A New Vector Space Model Exploiting Semantic Correlations of Social Annotations for Web Page Clustering

Xiwu Gu[1], Xianbing Wang[1], Ruixuan Li[1], Kunmei Wen[1], Yufei Yang[1],
and Weijun Xiao[2,*]

[1] Intelligent and Distributed Computing Lab, College of Computer Science and Technology
Huazhong University of Science and Technology, Wuhan 430074, P.R. China
{guxiwu,rxli,kmwen}@mail.hust.edu.cn,
{wxb1225,yfyang}@smail.hust.edu.cn
[2] Department of Electrical and Computer Engineering, University of Minnesota, twin cities
200 Union Street SE, Minneapolis, MN 55455, US
wxiao@umn.edu

Abstract. Text clustering can effectively improve search results and user experience of information retrieval system. Traditional text clustering approaches are based on vector space model, in which a document is represented as a vector using term frequency based weighting scheme. The main disadvantage of this model is that it cannot fully exploit semantic correlations between social annotations and document contents because term frequency based weighting scheme only captures the number of occurrences of terms in the document. However, social annotation of web pages implicates fundamental and valuable semantic information thus can be fully utilized to improve information retrieval system. In this paper, we investigate and evaluate several extended vector space models which can combine social annotation and web page text. In particular, we propose a novel vector space model by computing the semantic correlations between social annotations and web page words. Comparing with other vector space models, our experiments show that using semantic correlations between social tags and web page words improves the clustering accuracy with RI score increase of 4% ~ 7%.

Keywords: social annotation, clustering, information retrieval.

1 Introduction

Recent years the advance of Web 2.0 technology has influenced the ways that users interact with Internet. Unlike the traditional publishing-browsing style of interaction between users and Internet, Web 2.0 facilitates users the abilities of information sharing, exploration and collaboration on the Internet. This brings on the evolution of new web applications such as social annotations and social networking.

[*] Correspondence author.

H. Wang et al. (Eds.): WAIM 2011, LNCS 6897, pp. 106–117, 2011.

In particular, a social annotation is referred to as an online annotation associated with a Web resource (typically a Web page, Web image and Web video). By means of online Web annotation tools, users can collaboratively add and modify text labels that describe or categorize Web resource without modifying the resource itself. At present, widely used online Web annotation tools include Delicious (annotating Web pages), Flickr (annotating Web images), YouTube (annotating Web videos), etc.

Essentially, social annotations can be viewed as socially classification layer building on the top of existing annotated Web resources. Moreover, social annotation is user-provided metadata implicating fundamental and valuable semantic information of Web resources. In recent years, many of academic research have been focused on social annotations [1]. The main purpose of these research works is to combine the analytical technique of social annotations with traditional information retrieval technologies to enhance the performance and user experience of information retrieval systems.

Generally, a social annotation is a triple consists of three elements: tag, user, and resource. In the triple of a social annotation, the resource is tagged by user based on the resource content; the tag (also referred as annotation) reflects the category of the resource; the user, by which the tag is provided, hides the latent social community. The semantic correlation among of elements of the triple can be utilized to improve the information retrieval system.

This paper aims to improve web page clustering accuracy for information retrieval system using social annotation. As the fundament of web page clustering, traditional vector space model (VSM) only takes the word occurrence frequency in the document into account. Due to the “Bag of Words” nature it cannot represent the contents of document more precisely. On the other hand, as a sort of user-provided metadata, social tag should be fully exploited to enhance the traditional VSM. Based on this consideration, we propose a social annotation based vector space model that makes use of the semantic of social annotations to cluster Web pages more accurately. In particular, our model is constructed by calculating the semantic correlations between social annotations and web page words. Different from the other enhanced VSM in which the words in a document and tags annotated to the document are simply treated with the same weighting, we project the calculated semantic correlations of social tags to the axis of vector space so that the axis with higher semantic correlation between tags and words has higher weight. This is the main contribution of our work.

