Mining author-tag multilayer graph for social book search

Mohamed Ettaleb, Patrice Bellot, Chiraz Latiri

Mining Author-Tag Multilayer Graph for Social Book Search

Mohamed ETTALEB * and Patrice BELLOT
Aix Marseille University, University of Toulonn, CNRS, LIS,
13397, Marseille, France
E-mail: {mohamed.ettaleb,patrice.bellot}@univ-amu.fr

Chiraz LATIRI
University of Tunis El Manar,Faculty of Sciences of Tunis, LIPAH research Laboratory,
Tunis ,Tunisia
E-mail: chiraz.latiri3@gmail.com

The emergence of social media allows users to get opinions, suggestions, or recommendations from other users about complex information needs. Tasks, as the CLEF Social Book Search 2016 Suggestion Track, propose to pursue this issue. The originality is to deal with verbose queries of book recommendation in order to support users in searching for books in catalogues of professional metadata and complementary social media(*i.e* tags, authors, similar products). In this context, a new technique for *community-of-books* discovery based on frequent social information(*i.e* tags, authors) of similar books are proposed for book recommendation. Our method allows detecting frequent sub-graphs of similar books and using them to enrich the results returned by a traditional information retrieval system. This approach is tested on a collection containing Amazon/LibraryThing book descriptions and a set of queries, extracted from the LibraryThing discussion forums.

Keywords: graph mining; information retrieval; social book search; recommendation system

1. Introduction

With the fast growth of e-commerce and social network services, social information is becoming increasingly important in describing the properties, content and attributes of items or products [1]. In addition, social information is included in a large part of corpus, and is therefore largely discussed in the research areas of IR and recommendation systems. It seems that social information has a large amount of user-generated contents that can have a critical influence on recommendation applications. Although the question has been cited by some researchers, there has been limited work done to respond to the question of how to systematically survey social information to promote the recommendation. However, social information has numerical superiority over business data in selecting effective terms for linking similar products (i.e., the use of tags to link similar books). Typically, they are generated to reflect

preferences, creators' opinions about content, or relationships (e.g., products purchased together) with other products in the collection, or certain similar products. Consequently, it is intended to build a generic framework for diverse and heterogeneous types of social information in order to enhance conventional IR or recommendation models.

The main purpose of this paper is to propose a new framework to improve the performance of the recommendation system with social information (i.e. reviews, tags, ratings, etc). Within this framework, existing forms of social information that might appear can be easily categorized and sufficiently used to construct a multi-layer graph of books to discover frequent sub-graphs of similar books.

Our approach presents a collection of books as hierarchical multilayer graph for books and utilizes an algorithm of graph mining to find the frequent sub-graphs, which reflect frequent similar books(books often presented as similar by the readers). A multilayer graph is defined on the same set of vertices and edges, but each layer has a different set of labels. In order to validate our proposal, we evaluated it in the context of the Social Book Search (SBS) CLEF Labs task[2].

2. Related work

We are inspired in our paper to find communities and link sub-graphs. We briefly review the two areas of work below; but given the large number of literature and the limited space available, we can not hope to be comprehensive.

Connectivity sub-graphs: [3] and[4] solved the problem of finding a sub-graph connecting a set of query nodes in a graph. The purpose was to develop proximity measurements between the nodes of the graph that depended on the overall structure of the graph. Then, the task was to extract sub-graphs close to the query nodes, according to the developed similarity measure, and to link the query nodes. In subsequent work,[5] refined proximity measurements using the concept of *Cycle-Free Effective Conductance*(CFEC) and proposed a brunch and bound algorithm in order to find a sub-graph that would optimize the CFEC measurement.

Community detection: More recently, a related but different problem called community search has attracted considerable interest. It has been motivated by the need to give more meaningful and personalized responses to the user[6]. For a given set of query nodes, the community search looks to find the communities containing the query nodes. The entities modeled by the network nodes often have properties that are important to make sense to communities. The main difference of our approach is that we are more interested in extracting the cmost relevant community which has the highest scores (see Section 4) for book recommendation. We select for each query the $TopN$ books retrieved

by an initial information retrieval system. After that, the $TopN$ books are considered as contextually similar books and are represented by a graph G_q.

