Improving Item Recommendation Based on Social Tag Ranking

Taiga Yoshida[1], Go Irie[1], Takashi Satou[1],
Akira Kojima[2], and Suguru Higashino[1]

[1] NTT Cyber Solutions Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka, Kanagawa, Japan
[2] NTT Cyber Space Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka, Kanagawa, Japan
{yoshida.taiga,irie.go,satou.takashi,
kojima.akira,higashino.suguru}@lab.ntt.co.jp

Abstract. Content-based filtering is a popular framework for item recommendation. Typical methods determine items to be recommended by measuring the similarity between items based on the tags provided by users. However, because the usefulness of tags depends on the annotator's skills, vocabulary and feelings, many tags are irrelevant. This fact degrades the accuracy of simple content-based recommendation methods. To tackle this issue, this paper enhances content-based filtering by introducing the idea of tag ranking, a state-of-the-art framework that ranks tags according to their relevance levels. We conduct experiments on videos from a video-sharing site. The results show that tag ranking significantly improves item recommendation performance, despite its simplicity.

Keywords: recommendation, content-based filtering, tag ranking.

1 Introduction

The number of multimedia contents on the Web is dramatically increasing. This is making it more and more difficult for users to find interesting items. Many recommendation approaches have been proposed to support users in reaching their goal. The most popular approach, collaborative filtering, measures the similarity between items based on users' logs [1] [2]. If the log amount is sufficient, collaborative filtering works well. However, it fails when the log amount is small[3].

Content-based filtering that measures the similarity between item contents is a promising approach for resolving this issue. Multimedia contents sharing web services such as YouTube[1] and Flickr[2] use keywords associated with item

[1] http://www.youtube.com/
[2] http://www.flickr.com/

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 161–172, 2012.

roger federer novak djokovic us u.s. open 2009 tennis spectacular shot amazing phenomenal sport sports ny new york flushing meadows point match crazy

Fig. 1. Tags associated with a YouTube video

meta-data called "tags" for classification of items. Typically, these methods measure the similarity between a pair of items based on the tags associated with them. For instance, the number of co-occurring tags can be used [4]. As two items share a greater number of common tags, the similarity between them increases.

However, using tags in a naïve manner does not always work well because some are irrelevant to the item [5]. Because tags are provided by users, their quality depends on the users skills, vocabulary, and feelings. Fig. 1 shows an example of a list of tags associated with the video uploaded to YouTube titled "Federer Amazing Shot at the US Open 2009 Semifinal"[3]. For instance, "2009" and "crazy" are clearly less relevant to the video than "tennis" and "federer". The first two degrade the performance of item recommendation, because items sharing such irrelevant tags are not similar.

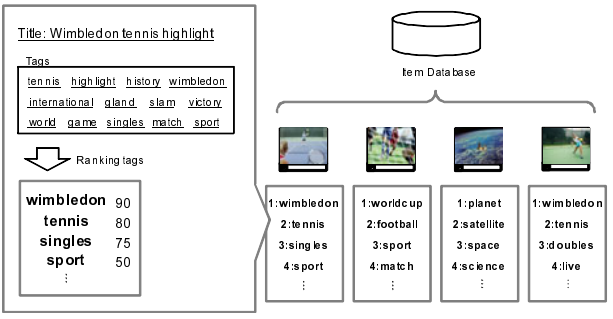
Recently, Liu [6] et al. reported that the performance of auto-tagging and image search can be improved by ranking the tags associated with an image according to their relevance levels to the image content. The idea is very simple and suitable for practical usage.

We here raise a question: is the tag ranking approach effective in the context of item recommendation? If tag ranking is effective for item recommendation, degradation by irrelevant tags can be suppressed in a simple manner. In this paper, we introduce tag ranking into content-based filtering. A ranking of tags is created based on item relevance, and the similarity between items is determined by comparing tag rankings of items. For validating the effectiveness of tag ranking, we conducted some experiments. The results show that tag ranking is effective for recommendation, and content-based filtering can be improved simply by introducing tag ranking.

