Improving a Hybrid Literary Book Recommendation System through Author Ranking

Paula Cristina Vaz
INESC-ID/IST-Portugal
paula.vaz@inesc-id.pt

David Martins de Matos
INESC-ID/IST-Portugal
david.matos@inesc-id.pt

Bruno Martins
IST/INESC-ID-Portugal
bruno.martins@ist.utl.pt

Pavel Calado
IST/INESC-ID-Portugal
pavel.calado@ist.utl.pt

ABSTRACT

Literary reading is an important activity for individuals and can be a long term commitment, making book choice an important task for book lovers and public library users. In this paper, we present a hybrid recommendation system to help readers decide which book to read next. We study book and author recommendations in a hybrid recommendation setting and test our algorithm on the LitRec data set. Our hybrid method combines two item-based collaborative filtering algorithms to predict books and authors that the user will like. Author predictions are expanded into a booklist that is subsequently aggregated with the former book predictions. Finally, the resulting booklist is used to yield the top-n book recommendations. By means of various experiments, we demonstrate that author recommendation can improve overall book recommendation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information filtering; H.3.7 [**Digital Libraries**]: System issues

General Terms

Algorithms, Experimentation

Keywords

Hybrid Recommender, Book Recommendation, Author Recommendation

1. INTRODUCTION

Literary reading is an important activity for individuals and public libraries enable users to participate in this activity free of charge by allowing them to borrow books for short periods. As a consequence, good recommendations in a public library can improve a reader's usability of the library.

Libraries have limited shelf space, but still have enough books to make book selection difficult and time consuming. However, the number of books and users is not enough to successfully use the traditional collaborative techniques that rely on large amounts of data to detect patterns.

Copyright is held by the author/owner(s).
JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

In this paper, we aim to (i) assess whether item-based collaborative filtering (CF) can be used to make good recommendations in a public library, and (ii) assess whether selecting books by author preferences can improve recommendations. We also propose HyBook: a weighted hybrid approach to recommend literary books. Our approach combines two item-based CF algorithms to improve recommendations, where one recommends books and the other recommends authors. Author recommendation is used to improve the book top-n recommendations through a fusion approach.

2. RELATED WORK

Literature on recommendation systems typically classifies systems such as collaborative-, content-, demographic-, and/or knowledge-based. Moreover, hybrid systems combine two or more algorithms to improve recommendations, overcoming limitations of the individual approaches [2].

There are several book recommendation sites that can be found on the Internet. We highlight: *gnooks*[1], a CF book and author recommender through a literature map; and *Similar authors*[2] a CF author recommender. In [4], the author investigates the effectiveness of author rankings in a library catalog to improve book retrieval. However, to the best of our knowledge, this is the first study attempting to improve book recommenders through author recommendation.

3. PROPOSED METHOD

HyBook algorithm uses two matrices: the $book \times user$ matrix, where $cell_{i,j}$ contains the rating that $user_j$ gave to $book_i$; and the $author \times user$ matrix, where $cell_{i,j}$ contains the average of ratings that $user_j$ gave to books of $author_i$. These matrices are used to calculate both $book \times book$ and $author \times author$ matrices where the $cell_{i,j}$ contains similarity between items i and j. Similarity is calculated using item co-occurrences.

HyBook uses the active user preference vector and the $book \times book / author \times author$ matrix to yield the book/author rank vector. Position i of the rank vector contains the rank predicted for $book_i / author_i$ according to the active user preferences. The author rank vector is then expanded into a book rank vector by assigning to each book its author predicted rank. Rank vectors are calculated by aggregating

[1]http://www.gnooks.com
[2]http://www.similarauthors.com

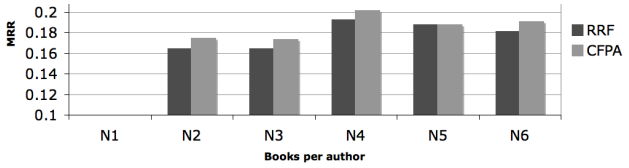

Figure 1: MRR per author neighborhood size.

the columns of the similarity matrices corresponding to the items preferred by the active user. We tested our CF predictions with two aggregation functions: the reciprocal rank fusion (RRF) [3]; and the traditional collaborative filtering preference aggregation (CFPA) [1].

