, USA, 2006. ACM Press.

[22] G.-R. Xue, H.-J. Zeng, Z. Chen, W.-Y. Ma, H.-J. Zhang, and C.-J. Lu. Implicit link analysis for small web search. In *SIGIR '03*, pages 56–63, New York, NY, USA, 2003. ACM.