3. Recommendation based on the frequent sub-graph mining

3.1. *Modeling links between books using a graph*

We have conducted an analysis of books to find a new way to link them. For the SBS task collection, we exploit a specific type of similarity based on several factors. This similarity is pre-calculated by Amazon between the different items using a collaborative item-item filtering method.
The result of the selection corresponds to the "Similar Product" field of each book in the SBS task collection. To model the collection in a graph, we extract the links from "Similar Products". In our graph, each node corresponds to a book (an Amazon book description) and has all the following properties:

(1) ID: representing the ISBN of the book
(2) $Author$: the author of a book
(3) $Tags$: set of labels from a book

The relationships in our graph are not oriented and correspond to Amazon's similarities. Given the two nodes $B_1, B_2 \in G$, if B_2 is suggested as a similar product of B_1 by Amazon or the other way around. The graph that we built has a total of 573,488 nodes (the rest is not included because they do not have "Similar Products" or "Tags" or "Author") and 2,140,947 relationships, with most documents having no more than 6 "Similar Products".

3.2. *Towards query based recommendation system using frequent sub-graphs*

The general architecture of our two-level book recommendation system is illustrated in the figure 3.2. In this system, SRI finds all initial relevant books for user's query. Then, to build the book graph, each node on a book-graph represents a book. An undirected edge represents the link of similarity between two books, that is the "Similar Products" links between them. Finally, the nodes in the book-graph are labeled with their own attributes (i.e. Tags or Authors in our case). Next, a sub-graph mining algorithm is applied to extract the frequent subsets of similar books. A *Scoring Ranking* module combines scores of books resulting from SRI and *Graph mining* modules and reclassifies them. In this section, the collection of books is noted by B. In B, each book B_i has a unique ID. The set of queries are denoted by T. Collection $B_i \subset C$ refers to the books returned by the initial information retrieval system. B_{t_i} indicates the books retrieved for request $T_i \in T$. $TopN_{B_{t_i}}$ represents the highest ranked books for a query T_i. The latter is then transformed as a sub-graph $G_{t_i} = (V_i, E_i, L_{vi}, L_{ei})$ of size N nodes. The set of books present in the labeled

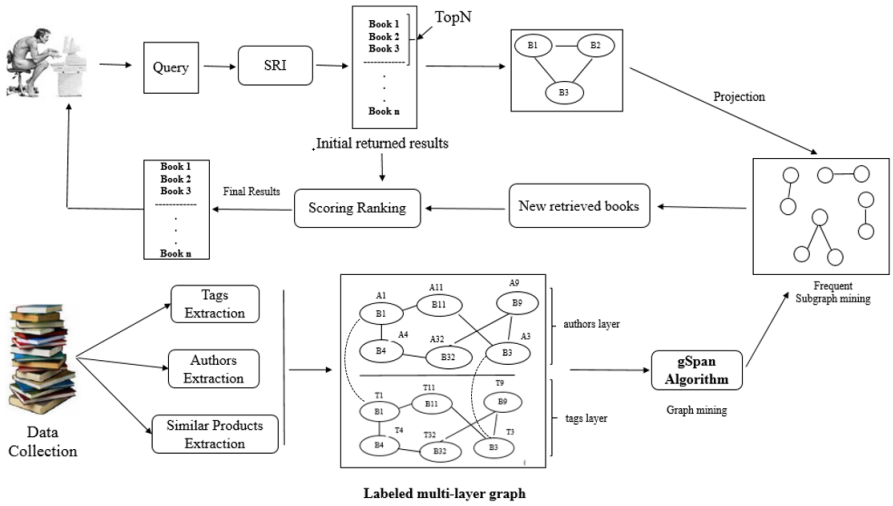

Fig. 1. Architecture of book retrieval approach based on frequent sub-graph mining

graph is noted by G. This labeled graph may be represented by quadruplet $G = (V, E, L_v, L_e)$, where:

- V: is a set of vertices, and each vertex represents the ID of the book.
- $E \subseteq V \times V$: is a set of edges, given the nodes $\{B_1, B_2\} \in E$ if and only if B_1 and B_2 are suggested as "Similar Products"
- L_v: is a set of vertex labels, and each vertex is tagged by book tags
- L_e: is a set of edge labels.