Our work can be summarized as: we firstly calculate a semantic correlation matrix between social tags and web page words. Then a web page can be represented by a vector in three ways: (1) the axes of vector space are social tags and the coordinate of axis t for document d is determined by the semantic correlations between document d and tag t, which is denoted by $P(d, t)$. (2) the axes are web page words and the coordinate of axis w for document d is determined by the semantic correlations between document d and word w, which is denoted by $P(d, w)$. (3) the axes are words plus tags and the coordinates is weighted sum of $P(d, t)$ and $P(d, w)$. In order to evaluate the performance of our method, we compare it with other vector space models with K-means clustering algorithm. Our experiments show that our proposed model can improve the clustering accuracy by 4% ~ 7%.

Different from the existing works, the main contributions of our work are: (1) we exploited the semantic correlations between social tags and the words of web pages.

(2) we proposed a new vector space model by making full use of exploited semantic correlations of social annotations. This work will be an effort to help clustering the tagged web.

The rest of this paper is organized as follows: Section 2 introduces the related research work of social annotations. Section 3 proposes an extended vector space model by exploiting semantic correlations of social annotations. Experiments and numerical results are presented in Section 4. The conclusion is drawn in Section 5.

2 Related Work

Applying social annotations to web-based information systems is a hot topic and has been received a lot of attentions in recent years [2]. We briefly summarize the existing related works into the following three directions.

2.1 Language Models of Social Annotation

The research of probabilistic language model of information retrieval system is based on the idea which is originally proposed by Pronte and Croft [3]: Given a specific query, a document is a good match to query if the generative probabilistic language model of for the document is more likely to generate the query. Along with the emergence of social annotation, later research works on probabilistic language model seek to combine the probabilistic model of social annotation and annotated document content. Zhou et al. proposed a unified framework to combine the modeling of social annotations with the language modeling-based methods for information retrieval. The framework enhances document and query language models by incorporating user domain interests as well as topical background models [1]. Xu et al. analyzed two properties of social annotation, namely keyword property and structure property. These two properties of social annotations are leveraged for the use of language modeling with a mixture language model LAM (Language Annotation Model) and LAM is utilized to strengthen existing smoothing methods for the language model for information retrieval [4] [5].

2.2 Semantics of Social Annotation

The semantics of social annotations are captured by the following ways: calculating similarity between social annotations, establishing probabilistic model of social annotations and measuring the relationships among social annotations or annotated Web resources. Wu et al. established a global semantic model from social annotations which can be inferred from social annotations statistically [2]. Markines et al. evaluated similarity measures for emergent semantics [6] [7] of social annotations by building an evaluation framework to compare various general social annotations-based similarity measures statistically [8] [9]. Cattuto et al. analyzed several measures of tag similarity and used validated measures of semantic distance to characterize the semantic relation between the mapped tags [10].

2.3 Social Annotation Applications

The probabilistic language model and semantics of social annotations can be employed in many aspects of information retrieval systems such as ranking, classification, clustering to improve the effectiveness and user experience.

Similarity ranking measures the relevance between user query and resulting Web resources. Some proposed ranking models seek to improve query-resource similarity more precisely by making use of social annotation. Bao et al. proposed two novel algorithms SocialSimRank and SocialPageRank to incorporate social annotations into search result ranking [11]. Liu et al. proposed a tag ranking scheme to automatically rank the tags of images. Tag ranking scheme is applied to tag-based image search [12]. Schenkel et al. are focused on search ranking in social networks [13].

Since Social annotation provides a natural way for people to classify Web resources, some other research works try to explore this feature of social annotation to enhance the accuracy of Web resource classification and clustering. Pedro and Siersdorfer proposed a novel multi-modal approach for automatically ranking and classifying photos by exploiting image features and social annotations [14]. M.G. Noll and C. Meinel explored and studied the characteristics of social annotations with regard to their usefulness for Web document classification [15]. Shepitsen et al. presented a personalization algorithm for recommendation in folksonomies which relies on hierarchical social annotations clustering [16] [17]. Begelman et al. clustered social tags by defining a set of similarity measures among tags [18]. Ramage et al. used social annotations as complementary data source to improve automatic clustering of Web pages by means of combining social annotations with Web pages in an extended vector space model [19].