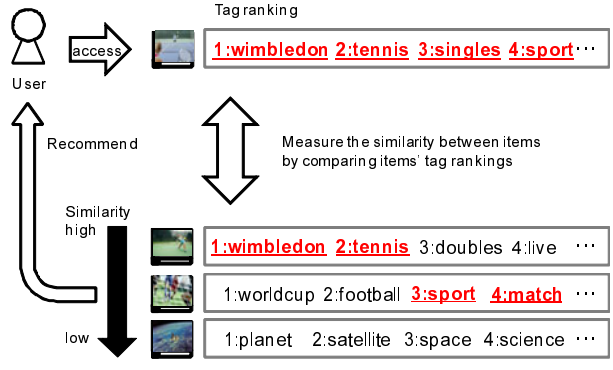
2 Recommendation Based on Tag Ranking

Fig. 2 overviews a recommendation scheme based on tag ranking. At first, the tags set on each item are ranked according to their importance. If items share several tags that have high ranking, the similarity between them is assessed to be high. Items that are highly similar to those in a user's log are recommended to the user.

[3] http://www.youtube.com/watch?v=RJuEzJEQ9N4



(a) Creating tag ranking on each item



(b) Measuring similarities between items based on tag ranking

Fig. 2. Overview of recommendation based on tag ranking

2.1 Creating Tag Ranking

There are several approaches on which tag ranking can be based. One simplest approach is TF-IDF [7], a general indicator of the importance of keywords based on the frequency of the keyword in the item (TF) and the inverse of the number of items that were assigned the keyword (IDF). In the case of social tagging, the same tag is not usually associated with an item more than once, so TF value is always $\{0, 1\}$. The IDF value of a keyword rises if it is contained in fewer items. However, such rare tags are not always relevant to the item. Therefore, TF-IDF is not appropriate for social tags.

The tag ranking proposed by Liu [6] estimates tag relevance levels by applying a probabilistic method and random walk-based refinement. This approach is effective for ranking tags according to their relevance to items. However, it requires image features for ranking tags, so some modifications would be needed to apply this approach to other media such as videos.

A more typical approach for capturing the relevance of tags is co-occurrence. This approach is based on the idea that semantically related tags co-occur frequently. For instance, "wimbledon" frequently co-occurs with "tennis", because these tags are related to each other. This co-occurrence based approach is very simple and is applicable to various media, so we adopt it for tag ranking. We extract two different bits of information from tag co-occurrence data: Co-occurrence Depth and Co-occurrence Width.

If many tags that are semantically related to each other are associated with an item, they may be important tags for it. This is because if these tags are strongly related to the item, users may associate it with the item even when other semantically related tags are already associated with the item.

Co-occurrence Depth scores are calculated based on the co-occurrence between tags associated with the same item. As an example, we describe the calculation of Co-occurrence Depth score for the tags associated with the item "Wimbledon tennis highlight". Some tags such as "tennis", "wimbledon", "game", "singles" or "sport" are associated with the item. "Tennis" often co-occurs with many tags associated with the same item such as "wimbledon" , "singles" or "sport". On the other hand, "game" often co-occurs with tags for video games and more rarely found with tags about tennis. The score of "tennis" should be higher than that of "game", because "tennis" is more relevant to the item than "game". Co-occurrence Depth score increases if the tag co-occurs with tags on the same item more frequently.

We denote a certain item in the item database as i_n, and tags associated with i_n as $T^{i_n} = \{t_m^{i_n} | m = 1, 2, ..\}$. Co-occurrence Depth score $S_d(i_n, t_m^{i_n})$ of tag $t_m^{i_n}$, which is associated with item i_n, is calculated as follows.

$$S_d(i_n, t_m^{i_n}) = \sum_{t_l \in T^{i_n} s.t. t_l \neq t_m^{i_n}} C(t_l, t_m^{i_n}) \tag{1}$$

$C(t_x, t_y)$ is a chi-square value whose null hypothesis is that tags appear independently in the item database. The value of $C(t_x, t_y)$ increases when tags t_x and t_y have high positive correlation. If there is a negative correlation between t_x and t_y, the sign of $C(t_x, t_y)$ flips to minus.

By calculating scores based on only Co-occurrence Width, both scores of "wimbledon" and "sport" can be high in the same way. However, "wimbledon" is more relevant to the item than "sport", because "wimbledon" is more specific than "sport". We also use the specificity of the tag for creating tag ranking.

Co-occurrence Width scores are calculated based on the variety of co-occurring tags in the item database. For instance, the score of "wimbledon" should be high because it is associated with items only about tennis and the variety of co-occurrence tags is small. On the other hand, the score of "sport" should be low because it is also associated with items other than tennis such as football, baseball, golf and so on. Co-occurrence Width score increases if the tag co-occurrence with other tags in the database do not vary widely.