Finally, the collection of frequent sub-graphs returned by *gSpan* is denoted by G_{gSpan}. *gSpan* uses frequent sub-graphs as a representation of common similarity among graph G. Two books in G, containing some common frequent sub-graphs, must have common labels. Our process takes as input B_i the list of books returned for each request by an information retrieval system. We select for each request the $TopN$ ranked books, $TopN$ is fixed at 3. B_f represents the list of recommendations for each query. The process processes query by query and extracts the list of all the frequent sub-graphs g' in which g and g' are isomorphic. A concatenation is performed between the list of $TopN$ books and B_{graph} in a new list B_f. Finally, B_f is reclassified using the reclassification scheme.

4. Combining Search Systems

Using different recovery systems can extract different sets of documents. Combining the results of many research systems, as opposed to using a single recovery technique, can improve the efficiency of recovery, as Belkin showed in [7], where he combined the results of probabilistic and vector space models. In our work, we combine the results of the initial retrieval system with the results returned by the sub-graph mining method. We use the maximum and minimum scores according to Lee's formula [8] to calculate the new score:

$$Score(B_i) = P(B_i) + normalizedScore(B_i) \tag{1}$$

$$normalizedScore(B_i) = \frac{oldScore(B_i) - minScore}{maxScore - minScore} \tag{2}$$

Where:
$P(B_i)$ is the score of the book B_i attributed by the the initial retrieval system. $oldScore(B_i)$ is the *support* of the sub-graph returned by *gsPan* wich the book B_i appears. $maxSxore$ and $minScore$ are respectively the maximum and minimum *support* from all extracted sub-graphs .

5. Experimental results and discussion

5.1. *Test collection description*

First, to extract the documents returned by an information retrieval system, we used the Terrier information retrieval platform that is being developed at the University of Glasgow [9]. It implements the various modules involved in the traditional IR process and also provides a framework for evaluating research results for different applications. The BM25 model was used to retrieve the relevant documents with the default parameter values ($b = 0; k3 = 1000; k1 = 2$).
In Section 3, we used the *gSpan* algorithm to extract all the common patterns. As a parameter, *gSpan* takes as minimal support $minsupp = 3$.
We conducted three different runs, namely:

(1) **Recommendation based on the Authors graph:** We used the layer *Author* to extract the frequent subsets of similar books.
(2) **Recommendation based on the graph of the Tags:** the layer *Tags* was used to recover new frequent subsets.
(3) Run-**Recommendation based on the multi-layer graph (Graph-Tags+Authors):** the two layers(*Author* and *Tags*) are combined.

5.2. *Results and Significance evaluation*

We first compare the search results (baseline) that are presented in [10], in which authors have used a combination of reduction and expansion approaches.

After several experiments on the two years topics, Table 1 shows the results obtained. From this table, we notice that on the 2-year SBS datasets with book collections, our approach outperforms than the baseline model to a large extent. In 2014 collection, the relative improvements of overall performances are 5.92%, 11.59% and 13.63% respectively with the metric $NDCG@10$, which is the official evaluation metric. The scores of other metric, MAP, also present similar results.

About the 2016 collection, the best performance is obtained by $Graph-Tags+Authors$ (when we combined the graphs with the two layers $Authors$ and $Tags$ together) strategy with the greatest improvements of the score of 13.63% and 14.34% in the score of NDCG@10. Tracing it to its cause, in many cases, the selected frequent sub-graphs of our method have a higher quality.

Hence, we compare our results with the baseline model on the SBS datasets according to the t-test experiments, we notice that when applying Multi-Layers Graph($Tags$ and $Authors$), our system statistically significantly outperforms the baseline model ($p - value = 0.0016$ in 2014 collections and $p - value = 0.0011$ in 2016). The other two pairs of comparison shows similar results ans the exact result value are shown in Table 1. With the statistical significant testings, our proposed frequent sub-graphs mining provides a unified framework for dealing with a variety of social information.

To deepen the analysis of efficiency, we present an analysis of the gains of our approach. Table1 shows the percentages of queries Q^+ for which the sub-graph mining techniques perform better in term of $NDCG@10$. As described in Table 1, the average percentage for the set of queries Q^+ is about 28.7%. these results confirm the effectiveness of using sub-graph mining for the recommendation.