3 Vector Space Model Based on Semantic Correlation of Social Annotation

The objective of our work is to investigate how to exploit the semantic relationship between social annotations and annotated documents and how this semantic relationship can be employed to enhance vector space model so that the accuracy of web page clustering can be improved. Our work is mainly inspired by the previous work [16] [17] and [19].

In research work of [16] and [17], a hierarchical tag clustering algorithm is proposed based on tag's vector space model. In this tag-based vector space model, an annotated Web resource is modeled as a vector over the set of tags, and a user is modeled in the same way. Each component of a vector is calculated based on the well-known TF-IDF measure [20] of tags. To calculate the similarity between a user query and a Web resource, a query is also modeled as a unit vector consisting of a single tag, which is based on the assumption that the user interacts with the system by selecting a query tag and expects to receive resource recommendations. We argue that the content of annotated Web resource should also been considered in order to model the annotated web resource in tag-based vector space model more precisely, and in the case that user query consists of arbitrary terms, how to model user query in tag-based vector space model is still an unsolved problem.

On the other hand, the research work [19] proposed an extended vector space model, which is constructed in following ways with bags of words and tags: Words only, Tags Only, Words combining Tags. The combination of words and tags can be concatenation of l_2-normalized word and tag vector with equal weight, treating term of tag as n terms of word or treat tags simply as additional words. We also argue that the combination of words and tags should consider the correlation between words and tags rather than treating them with equal weight. Another problem of word term based vector space model is that a document may contain many informationless terms, and these informationless terms are more likely to act as noises which will increase the errors of clustering.

3.1 Problem Definition

Since an annotated web page is associated with a group of tags, we describe an annotated web page D as a tuple:

$$D = < W, T > \tag{1}$$

where $W = \{w_1, w_2, \dots, w_l\}$ is the set of words occurred in the web page D, $T = \{t_1, t_2, \dots, t_m\}$ is the set of social tags of the web page D.

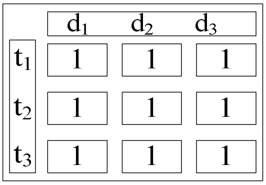
With this document description, the goal of social annotation based web page clustering task can be defined as: Given a set of documents $\boldsymbol{\mathcal{D}} = \{D_1, D_2, \dots, D_N\}$ and a target number of clusters K, we want to find an assignment function C:

$$C : \mathcal{D} \longrightarrow \{1,2, \dots, K\} \tag{2}$$

which maps each document in $\boldsymbol{\mathcal{D}}$ to a cluster number $\mathcal{x} \in \{1,2, \dots, K\}$, where $D_i (i = 1, \dots, N)$ is represented in a bi-tuple defined in (1).

3.2 Weighted Matrix of Social Annotation

There are two kinds of elements related to the social annotations: web page and social tag. Their relationship can be described by weighted tag-document matrix, which is illustrated in Figure 1.



	d_1	d_2	d_3
t_1	1	1	1
t_2	1	1	1
t_3	1	1	1

Fig. 1. The illustration of the relationship between web page and tag

As shown in Figure 1, the relationship between web page and tag can be described by binary matrixes, in which "1" element of ith row and jth column represents a relationship exists between corresponding element i and j. However, the matrix doesn't consider the factor of frequency, that is, how many times a tag is annotated to a web page. In order to count the factor of frequency, the tf-idf measure is applied to the $m \times n$ tag-document matrix A_{TD}.

$$A_{TD} = \begin{bmatrix} A_{TD}(1,1) & \cdots & A_{TD}(1,n) \\ \vdots & \ddots & \vdots \\ A_{TD}(m,1) & \cdots & A_{TD}(m,n) \end{bmatrix} \tag{3}$$

where m is the number of social tags and n is the number of annotated web pages. The matrix element $A_{TD}(i,j)$ denotes the relationship between the *i*th social tag and *j*th web page, which can be obtained as:

$$A_{TD}(i,j) = tf_{t_i,d_j} \times idf_{t_i} \tag{4}$$

where

$$idf_{t_i} = log \frac{N}{df_{t_i}} \tag{5}$$

In formula (4) and (5), tf_{t_i,d_j} denotes how many times the tag t_i is applied to the web page d_j; idf_{t_i} is the measure of the importance of the tag t_i in which *N* is the total number of annotated web pages and df_{t_i} denotes the number of web pages to which the tag t_i is applied.

