PubRec: Recommending Publications Based On Publicly Available Meta-Data

Anas Alzoghbi, Victor Anthony Arrascue Ayala,
Peter M. Fischer, and Georg Lausen

Department of Computer Science, University of Freiburg
Georges-Köhler-Allee 051, 79110 Freiburg, Germany
{alzoghba,arrascue,peter.fischer,lausen}@informatik.uni-freiburg.de

Abstract. In recent years we can observe a steady growth of scientific publications in increasingly diverse scientific fields. Current digital libraries and retrieval systems make searching these publications easy, but determining which of these are relevant for a specific person remains a challenge. This becomes even harder if we constrain ourselves to publicly available meta-data, as complete information (in particular the fulltext) is rarely accessible due to licensing issues. In this paper we propose to model researcher profile as a multivariate linear regression problem leveraging meta-data like abstracts and titles in order to achieve effective publication recommendation. We also evaluate the proposed approach and show its effectiveness compared with competing approaches.

Keywords: Recommender System, Scientific Paper Recommendation, Content-based Filtering, Multivariate Linear Regression, User Modelling

1 Introduction

Modern research is remarkably boosted by contemporary research-supporting tools. Thanks to digital libraries, researchers support their work by accessing a large part of the complete human knowledge with little effort. However, the sheer amount of rapidly published scientific publications overwhelms researchers with a large number of potentially relevant pieces of information. Recommender systems have been introduced as an effective tool in pointing researchers to important publications [5, 9, 10]. An approach that gained a lot of interest [2] extracts the interests of a user from the text of his/her publication list. In order to do so in an effective manner, full access to the textual content of research papers is needed. Yet, digital libraries typically provide only meta-data for publications including the publication date, title, keywords list and abstract. Although the availability of such information facilitates the problem, the usefulness of such a limited amount of information for paper recommendation is still unclear.

 In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org

In this work we explore an approach to effectively perform paper recommendation utilizing such limited information. We present an adaptive factor to measure the interest extent of the active researcher in each of her/his previous publications; we apply a learning algorithm to fit a user model which in turn can be used to calculate the possible interest in a potential paper. Our contributions can be summarized as follows:

- An effective approach for modeling researchers interest that does not require access to the fulltext of the publication, but only freely available meta-data.
- An adaptive *anti-aging* factor that defines, for each researcher and publication, a personalized interest extent, so that older contributions have less impact.
- Preliminary results of comparing our approach against two state of the art recommendation techniques that considers full textual content.

The rest of this paper is organized as follows. In Section 2 we review work related to our approach. Section 3 presents the problem definition and outlines the presented approach. Section 4 demonstrates the profile building model employing the *anti-aging* factor. In Section 5 we explain the conducted experiments and discuss the results. Finally we conclude the paper in Section 6

2 Related Work

Research paper recommendation has been a hot topic for more than a decade. Several works addressed this problem proposing ideas from different recommendation directions [2]. Publication title and abstract were employed in [10] to build a user model using collaborative topic regression combining ideas from Collaborative Filtering and content analysis, but results were of varying quality. Nascimento et al. [5] use titles and abstracts as well. Users provide a *representative* paper that fits their interests, out of which keywords are extracted from the title and abstract. These keywords are then used to retrieve similar papers from digital libraries. We believe this is a limited approach as keywords from one publication are not enough to capture user interests. Sugiyama and Kan in [8, 9] employ a simplified variation of the Rocchio algorithm [7] to build a user profile utilizing all terms which appear in the fulltext of the user's authored publications, while they also incorporate terms from the citing and referenced papers. However, this approach suffers from the poor quality of the terms used and from the dependency on tools to extract text from pdf files which have well-known limitations. Above all, the authors assumed the availability of the full text of the publications which is rarely the case. In this work we optimize the use of the publicly available meta-data rather than relying on the full text of the publication. Moreover, we build a researcher interest model that can depict different affinity models of researchers.

3 PubRec Model

We propose a content-based research publications recommender (PubRec) that models both the active user (the researcher) and the candidate publications in

terms of domain-related keywords. This section introduces the basic concepts of PubRec along with the formal problem definition.

3.1 Research Publication Profile

Digital libraries like ACM, IEEE, Springer, etc. publish meta-data about research publications publicly. Out of this meta-data, we are interested in title, abstract, keyword list and publication year. The first three can be effectively exploited to build a profile for each publication p as a keyword vector, which represents p in terms of domain-related keywords: $\overrightarrow{V_p} = \langle w_{p,k_1}, w_{p,k_2}, ..., w_{pk_n} \rangle$, where k_i is a domain-related keyword from the set of all keywords K, and w_{p,k_i} is the weight of k_i in p with range of $[0, 1]$.
All keywords from the keyword list are added to $\overrightarrow{V_p}$ with the maximum weight value of 1 by virtue of their source. As they are assigned to publications explicitly by the authors, we consider them the most precise domain description for the underlying publication. This list, however, contains usually up to 10 keywords, which is a small number for modeling a publication, thus, we aim to extend this list. Titles and abstracts hold a great essence of the ideas presented in publications. Therefore, we treat them as the second source of keywords and for each publication we apply keyword extraction from the concatenation of its title and abstract with weights correspond to the TF-IDF weighting scheme.

