Collaborative Filtering for Social Tagging Systems: An Experiment with CiteULike

Denis Parra
School of Information Sciences
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
dap89@pitt.edu

Peter Brusilovsky
School of Information Sciences
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
peterb@ pitt.edu

ABSTRACT

Collaborative tagging systems pose new challenges to the developers of recommender systems. As observed by recent research, traditional implementations of classic recommender approaches, such as collaborative filtering, are not working well in this new context. To address these challenges, a number of research groups worldwide work on adapting these approaches to the specific nature of collaborative tagging systems. In joining this stream of research, we have developed and compared three variants of user-based collaborative filtering algorithms to provide recommendations of articles on CiteULike. The first approach, Classic Collaborative filtering (CCF) uses Pearson correlation to calculate similarity between users and a classic adjusted ratings formula to rank the recommendations. The second approach, Neighbor-weighted Collaborative Filtering, takes into account the number of raters in the ranking formula of the recommendations. The third approach explores an innovative way to form the user neighborhood based on a modified version of the Okapi BM25 model over users' tags. Our results suggest that both alterations of CCF are beneficial. Incorporating the number of raters into the algorithms leads to an improvement of precision, while tag-based BM25 can be considered as an alternative to Pearson correlation to calculate the similarity between users and their neighbors.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*information filtering*; Information Search and Retrieval–*selection process*.

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Collaborative-filtering, recommender systems, tagging.

1. INTRODUCTION

The new generation of collaborative tagging systems such as Delicious or CiteULike presented a new challenge to researchers and practitioners in the area of recommender systems. While both content-based [1] and collaborative filtering [2] recommender systems achieved remarkable success in traditional information repositories, social tagging systems may need novel approaches to recommendation. First of all, user-contributed content is more diverse in its nature and quality than the centrally created and structured content of traditional repositories. Second, traditional 5-10 point ratings are typically not available – only the fact that an item was contributed or bookmarked by the user is present in the system. At the same time, the loss of quality control and fine-grained ratings in collaborative tagging systems is compensated by the presence of tags and (in most systems) explicit connections between users. It is evident that recommendation approaches for collaborative tagging systems should capitalize on the success of classic recommender system, while trying to harness the new power provided by tags and social links. However, there is no shared understanding of how these features have to be taken into account to improve the quality of personalization. A few pioneer projects explored different ways to integrate social links or social tags into collaborative recommendation [3, 4, 6], and content-based recommendation [5] approaches. To some extent, the results are encouraging – both social links and tags do indeed improve the personalization quality. At the same time, the overall recommendation quality is unusually low – the precision for both content based and collaborative "tag-aware" recommendation reported in [4, 6] stays in the range of 0.1-0.3. The lack of reliable success calls for further research on recommendation in social tagging systems. This paper contributes to this stream of research by exploring two extensions of the traditional collaborative filtering approaches. First, we argue that the diverse user-contributed nature of content in collaborative tagging systems requires more evidence of relevance and quality than in traditional systems where the content is co-rated by the site developers. In this context, recommender algorithms should favor items bookmarked by more users. However, classic algorithms do not take the number of raters into account. Second, we argue that due to the large volume of items and low overlap between user bookmarks, the traditional approaches to neighborhood calculation may not be the most efficient. Two users who are very similar in their interests may still have too few common items bookmarked. In this context, tags applied by users can provide a more reliable approach to find similar users and this can be used to get better recommendation. To assess our hypotheses we developed variants of user-based collaborative filtering, which take into account the number of users who bookmarked an item and one approach uses tag-level similarity instead of traditional Pearson correlation to form user neighborhood.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
RecSys'09, October 23–25, 2009, New York, New York, USA.
Copyright 2009 ACM 978-1-60558-435-5/09/10...$10.00.

The remainder of this paper is structured as follows. Section 2 describes the characteristics of the data and how it was collected. Section 3 describes the three recommender approaches developed: Classic Collaborative Filtering (CCF), Neighbor-weighted Collaborative Filtering (NwCF) and BM25-based similarity (BM25). In section 4 we describe the study conducted and present the results. Section 5 introduces relevant related work, in Section 6 we address the discussion and in Section 7 we summarize conclusions and future work.