Because social tags can be viewed as the terms marking social categorization of web pages, so an alternative choice is to assign the same importance for each tag. In this case idf_{t_i} can be ignored and $A_{TD}(i,j)$ is obtained as:

$$A_{TD}(i,j) = tf_{t_i,d_j} \tag{6}$$

3.3 Tag Similarity

The similarity measure between tags can be elicited by exploiting underline tags co-occurrence of matrix A_{TD}, in which element $A_{TD}(i,j)$ is a relationship measure for tag i and document *j*. Based on the assumption that semantically similar annotations are more likely assigned to the same documents, we can derive similarity matrix of tags from tag-document matrix A_{TD}, and we denote the similarity matrix of tags derived from A_{TD} by S_T:

$$S_T = A_{TD} \times {A_{TD}}^T \tag{7}$$

where the element $S_T(i,j)$ measure the similarity between tag t_i and tag t_j.

3.4 Semantic Correlation between Tags and Words

The matrix A_{TD} can be viewed as a semantic correlation matrix between social tags and documents. However, A_{TD} is only based on the statistical feature of tags. In order to characterize semantic correlation between tags and documents more precisely, the contents of documents should also be exploited.

An intuitive approach to exploit the contents of a document is based on the total times a tag annotated to the document. But this approach ignores the semantic correlation between the tag and word occurred in document. Our approach to calculate the semantic correlation between tag and document is based on the tag-word correlation matrix.

In vector space mode, each document is viewed as vector in which each component corresponds to a word in dictionary with TF-IDF weight. We denote the $l \times n$ word-document matrix as A_{WD}:

$$A_{WD} = \begin{bmatrix} A_{WD}(1,1) & \cdots & A_{WD}(1,n) \\ \vdots & \ddots & \vdots \\ A_{WD}(l,1) & \cdots & A_{WD}(l,n) \end{bmatrix} \tag{8}$$

where l is the number of words in dictionary and n is the number of web pages. The word w_j can be represented as the *j*th row vector $\overrightarrow{V_{WD}}(w_j)$ in A_{WD}:

$$\overrightarrow{V_{WD}}(w_j) = [A_{WD}(j,1), \dots, A_{WD}(j,n\ )] \tag{9}$$

Similarly, the tag t_i can be represented as the *i*th row vector $\overrightarrow{V_{TD}}(t_i)$ in A_{TD}:

$$\overrightarrow{V_{TD}}(t_i) = [A_{TD}(i,1), \dots, A_{TD}(i,n)] \tag{10}$$

Since $\overrightarrow{V_{WD}}(w_j)$ and $\overrightarrow{V_{TD}}(t_i)$ are vectors in document based vector space, the semantic correlation between tag t_i and word w_j, which is denoted by $S_{TW}(t_i, w_j)$, can be measured as:

$$S_{TW}(t_i, w_j) = \overrightarrow{V_{TD}}(t_i) \times \overrightarrow{V_{WD}}(w_j)^T \tag{11}$$

Finally, we denote the semantic correlation matrix between tag and word by S_{TW}, and S_{TW} can be calculated as:

$$S_{TW} = A_{TD} \times A_{WD}{}^T \tag{12}$$

where the element $S_{TW}(i,j)$ measure the semantic correlation between tag t_i and word w_j.

3.5 Extended Vector Space Model

Based on the similarity matrix of tags and the semantic correlation matrix between tag and word, we can model web page as a vector in tag-based vector space, in which each social tag is axis and each component of the vector is determined by the projection of the web page with respect to each tag.

The projection of the web page with respect to each axis is calculated according to semantic correlation between web page and tag. Given a web page *d* and tag *t*, we denote the projection of document *d* with respect to axis *t* by $P(d,t)$. The vector of document *d*, denoted by $\overrightarrow{V_T}(d)$, is defined as:

$$\overrightarrow{V_T}(d) = [P(d,t_1), \dots, P(d,t_m)] \tag{13}$$

where *m* is total numbers of tags, and each component of $\overrightarrow{V_T}(d)$ is calculated by $P(d,t)$:

$$P(d,t) = \mu \times A_{DT}(d,\mathrm{t}) + \varphi \times C(d,t) \tag{14}$$

μ and φ are all real constant between 0 and 1. Formula (14) shows that $P(d,t)$ is the linear combination of two parts: the first part, which denoted by $A_{TD}(t,d)$, is determined by how many times the tag *t* is applied to the document *d*; another part, denoted by $C(d,t)$ is determined by the semantic correlation between tag *t* and document *d*.