The Co-occurrence Width score of a tag is calculated from entropy, which is an indicator of degree of variability. Co-occurrence Width score $S_w(t_m^{i_n})$ of tag $t_m^{i_n}$ is calculated as follows.

$$E(t_m^{i_n}) = -\sum_{t_l \in T_m^{i_n}} \frac{N_{t_m^{i_n},t_l}}{N_m^{i_n}} \log \frac{N_{t_m^{i_n},t_l}}{N_m^{i_n}} \tag{2}$$

$$S_w(t_m^{i_n}) = e^{-E(t_m^{i_n})} \tag{3}$$

$T_m^{i_n}$ is a set of tags associated with the same items with $t_m^{i_n}$. $N_{t_m^{i_n},t_l}$ is the number of items with which both $t_m^{i_n}$ and t_l are associated. $N_m^{i_n}$ is calculated by $\sum_{t_l \in T_m^{i_n}} N_{t_m^{i_n},t_l}$.

Although tag ranking can be created from either of these co-occurrence scores, we simply combine them by multiplying them first. The importance score $S(i_n, t_m^{i_n})$ of tag $t_m^{i_n}$ associated with item i_n is calculated as follows.

$$S(i_n, t_m^{i_n}) = S_d(i_n, t_m^{i_n}) S_w(t_m^{i_n}) \tag{4}$$

Tag ranking for item i_n is created by ordering the tags associated with it in descending order of importance score.

2.2 Ordering Items Based on Tag Ranking

If two items share tags that are placed high in their tag rankings, those items may be similar. Thus similarity $R_i(i_m, i_n)$ between items i_m and i_n is calculated as follows.

$$R_i(i_m, i_n) = \sum_i \sum_j \frac{1}{ij} \delta(t_i^{i_m}, t_j^{i_n}) \tag{5}$$

$t_i^{i_m}$ is the i-th tag in the tag ranking associated with i_m. $\delta(t_i^{i_m}, t_j^{i_n})$ is a function whose value is 1 when $t_i^{i_m}$ is equal to $t_j^{i_n}$ and 0 otherwise. The value of $R_i(i_m, i_n)$ increases when i_m and i_n share many common tags with high ranking.

Items that have high similarity with items in the user's access log are recommended to the user. When items $I = \{i_k | k = 1, 2, ..\}$ are in the user's access log, the recommend score $R(I, i_n)$ of item i_n is calculated by the following equation.

$$R(I, i_n) = \sum_{i_k \in I} R_i(i_k, i_n) \tag{6}$$

Items are ordered in descending order of their recommend scores and recommended to the user.

3 Experiments

3.1 Experimental Conditions

We conducted experiments to validate the performance of tag ranking in the context of item recommendation. We used 14,159 videos downloaded from a

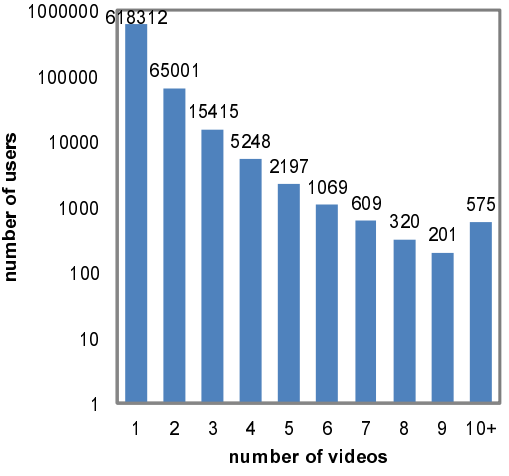


Fig. 3. Distribution of number of videos user viewed

popular video sharing site in the experiments. They contained 20 categories of genre, and one genre was assigned to each video. Evaluations of recommendation methods generally use the precision of predictions as discerned from users' access logs [9]. In this work, we performed an evaluation by taking the comment logs of users as access logs. In our dataset, 850,881 comments were attached to videos by 708,947 users. Fig. 3 shows the distribution of the number of videos viewed by individual users. In the experiments, we used the logs of 2,774 users, each of whom watched over 6 videos, for evaluating the dependency of precision on the amount of users' logs. We divided them into 1,387 learning users and 1,387 test users. Recommendation precision was taken to be the precision with which the user's 6th video (called target video) in the test user's log was predicted; each method assessed the 1st to the 5th item in the user's log to predict the 6th item, which is viewed next by the user. We evaluated the precision of each method by mean average precision (MAP) [10]. We compared the following 6 methods.