2. DATASET

We performed our study based on data that we *crawled* from CiteULike[1]. The daily datasets provided by CiteULike lack a lot of relevant information necessary to develop our algorithms, as the title and the authors of each article.

We selected a group of users to be our *center users*, i.e., those who would receive the recommendations. For each one of these center users, we *crawled* their posted articles (id, title, authors, post timestamp, and tags associated), the neighborhood of users who posted her same articles, and the neighborhood of users who share the same tags. To avoid limiting the neighborhood due to tag variations as hyphens, underscores and plurals, we enhanced the spreading of tags by adding stemmed tags using Krovetz algorithm, and modified tags changing hyphens and underscores to eventually be added to the set of tags to be crawled.

The details of the final dataset are described in Table 1. We chose 10 center users and we crawled all their articles, tags and potential neighbors. In total, we crawled 5,118 users. We ensured that all the potential neighbors, i.e., users sharing any of the articles or tags of our 10 centers users, were part of the dataset. For each of these neighbors we also crawled all their articles and tags. In Table 1, annotations correspond to tuples of the style {user, article, tag}

Table 1. Description of the dataset

Item	# of unique instances
users	5,849
articles	574,907
tags	139,993
annotations	2,337,571

3. ALGORITHMS

To create user-based recommendations using collaborative filtering, two steps are necessary. The first step is finding the neighborhood of the center user, i.e., a set of the most similar users. Once the most similar users are identified, the second step is to rank the articles to be recommended. These articles will be taken from the set of articles which the neighbors rated positively, and which the center user has not posted yet.

We implemented three variations of user-based collaborative filtering approaches: Classic Collaborative Filtering (CCF), Neighbor-weighted Collaborative Filtering (NwCF) and BM25-based similarity.

[1] www.citeulike.org

3.1 Classic Collaborative Filtering (CCF)

This approach is described in detail in [2]. In the classic CF model, the similarity between two users is calculated using the Pearson correlation over the ratings of their common items. The formula for the Pearson correlation, as stated in [2], is:

$$userSim(u,n) = \frac{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)(r_{ni} - \bar{r}_n)}{\sqrt{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \subset CR_{u,n}} (r_{ni} - \bar{r}_n)^2}} \quad (1)$$

In the formula, r stands for rating, u denotes the center user and n a neighbor. $CR_{u,n}$ denotes the set of co-rated items between u and n. After performing this calculation, we select the top ten most similar users. Next, we rank the articles of these users to recommend to the center user, using the formula of predicted rating for user u with average adjusts described in [2]

$$pred(u,i) = \bar{r}_u + \frac{\sum_{n \subset neighbors(u)} userSim(u,n) \cdot (r_{ni} - \bar{r}_n)}{\sum_{n \subset neighbors(u)} userSim(u,n)} \quad (2)$$

3.2 Neighbor-weighted Collaborative Filtering (NwCF)

This method is an enhancement of our CCF implementation. The neighborhood of ten users is obtained in exactly the same way, using the Pearson correlation. However, we attempted to take into account the number of raters in the calculation of the ranking of the articles. We do it due to a large amount of the articles have been rated by only one or at most two users. In this way, we push up in the recommendation list those articles rated by a larger number of neighbors. The new predicted rating is given by

$$pred'(u,i) = \log_{10}(1 + nbr(i)) \cdot pred(u,i) \quad (3)$$

3.3 BM25-based Similarity (BM25)

BM25, also known as Okapi BM25, is a non-binary probabilistic model used in information retrieval [7]. It calculates the relevance that the documents of one collection have given a query. As we try to take advantage of the set of tags of each user, we made two analogies, comparing the tags of the center user with a query, and the set of tags of each neighbor as a document. Based on this idea, we performed a similarity calculation based on the BM25 model and thus we obtained her neighborhood. Our proposed BM25-based similarity model is taken from the calculation of the Retrieval Status Value of a document (RSV_d) of a collection given a query [7]:

$$RSV_d = \sum_{t \in q} IDF \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b \times (L_d / L_{ave})) + tf_{td}} \cdot \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}} \quad (4)$$