Before giving the definition of $C(d,t)$, we firstly denote the set of tags by T and the set of words occurred in web pages by *W*, thus $C(d,t)$ is defined as:

$$C(d,t) = \sum_{w \in W} Sim(w,t) \tag{15}$$

and $Sim(w,t)$ is defined as:

$$Sim(w,t) = \begin{cases} S_T(w,t), & w \in T \\ S_{TW}(w,t), & w \in W - T \end{cases} \tag{16}$$

Similarly, we can also model web page as a vector in word-based vector space, in which each word occurred in web pages is axis and each component of the document vector is determined by the projection of the document with respect to each word. Given a document *d* and word *w*, we denote the projection of document *d* with respect to axis *w* by $P(d,w)$. The vector of document *d*, denoted by $\overrightarrow{V_W}(d)$, is defined as:

$$\overrightarrow{V_W}(d) = [P(d,w_1),\dots,P(d,w_n)] \tag{17}$$

where *n* is total numbers of words, and $P(d,w)$ is calculated by:

$$P(d,w) = \mu \times A_{DW}(d,\mathrm{w}) + \varphi \times C(d,w) \tag{18}$$

where A_{DW} is the transpose of A_{WD} and $C(d,w)$ is calculated by

$$C(d,w) = \sum_{t \in T} Sim(w,t) \tag{19}$$

Another way to represent a web page is to combine the projection of the web page with respect to tag and word. Given a web page *d*, we calculate tag-based vector $\overrightarrow{V_T}(d)$ and word-based vector $\overrightarrow{V_W}(d)$, respectively. Then the web page d is modeled by a vector $\overrightarrow{V_{T+W}}(d)$, which is calculated by:

$$\overrightarrow{V_{T+W}}(d) = \mu \times \overrightarrow{V_T}(d) + \varphi \times \overrightarrow{V_W}(d) \tag{20}$$

When combining the $\overrightarrow{V_T}(d)$ and $\overrightarrow{V_W}(d)$ vector, we need to utilize feature selection algorithm to select words and tags and make the number of words and tags equal to each other.

4 Experiments and Numerical Results

In order to investigate the effectiveness of the clustering method under our social annotation based vector space model, we apply flat clustering algorithm K-means to our model to evaluate clustering accuracy. The evaluation is done by extensive experiments on real world data collections.

4.1 Data Collections

The target data collection is partly crawled from the social annotation site del.icio.us during July, 2010. Our data collections consist of 1502825 tags and 466871 web pages.

4.2 Evaluation Measure and Gold Standard

To evaluate the accuracy of clustering, we use subset of partly crawled data collection from del.icio.us which is also presented in Open Directory Project (ODP). ODP is the largest, most comprehensive human-edited directory of the Web. For every web page presented in ODP, we use the root category of that page as our evaluation gold standard. The evaluation measure we adopted is Rand Index (RI) [21]. RI penalizes both false positive and false negative decisions during clustering and it measures the accuracy of clustering result. If define *A* as the number of true positive documents, *B* as the number of false negative documents, *C* as the number of false positive documents, *D* as the number of true negative documents, and *A*+*B*+*C*+*D* is the total number of document ,then RI is defined by

$$RI = \frac{A+D}{A+B+C+D} \tag{21}$$

4.3 Experiment Settings

The total category of crawled web page presented is 17 and we select 100, 200 and all documents in each category for evaluating. We also use feature selection algorithm to select 1000, 1500, 2000 tags for testing, respectively. And the number of words is set to be equal to the number of tags. The parameters μ and φ are all set to 0.5 in formula (14), (18), and (20). For the performance comparing, we also apply the K-means to the following models:

(1) **Tag-only:** a document is modeled by tag-based tf-idf weighting vector.