Table 1. Results of SBS 2014 and 2016 with different strategies, The p-value obtained using (two-tailed T-Test), and Percentage of queries $Q+$ in terms of $NDCG@10$

Strategy	NDCG@10	MAP	Improved	Q+	P-values
SBS 2014 COLLECTIONS					
Baseline	0.1518	0.1194			
Graph-Authors	0.1608	0.0.1251	5.92	20.7%	0.059%
Graph-Tags	0.1694	0.1283	11.59	24.2%	0.00041%
Graph-Tags+Authors	**0.1725**	**0.1309**	**13.63%**	**27.1%**	0.0016
SBS 2016 COLLECTIONS					
Baseline	0.1688	0.1055			
Graph-Authors	0.1776	0.0.1113	5.21	19.1%	0.052%
Graph-Tags	0.1883	0.1153	11.55%	25.3%	0.0003
Graph-Tags+Authors	**0.1930**	**0.1184**	**14.34%**	**28.7%**	0.0011

We also perform the statistical comparison in terms of $NDCG@10$ and MAP with the participants. The chosen effective baselines are the run from [10] that used query expansion approaches based on association rules between sets of terms, the runs from Hafsi [11] when authors exploited some user generated content such as reviews and ratings to recommend some books, Benkoussas [12] which are proposed a method that combines the outputs of probabilistic model (InL2) and Language Model (SDM), Chaa [13] that are presented a new technique to automatically generate a stopword list in order to reduce verbose queries. According to the $NDCG@10$ and MAP experiments, our system is outperforms Benkoussas system ($NDCG@10$ =0.128, MAP=0.101) The other three pairs of comparison shows similar results and the exact result values are shown in Table 2. These comparative experiments also prove that the proposed models are more effective on both datasets compared with the participants.

Table 2. Comparison results on SBS 2014

Run	**NDCG@10**	**MAP**
Our proposal	**0.1725**	**0.1309**
From [12]	0.128	0.101
From [13]	0.1565	0.1100
From [11]	0.1424	0.1070
From [10]	0.1518	0.1194

6. Conclusion

This paper proposes and evaluates new approach based on frequent sub-graph mining of similar books using social information(*i.e* Tags, Authors) in the context of book recommendation. This method presents books as hierarchical graphs and utilizes an algorithm of graph mining to find the frequent sub-graphs. It allows to find a subsets of relevant book in the graph based on the sub-graphs mining technique to enrich the books list returned by a traditional recovery model. The results obtained confirmed that the use of the graph mining for the recommendation is fruitful.

References

1. B. Zhang, X. Yin, X. Cui, J. Qu, B. Geng, F. Zhou, L. Song and H. Hao, Social book search reranking with generalized content-based filtering, in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, 2014.
2. M. Koolen, T. Bogers and J. Kamps, Overview of the SBS 2015 suggestion track, in *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015.

3. C. Faloutsos, K. S. McCurley and A. Tomkins, Fast discovery of connection subgraphs, in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, 2004.
4. H. Tong and C. Faloutsos, Center-piece subgraphs: problem definition and fast solutions, in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, 2006.
5. Y. Koren, S. C. North and C. Volinsky, Measuring and extracting proximity graphs in networks, *TKDD* **1**, p. 12 (2007).
6. X. Huang, H. Cheng, L. Qin, W. Tian and J. X. Yu, Querying k-truss community in large and dynamic graphs, in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '142014.
7. N. J. Belkin, P. B. Kantor, E. A. Fox and J. A. Shaw, Combining the evidence of multiple query representations for information retrieval, *Inf. Process. Manage.* **31**, 431 (1995).
8. J. H. Lee, Combining multiple evidence from different properties of weighting schemes, in *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*, 1995.
9. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald and C. Lioma, Terrier: A high performance and scalable information retrieval platform, *Proceedings of the OSIR Workshop* , 18 (2006).
10. M. Ettaleb, C. Latiri and P. Bellot, A combination of reduction and expansion approaches to deal with long natural language queries, in *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference KES-2018.*, 2018.
11. M. Hafsi, M. Gery and M. Beigbeder, Lahc at INEX 2014: Social book search track, in *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, 2014.
12. C. Benkoussas, H. Hamdan, S. Albitar, A. Ollagnier and P. Bellot, Collaborative filtering for book recommandation, in *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, 2014.
13. M. Chaa, O. Nouali and P. Bellot, New technique to deal with verbose queries in social book search, in *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, 2017.