In our model RSV_d represents the similarity score between the center user (the terms of the query q) and one neighbor (the terms of the document d). This similarity is calculated as a sum over every tag t posted by the center user. The neighbor d is represented as her set of tags with their respective frequencies. L_d is the document length, in our case is the sum of the frequencies of each tag of the neighbor d. L_{ave} is the average of the L_d of every neighbor. The term tf_{td} is the frequency of the tag t into the set of tags of the neighbor d. tf_{tq} represents the frequency of the tag t into the query, i.e., the set of tags of the center user. Finally, $k1$, $k3$ and b are parameters that we have been set in 1.2, 1.2 and 0.8

respectively, values slightly different from those suggested by default in [7].

After calculating the similarity between the center user and each neighbor, we choose the top N similar neighbors, and then we calculate the ranking of the recommended articles using the formula (3).

4. THE STUDY

To perform our study, we selected ten active CiteULike users which had posted at least 50 articles each. Four of the subjects are part of the Personalized Adaptive Web Systems (PAWS) lab of the School of Information Sciences at the University of Pittsburgh. Six additional subjects were selected randomly from a list of active CiteULike users.

For each subject we generated 4 sets with 10 ranked articles each one. The first three lists were generated using the methods CCF, NwCF and BM25, considering 10 neighbors for each center user. The fourth list was generated using BM25, yet considering 20 neighbors. To avoid pitfalls in the evaluation [8], for each subject we combined the 4 sets of recommendations into one set, we changed the order of the articles randomly and we ask them to evaluate each article relevancy (relevant, somewhat relevant, and not relevant), and novelty (novel, somewhat novel, and not novel) using a 3-point scale. For example, one article can be evaluated as relevant but not novel (because it was already known), and another article can be judged to be relevant and also novel, because the user just discovered and found it to be important to her interests.

Another issue considered to make the comparison more reliable was the amount of information about an article used to make relevance judgment. For each article, we provided a URL to its CiteULike record, which provides basic bibliographic information and frequently an abstract for each article. We requested each subject to evaluate the articles based on that information or, if the abstract was not available, looking for the abstract in the paper source, but do not look beyond the abstract.

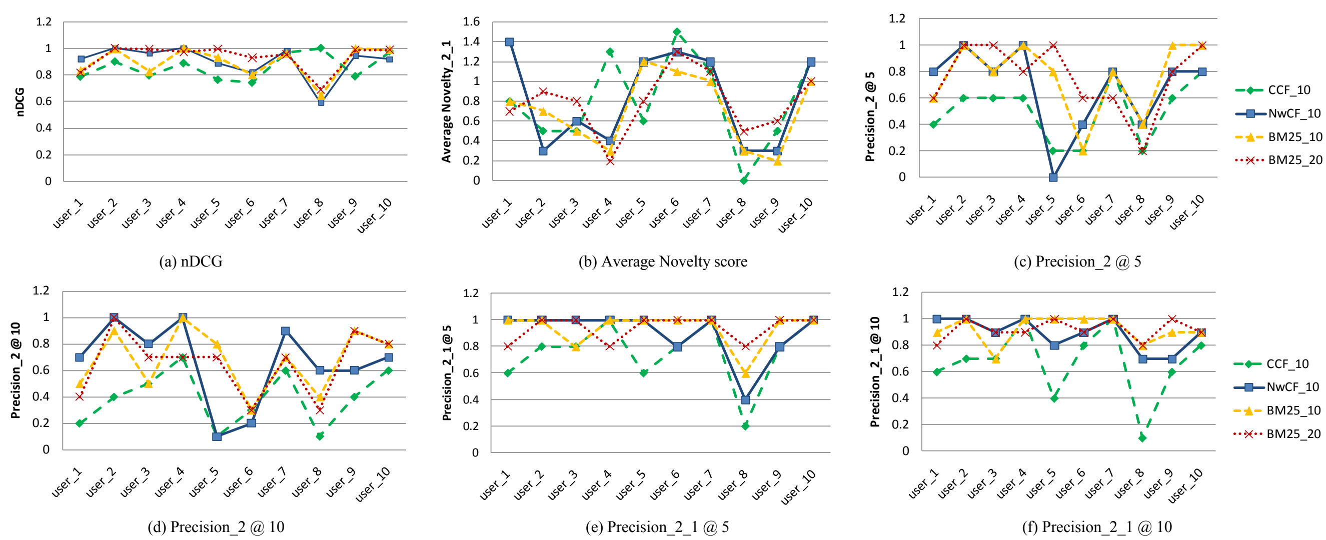
For each subject, we calculated normalized Discounted Cumulative Gain (nDCG) [7], Precision_2@5, Precision_2@10, Precision_2_1@5 and Precision_2_1@10 over the different initial four lists of recommendations. In Precision_2_1, we consider relevant those articles evaluated as *Relevant* and *Somewhat Relevant*. In Precision_2, we only consider relevant the articles evaluated as *Relevant*. Besides, we calculated the average novelty for each user on each method. To calculate the average novelty, we considered only items evaluated as relevant or somewhat relevant, disregarding novelty of not relevant items.