(2) **Word-only:** a document is modeled by word-based tf-idf weighting vector.

(3) **Tag+Word:** a document is modeled by concatenation of l_2-normalized word and tag vector with equal weight, which is proposed by Ramage et al. [19] and used as accuracy comparing baseline with our work.

(4) **V_T:** a document is modeled by the projection of web page with respect to each tag, which is defined by formula (13)-(16).

(5) **V_W:** a document is modeled by the projection of web page with respect to each word, which is defined by formula (17)-(19).

(6) **V_{T+W}:** a document is modeled by combining the projection of web page with respect to tag and word, which is defined by formula (20).

4.4 Experimental Results and Analysis

Our first experiment is to randomly select 100 documents from each category in our data collections. Based on selected documents, we perform K-means tests for six different vector models described in Section 4.3. For each of tests, we use RI values to measure our clustering performance.

As shown in Table 1, Tag-only has the worst RI value, which means Tag-only has the lowest clustering accuracy. The reason is that social tags have limited information and bad quality. Another possible reason is social tags are so sparse that it is bad for our clustering performance. For Word-only and Tag+Word, both of them have better RI values than Tag-only under our expectations because they have detailed information for documents. More importantly, our proposed V_{T+W} model improves RI value further beyond Word-only and Tag+Word. The main reason is that V_{T+W} model considers the semantic correlations between social tags and web page words and it has much richer semantic information than other models. Another reason is there are noise words in either Word-only or Tag+Word models. Compared to other models, our model has up 29% better RI than Word-only and 22% better RI than Tag+Word. The increased value is calculated by averaging RI increase with tags number 1000, 1500 and 2000.

Table 1. Select 100 documents randomly from every category

	tag, word = 1000	**tag, word = 1500**	**tag, word = 2000**
Tag-only	0.1148	0.1064	0.1002
Word-only	0.7125	0.5533	0.7331
Tag+Word	0.8065	0.6502	0.6411
V_T	0.8458	0.8361	0.8380
V_w	0.6831	0.5925	0.6337
V_{T+W}	0.8450	0.8370	0.8553

Similarly, we picked up more documents for each category and conducted the second experiments. The result is shown in Table 2. Compared to other models, our model has up 42% better RI than Word-only and 34% better RI than Tag+Word averagely. Again, our model shows much better RI values than other models.

Table 2. Select 200 documents randomly from every category

	tag, word = 1000	**tag, word = 1500**	**tag, word = 2000**
Tag-only	0.0947	0.0899	0.0840
Word-only	0.8216	0.6388	0.6161
Tag+Word	0.7731	0.6342	0.5813
V_T	0.8574	0.8620	0.8635
V_w	0.8349	0.7483	0.7175
V_{T+W}	0.8677	0.8727	0.8765

Besides investigating how V_{T+W} can achieve much better clustering accuracy, we also conduct an experiment how sampling documents affect clustering accuracy. Without sampling documents, we select all the documents to conduct an experiment as shown in Table 3. This table shows that our model has 2% better RI value than Word-only averagely, and 4% ~7% better RI value than Tag+Word .

Table 3. Select all documents from every category

	tag, word = 1000	tag, word = 1500	tag, word = 2000
Tag-only	0.1402	0.1623	0.1317
Word-only	0.8277	0.8201	0.8262
Tag+Word	0.8163	0.7857	0.7809
V_T	0.8316	0.8333	0.8291
V_w	0.8230	0.8131	0.7676
V_{T+W}	0.8484	0.8400	0.8327

From Table 1, 2, and 3, one can find increasing samples can benefit RI values. However, when oversampling documents, it will be against the RI values. The possible reason is that some samples have noises information. This observation is consistent with the conclusion that noisy information is not so good for the improvement of clustering performance [22].

5 Conclusion

Social annotations of web pages contain usefully semantic information and thus can be utilized to improve the performance of information retrieval system. In this paper we exploit the semantic of social annotations by calculating the semantic correlation between social annotations and words in web pages. Furthermore, we propose a novel vector space model based on semantic correlation between social annotations and words, and apply the K-means clustering algorithm to our model to evaluate the clustering accuracy. Comparing with other vector space models, our model can achieve the best clustering accuracy by 4% ~ 7%.