- **Content-based Methods**
  Tag-Rank: bases recommendations on tag ranking
  Jaccard: calculates similarities between items from Jaccard coefficients of tags
  Genre: recommends items whose genre is the same as the item randomly selected from the user's access log
- **Log-based Methods**
  Item-CF: calculates similarities between items using accessed user logs of items (item-based collaborative filtering)
  User-CF: calculates similarities between users using accessed item logs of users (user-based collaborative filtering)
  Ranking: recommends items that have higher rank in access ranking but have not yet been accessed by the user

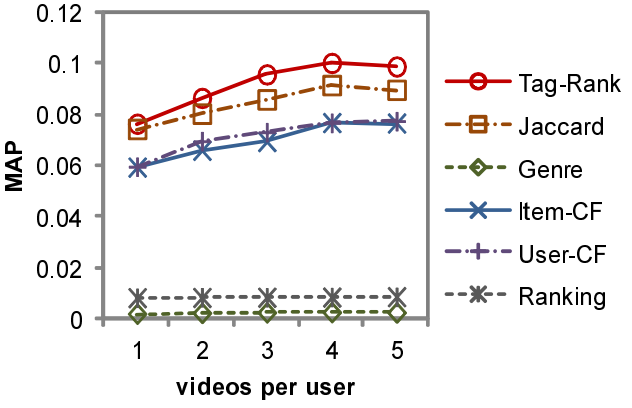


Fig. 4. MAP vs. user log number (single)

We conduct two different experiments. First, we compare each single method. Next, we validate accuracy when combining Tag-Rank with log-based methods. In each experiment, we measure MAP under the conditions of changing the number of items accessed by a user and MAP under the condition of changing the popularity of the target video.

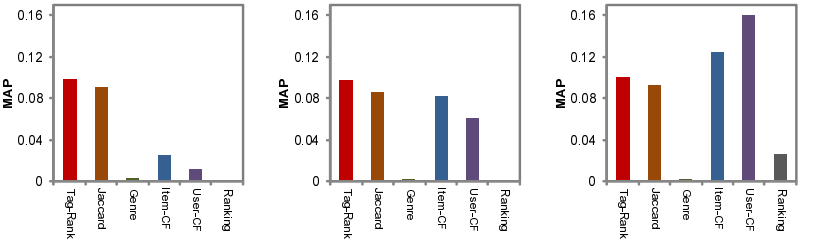
3.2 Comparing Single Methods

In the experiments, we compared three content-based methods and three log-based methods.

MAP vs. User Log Number. We evaluated MAP values of the target video while varying the number of entries in each user's log from 1 to 5. Fig. 4 shows the results of the experiment. The horizontal axis of the figure is the number of accessed items per user, and the vertical axis is MAP value.

The results show that Tag-Rank was the best of the 6 methods. The 3 log-based methods do not work when the users' access logs had few items. On the other hand, because Tag-Rank uses tags associated with videos for recommendation, it is able to measure similarity between videos precisely even when the users' access logs have few items. If each user's access log held many items, log-based methods are expected to top Tag-Rank. However, this situation is not common and Tag-Rank provides adequate performance for practical numbers of items. Moreover, precision can be improved by combining Tag-Rank with a log-based method as described below.

Among the 3 content-based methods, Tag-Rank and Jaccard are better than Genre. Since many videos were assigned the same genre, genre information was not discriminatory enough to assess the similarities between videos. Tag-Rank had higher MAP than Jaccard. We performed the Wilcoxon signed rank test for



(a) Tail (b) Mid (c) Head

Fig. 5. MAP vs. view number (single)

Tag-Rank and Jaccard while varying the items in each user's log from 1 to 5. There was a significant difference at the 1% significant level in all conditions.

MAP vs. View Number. We evaluated MAP values of target videos with different levels of popularity.

In this experiment, we divided test users into 3 groups according to the number of the target videos viewed.