Figure 1 a) shows us smooth results on different subjects and not so different results on the values of nDCG between different algorithms. However, if we compare them further, we can see that CCF performed the worst and is not so clear which one, BM25_10, BM25_20 or NwCF is significantly the best. This result suggests us that the ranking order of the recommendations, in general, is very close to the optimal one, where the most relevant articles are at the top and the less ones at the bottom. In figure 1 b) is not possible to see any clear trend about which algorithm performs the best on novelty.

The results on Precision_2 and Precision_2_1 do not let us infer easily some ideas, but we can see some trends. In general, CCF has the worst results, suggesting that including the amount of raters in the ranking formula is an important factor to consider in the success of these recommendations.

5. RELATED WORK

A few pioneer projects explored different ways to integrate social links or social tags. In [3], the authors incorporate social tags and also the concept of *web of trust* for the issue of quality assessment into a collaborative recommendation approach. The study in [4] investigates the effect of incorporating tags to different CF algorithms, testing their algorithms on last.fm, a musical social tagging system, obtaining promising results. The approach presented in [5] compared a pure content-based with a tag-enhanced recommender, showing an improvement in predicted accuracy in the context of cultural heritage personalization.



(a) nDCG (b) Average Novelty score (c) Precision_2 @ 5

(d) Precision_2 @ 10 (e) Precision_2_1 @ 5 (f) Precision_2_1 @ 10

Figure 1: Metrics showing the results of each user on each method of the experiment (a) nDCG, (b) Average Novelty, (c) Precision_2 @ 5, (d) Precision_2 @ 10, (e) Precision_2_1 @ 5, (f) Precision_2_1 @ 10

The study presented in [6] describes the use of CiteULike for recommending articles to users. They compared three different collaborative filtering algorithms, two item-based and one user-based, and they found that the latter performed the best. They evaluated their algorithms using accuracy metrics as MAP, MMR and Precision@10, with low accuracy levels, in the range 0.1-0.3.

In [8] McNee et al. developed three algorithms to recommend articles to users, and they assessed them with a detailed survey on real users. In some algorithms, the subjects provided strong negative results, and the authors concluded that when evaluating a recommender system "*the evaluation must be done with real users, as current accuracy metrics cannot detect these problems*". Based on this study we decided to ask the subjects to evaluate the novelty in addition to the relevance. Five of our ten subjects commented at the end of the survey that they found very interesting articles in their recommendation list.

6. DISCUSSION

The results of our study, as well as the experience of other teams, show that one must approach the problem of recommendation in social tagging systems with open mind. Pragmatic implementation of traditional approaches may deliver relatively poor results in this new context. Our work shows that both traditional approaches to collaborative filtering behave sub-optimally in CiteULike. First, the use of Pearson correlation to form user neighborhood delivers poor results. While CiteULike with its 5-star-plus-one rating of bookmarked papers appears, at first sight, to be a good case for using Pearson formula, we found that in a bookmarking context this rating is not reliable. We started our CiteULike study using Pearson over five star-based rating, but were puzzled with low quality of recommendations in the pilot study. To address it, we moved from 6-point to 3-point rating. Since many users post articles without taking care of the ratings (by default it is 2 stars), and their evaluation criteria can vary, we decided to treat default 2-star rating as considerable interest (1 point), explicit change to one star as low interest (0 points), and explicit change to 3-5 stars as high interest (2 points). Afterwards, the results showed a significant improvement. This highlights the importance of the rating scale in recommender algorithms for social bookmarking systems, where the meaning of stars could be different, since tags, not stars are the primary product of bookmarking.