Finally, book rank vectors are aggregated and sorted, generating the top-n book list for the active user. Rank vectors are aggregated using a weighted arithmetic mean (WAM) as shown in Equation 1.

$$WAM(u) = \frac{\alpha\ CF_{author}(u) + (1-\alpha)\ CF_{book}(u)}{2} \qquad (1)$$

4. EXPERIMENTS

This section describes the LitRec data set and discusses experiments with HyBook. To evaluate the quality of top-n lists generated by HyBook, we used the mean reciprocal rank (MRR). In Equation 2, p_i is the position, at the top$_i$, of the first relevant document.

$$MRR = \frac{\sum_{i=1}^{number\ of\ tops} \frac{1}{p_i}}{number\ of\ tops} \qquad (2)$$

4.1 LitRec data set

LitRec is a data set built by the authors for recommendation purposes. It combines documents from Project Gutenberg[3] with ratings from *Goodreads*[4] and contains 38,591 ratings from 1,927 users and 3,710 documents. The data set also contains book authors (1,627 different authors) and the review date among others. The review date was used to sort and divide ratings in a train-test set of 90%-10%.

4.2 CF Aggregating function

We used RRF and CFPA to aggregate book and author predictions. Results showed that the best predictions were achieved using CFPA for books (MRR = 0.28) and RRF for authors (MRR = 0.22).

4.3 Author neighborhood

The author neighborhood is the number of books written by the author and used in predictions. Experiments have shown that if the number of books per author is not restricted, the expanded book rank vector will be saturated by authors with more books, leading to less accurate predictions. As shown in figure 1, predictions improve when the book neighborhood varies from 1 to 4 and decreases after 4. Based on this experiment, we set author neighborhood to 4 books.

[3]http://www.gutenberg.org
[4]http://www.goodreads.com

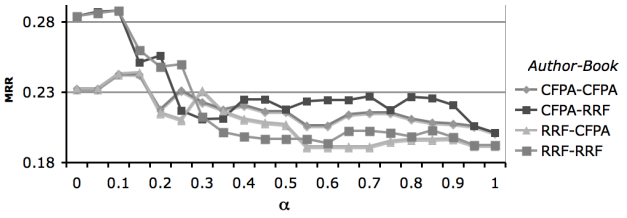

Figure 2: HyBook MRR for all combinations of author-book rank aggregation.

4.4 Aggregating book ranks

HyBook aggregates rank vectors yield by book and author predictions using WAM (Equation 1). We also experimented with RRF, but results were not promising. We varied the α parameter between 0 and 1 in order to assess the importance of the author in final recommendations. As shown in figure 2, HyBook yields the best predictions when ranks have the combination of 10% author and 90% book.

The graphic also outlines the evolution of all combinations of score aggregations. As expected from results obtained in the previous experiments, when author predictions use CFPA and books predictions use RRF, overall results are better. However, at the 10%-90% author-book combination the RRF-RRF combination can achieve the same results.

5. CONCLUSIONS & FUTURE WORK

In this paper we describe a HyBook: a hybrid book recommendation algorithm. HyBook combines author and book predictions to yield the top-n book recommendations.

Regarding the first of the initial goals presented in Section 1, experiments led us to conclude that the common CF approaches yield poor predictions due to limitations of this version of LitRec. Regarding the second goal experiments in LitRec have shown that overall predictions can be improved using author prediction. However, a maximum number of books per author must be established, otherwise authors with more books will appear to have an advantage over less productive authors.

We also observed that for the LitRec data set, the contribution of choosing books by author must be smaller than that of choosing books by their popularity.

This paper describes exploratory work in LitRec data set that opens a path for further research. We intend to continue exploring LitRec. We will try to assess if book choice is related to content, user location, and the month in which the book was read. Finally, the use of feature augmentation and dimensionality reduction techniques will also be considered.

6. REFERENCES

[1] G. Beliakov, T. Calvo, and S. James. Aggregation of preferences in recommender systems. In *Recommender Systems Handbook*, pages 705–734. Springer, 2011.

[2] D. Burke. Hybrid web recommender systems. In *The Adaptive Web*, pages 377–408, 2007.

[3] G. Cormack, C. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, New York, NY, USA, 2009.

[4] J. Kamps. The impact of author ranking in a library catalogue. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, BooksOnline '11, pages 35–40, New York, NY, USA, 2011. ACM.