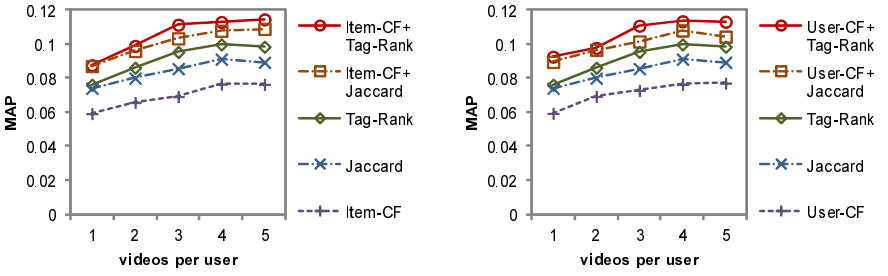
Tail: the target item is in the bottom one third of all videos
Mid: the target item is in the mid third of all videos
Head: the target item is in the top one third of all videos

We evaluated MAP values of the target videos in each group. Fig. 5 shows the results. The vertical axis is MAP value calculated from the 1st to the 5th videos in each user's log.

The results show that Tag-Rank had the best MAP values for the Tail group and the Mid group. User-CF had the best MAP values for the Head group. When using log-based methods for item recommendation, prediction accuracy is relative high if user logs contain many items, i.e. training data is sufficient. Since Tag-Rank does not depend on the number of times the target video is viewed, MAP offers high performance even when log amount is small.

Comparing the methods, Tag-Rank has higher MAP than Jaccard in all groups. Because Jaccard does not consider relevance levels of tags, it frequently recommends unsuitable videos. On the other hand, Tag-Rank emphasizes tags relevant to the video, so Tag-Rank offers high MAP. Item-CF and User-CF have high MAP when recommending videos in the Head group, but low MAP when recommending videos in the Tail group.

The precision of log-based methods varies according to log amount, but that of Tag-Rank is high and does not depend on item popularity. Tag-Rank is especially effective when recommending items with low view counts.



(a) Combination with Item-CF (b) Combination with User-CF

Fig. 6. MAP vs. user log number (hybrid)

3.3 Hybrid Methods

Combining log-based methods with content-based filtering approaches is a simple way of achieving high performance in a wider variety of situations [11] [12].

We evaluated MAP values when combining a content-based method: Tag-Rank or Jaccard with a log-based method: Item-CF or User-CF. The scores were calculated by summing up the normalized scores of the methods used.

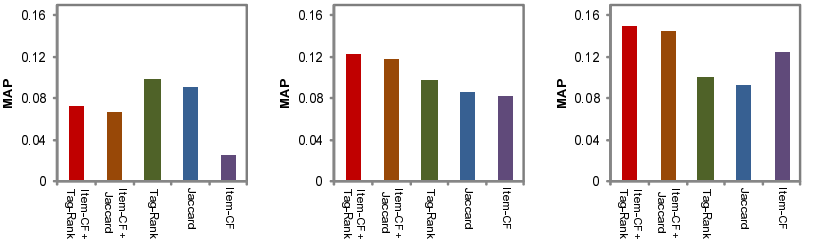
MAP vs. User Log Number. We evaluated MAP values of the target video while varying the number of items in each user's log from 1 to 5. Fig. 6 shows the results. The horizontal axis is the number of items accessed per user, and the vertical axis is MAP value.

The MAP values show that Tag-Rank combinations are superior to the Jaccard combinations in all conditions. Combining Item-CF or User-CF with Tag-Rank yields MAP values above the MAP values of the constituent methods used in isolation.

MAP vs. View Number. We also evaluated the MAP values of target videos with different levels of popularity. Fig. 7 shows the results for a content-based method with Item-CF. Fig. 8 shows the results for a content-based method with User-CF. The vertical axis is the MAP value when each user's log contains from 1 to 5 videos.

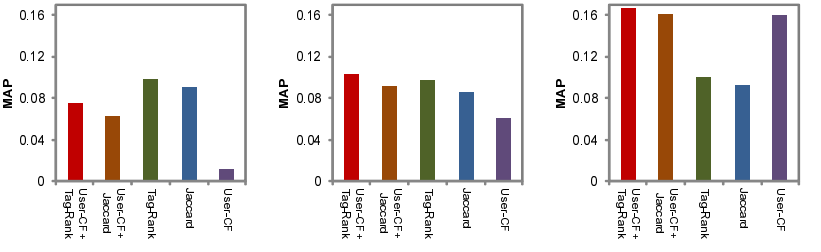
The results show that the Tag-Rank combinations are superior to the Jaccard combinations in all conditions. Tag-Rank well compensates the weak point of log-based methods with regard to recommending items in the Tail group. Even in the Mid group and the Head group, Tag-Rank can improve the MAP values of log-based methods.