Yet, we believe that even the reduced-scale rating is not reliable enough to use Pearson correlation due to nature of bookmarking systems. While traditional recommender systems use a fair mixture of positive and negative ratings, a presence of a bookmark is mostly a positive sign. In this context, any additional star ratings as used in CiteULike represent different shades of positive and become less reliable. While some users may do their best by distinguishing "I want to read it" and "I really want to read it", a good fraction simply gives up and become *single-value raters*. In our case, 21% of the users on this study had rated all their articles with 2 stars (the default rating), and 34% have used the same rating (either with 1, 2, 3, 4, or 5 stars) over all their articles. In this context, Pearson correlation becomes too noisy. A small fix is to move from Pearson to some binary item-based similarity measure such as Jaccard. A more radical approach, which we explored, is a switch from an item-based to a tag-based approach to calculate similarity between users. In our case it paid off: BM25-based similarity performed better than CCF.

Our experience with NwCF demonstrates that the inclusion of the amount of raters in the ranking formula is an important contribution. Our data shows that both nDCG and precision metrics have better results for NwCF than for CCF. This result hints that the number of raters is a part of the "social knowledge", which can increase the quality of outcome that CCF ignores.

7. CONCLUSIONS

We explored four variations of user-based collaborative filtering algorithms in the context of a collaborative tagging system for scientific articles, CiteULike. We can summarize the results of our study in three observations. First, classic rating-based collaborative filtering algorithms implemented on social tagging systems must carefully consider the rating scale to avoid noise on the recommendation lists. Second, incorporating the amount of raters in the recommender algorithms can help to decrease the uncertainty produced by items with too few ratings. Third, a tag-based approach to obtain user neighborhood in social tagging systems can be a suitable alternative to Pearson correlation.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0840597.

9. REFERENCES

[1] Pazzani, M. and Billsus, D. 2007 Content-Based Recommendation Systems. The Adaptive Web. (May 2007), 325-341.

[2] Schafer, J., Frankowski, D., Herlocker, J. and Sen, S. 2007 Collaborative Filtering Recommender Systems. The Adaptive Web. (May 2007), 291-324.

[3] Massa, P. and Avesani, P. 2004 Trust-Aware Collaborative Filtering for Recommender Systems. In Proceedings of OTM Confederated International Conferences, CoopIS, DOA, and ODBASE (Agia Napa, Cyprus, Oct. 25-29, 2004). 492-508.

[4] Tso-Sutter, K. H., Marinho, L. B., and Schmidt-Thieme, L. 2008. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In Proceedings of ACM Symposium on Applied Computing (Fortaleza, Brazil, March 16 - 20, 2008). SAC '08. ACM, New York, NY, 1995-1999.

[5] de Gemmis, M., Lops, P., Semeraro, G., and Basile, P. 2008. Integrating tags in a semantic content-based recommender. In Proceedings of the 2008 ACM Conference on Recommender Systems (Lausanne, Switzerland, October 23 - 25, 2008). RecSys '08. ACM, New York, NY, 163-170.

[6] Bogers, T. and van den Bosch, A. 2008. Recommending scientific articles using citeulike. In Proceedings of the 2008 ACM Conference on Recommender Systems (Lausanne, Switzerland, October 23 - 25, 2008). RecSys '08. ACM, New York, NY, 287-290.

[7] Manning, C., Raghavan, P. and Schutze, H. 2008 Introduction to Information Retrieval. Cambridge University Press.

[8] McNee, S. M., Kapoor, N., and Konstan, J. A. 2006. Don't look stupid: avoiding pitfalls when recommending research papers. In Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work (Banff, Alberta, Canada, November 04 - 08, 2006). CSCW '06. ACM, New York, NY, 171-180.