Acknowledgments. This work is supported by National Natural Science Foundation of China under Grant 60873225, 60773191, 70771043, National High Technology Research and Development Program of China under Grant 2007AA01Z403.

References

1. Zhou, D., Bian, J., Zheng, S., Zha, H., Giles, C.L.: Exploring social annotations for information retrieval. In: The 17th International Conference on World Wide Web (WWW 2008), pp. 715–724. ACM Press, Beijing (2008)
2. Wu, X., Zhang, L., Yu, Y.: Exploring social annotations for the semantic web. In: The 15th International Conference on World Wide Web (WWW 2006), pp. 417–426. ACM Press, Edinburgh (2006)
3. Ponte, J.M., Croft, W.B.: A Language Modeling Approach to Information Retrieval. In: The 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998), pp. 275–281. ACM Press, Melbourne (1998)
4. Xu, S., Bao, S., Cao, Y., Yu, Y.: Using social annotations to improve language model for information retrieval. In: The 16th International ACM Conference on Conference on Information and Knowledge Management (CIKM 2007), pp. 1003–1006. ACM Press, Lisboa (2007)

5. Xu, S., Bao, S., Yu, Y., Cao, Y.: Using Social Annotations to Smooth the Language Model for IR. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 1015–1021. Springer, Heidelberg (2007)
6. Aberer, K., Cudré-Mauroux, P., Hauswirth, M.: The chatty web: emergent semantics through gossiping. In: The 12th International Conference on World Wide Web (WWW 2003), pp. 197–206. ACM Press, Budapest (2003)
7. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)
8. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Stumme, G.: Evaluating similarity measures for emergent semantics of social tagging. In: The 18th International Conference on World Wide Web (WWW 2009), pp. 641–650. ACM Press, Marid (2009)
9. Markines, B., Menczer, F.: A scalable, collaborative similarity measure for social annotation systems. In: The 20th ACM Conference on Hypertext and Hypermedia, pp. 347–348. ACM Press, Torino (2009)
10. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
11. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: The 16th International Conference on World Wide Web (WWW 2007), pp. 501–510. ACM Press, Banff (2007)
12. Liu, D., Hua, X., Yang, L., Wang, M., Zhang, H.: Tag ranking. In: The 18th International Conference on World Wide Web (WWW 2009), pp. 351–360. ACM Press, Marid (2009)
13. Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Weikum, G.: Efficient top-k querying over social-tagging networks. In: The 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), pp. 523–530. ACM Press, Singapore (2008)
14. Pedro, J.S., Siersdorfer, S.: Ranking and classifying attractiveness of photos in folksonomies. In: The 18th International Conference on World Wide Web (WWW 2009), pp. 771–780. ACM Press, Marid (2009)
15. Noll, M.G., Meinel, C.: Exploring social annotations for web document classification. In: The 2008 ACM Symposium on Applied Computing (SAC 2008), pp. 2315–2320. ACM Press, Brazil (2008)
16. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.D.: Personalized recommendation in social tagging systems using hierarchical clustering. In: The 2008 ACM Conference on Recommender Systems, pp. 259–266. ACM Press, Lausanne (2008)
17. Gemmell, J., Shepitsen, A., Mobasher, B.: Personalization in Folksonomies Based on Tag Clustering. In: The AAAI 2008 Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, Chicago, pp. 37–48 (2008)
18. Begelman, G., Keller, P.: Automated Tag Clustering: Improving Search and Exploration in the Tag Space. In: The 15th International Conference on World Wide Web (WWW 2006), Workshop on Collaborative Web Tagging, Edinburgh, UK (2006)
19. Ramage, D., Heymann, P.: Clustering the Tagged Web. In: The Second ACM International Conference on Web Search and Data Mining, pp. 54–63. ACM Press, Barcelona (2009)
20. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management: an International Journal 24, 513–523 (1988)
21. Yeung, K.Y., Ruzzo, W.L.: Principal component analysis for clustering gene expression data. Bioinformatics 17, 763–774 (2001)
22. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press, Cambridge (2008)