From the results of the above experiments, we conclude that Tag-Rank improves recommendation performance despite its simplicity.



(a) Tail (b) Mid (c) Head

Fig. 7. MAP vs. view number (hybrid, combination with Item-CF)



(a) Tail (b) Mid (c) Head

Fig. 8. MAP vs. view number (hybrid, combination with User-CF)

3.4 Effect of Co-occurrence Depth Score and Co-occurrence Width Score

In this work, we create tag ranking based on two co-occurrence scores: Co-occurrence Depth score and Co-occurrence Width score. Co-occurrence Depth score indicates the relevance of the tag to the item. Co-occurrence Width score indicates the specificity of the tag. Tag ranking can be created based on combining them and also based on either of them. We compared the MAP value of each co-occurrence score when each user's log contained from 1 to 5 items. Fig. 9 shows the results.

For 1 or 2 items, Co-occurrence Width score yields higher MAP but above 2, the combination of Co-occurrence Depth score and Co-occurrence Width score offers the best performance. In this experiment the combination of co-occurrence scores is calculated by simple multiplication. For example, the MAP values might be improved by attaching a high weight to Co-occurrence Width scores when the log amount is small.

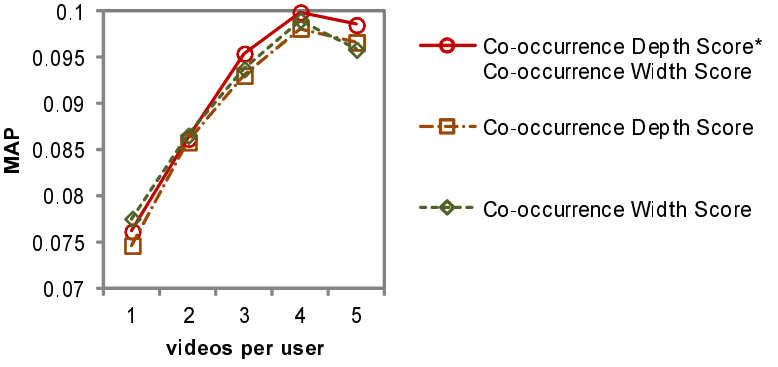


Fig. 9. Comparing co-occurrence scores

4 Conclusions

We proposed herein the idea of introducing tag ranking to improve recommendation precision. Tag ranking reflects tag importance as calculated by their co-occurrence. The similarity between items is measured by comparing their tag rankings. Items similar to those in the user's log are recommended to the user. In order to validate the effectiveness of tag ranking, we performed experiments on data from logs of a video sharing site. The experiments showed that our simple tag ranking approach can well improve the precision of content-based filtering. We also confirmed that the precision is improved by combining content-based methods with our proposed simple tag ranking method. We plan to validate the effect of tag ranking in detail by performing experiments on larger datasets and on other type of datasets. We also plan to examine other tag ranking methods.

References

1. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-Based Collaborative Filtering Recommendation Algorithms. In: 10th International Conference on World Wide Web (WWW), pp.285–295 (2001)
2. Breese, J.S., Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: 14th Conference on Uncertainty in Artificial Intelligence (UAI), pp. 43–52 (1998)
3. Maltz, D., Ehrlich, K.: Pointing the Way: Active Collaborative Filtering. In: SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 202–209 (1995)
4. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Stumme, G.: Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In: 18th International Conference on World Wide Web (WWW), pp. 641–650 (2009)
5. Golder, S., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)

6. Liu, D., Hua, X.-S., Yang, L., Wang, M., Zhang, H.-J.: Tag Ranking. In: 18th International Conference on World Wide Web (WWW), pp. 351–360 (2009)
7. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley (1989)
8. Sigurbjörnsson, B., Zwol, R.V.: Flickr tag recommendation based on collective knowledge. In: 17th International Conference on World Wide Web (WWW), pp. 327–336 (2008)
9. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems 22(1), 5–53 (2004)
10. Yates, R.-B., Neto, B.-R.: Modern Information Retrieval. Addison Wesley (1999)
11. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining Content-Based and Collaborative Filters in an Online Newspaper. In: ACM SIGIR Workshop on Recommender Systems (1999)
12. Pazzani, M.: A Framework for Collaborative, Content-Based, and Demographic Filtering. Artificial Intelligence Review, 393–408 (1999)