3.2 Researcher Profile

Given a researcher r with a set of her/his publications, we construct a researcher profile $\overrightarrow{V_r} = \langle s_{r,k_1}, s_{r,k_2}, ..., s_{r,k_n} \rangle$ such that $k_i \in K$ is a domain-related keyword, and s_{r,k_i} is the importance of k_i to r. Our proposed profile construction method ensures that r's *Interest Extent* (IE) in a publication p is achieved by computing the dot product between the researcher's vector and the publication's vector:

$$IE(\overrightarrow{V_r}, \overrightarrow{V_p}) = \overrightarrow{V_r} \cdot \overrightarrow{V_p} \tag{1}$$

3.3 Problem Definition

Our problem can be formally defined as:
Given a researcher r along with the corresponding set of publications P_r and a candidate set of publications P_{cand}, find k publications from P_{cand} with the maximum IE. The presented approach can be summarized in the following steps:

- First, we build the researcher profile $\overrightarrow{V_r}$ using previous publications by modeling the problem as a multivariate linear regression problem (Section 4)
- Each candidate publication $p \in P_{cand}$ is modeled as a keyword-vector $\overrightarrow{V_p}$
- We use Formula 1 to calculate $IE(\overrightarrow{V_r}, \overrightarrow{V_p})$ for candidate publication $p \in P_{cand}$
- Candidate publications are ordered by their Interest Extents and the top k are recommended to r.

4 Modeling Researcher Interest

We utilize researchers' publications to draw conclusions about their interests. A key aspect of PubRec consists in considering the different interest researchers have in their publications. After all, this interest might vary from paper to paper depending on several factors. Moreover, the importance of these factors vary among researchers. Thus, we believe that the publication age is an important factor in this regard since a five years old publication, for example, might not reflect the author's current interest as much a publication of the current year. Based on that, we introduce a scoring function for estimating the affinity of a researcher r towards one of her publications $p \in P_r$ by engaging the publication's age, which is expressed by the number of years elapsed after the publication's date and represented by σ in the following function:

$$IE_{r,p} = e^{\frac{-(\sigma)^2}{\lambda}}. \tag{2}$$

Here, λ is the researcher-specific *anti-aging factor*. As depicted in Figure 1, the curve of IE is plotted for three different values of λ: 4, 20 and 50. There we can see how λ regulates the steepness of this curve. As the values of λ increase, the curve becomes less steep and results in higher IE values for older publications. For example consider researcher r', the Interest Extent of r' for p', a 3 years old publication, can be modeled in three different ways upon three different values of λ: $IE_{r',p'} = 0.1$ for $\lambda = 4$, $IE_{r',p'} = 0.63$ for $\lambda = 20$ and $IE_{r',p'} = 0.83$ for $\lambda = 50$. This behavior helps in modeling different types of researchers based on their affinity model. Such that, researchers who tend to stick to the same research topics longer time are modeled using larger λ values compared to other researchers who tend to change their topics of interest more rapidly. Choosing the best λ for each researcher is done empirically in this work, but further investigations about the correlation between researcher characteristics and the optimal λ value are left for future work.

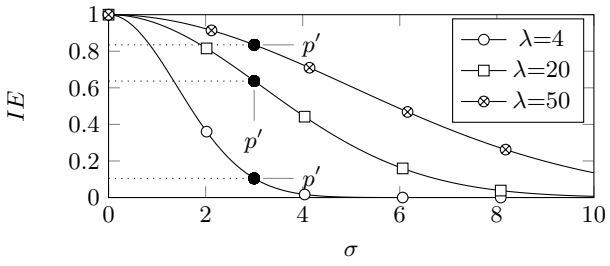

Fig. 1: *anti-aging factor* (*lambda*) impact on Interest Extent IE of researcher r'

4.1 Learning Researcher Profile

The second contribution in this work is to model the problem of measuring the importance of domain related keywords for a researcher r as a multivariate linear regression problem as follows: Given the set of r's publications P_r, for

each publication $p_i \in P_r$ we build the underlying publication profile as described in section 3.1: $\overrightarrow{V_{p_i}} = \langle w_{p_i,k_1}, w_{p_i,k_2}, ..., w_{p_i,k_n} \rangle$. Furthermore, the Interest Extent IE_{r,p_i} is calculated using Formula 2 as shown in Figure 2. Let the set of keywords' weights of the paper p_i: $w_{p_i,k_1}, w_{p_i,k_2}, ..., w_{p_i,k_n}$ be the set of predictors related to the response variable IE_{r,p_i}, then the multivariate linear regression model [6] for p_i is defined as: $IE_{r,p_i} = \overrightarrow{\theta} \cdot \overrightarrow{V_{p_i}} = \theta_0 + \theta_1 w_{k_1} + ... + \theta_n w_{k_n}$.

	k_1	k_2	k_3		k_n	IE
p_1 --→	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	...	$w_{1,n}$	IE_{r,p_1}
p_2 --→	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	...	$w_{2,n}$	IE_{r,p_1}
			...			
p_m --→	$w_{m,1}$	$w_{m,2}$	$w_{m,3}$	...	$w_{m,n}$	IE_{r,p_m}
θ --→	θ_1	θ_2	θ_3	...	θ_n	

Fig. 2: Publications keyword vectors and Interest Extents for one researcher

Where $\overrightarrow{\theta}$ is the regression coefficient vector and $\theta_0, \theta_1, \ldots, \theta_n$ are the regression coefficients. Each coefficient value $\theta_j, j \in 1, \ldots n$ defines the relation between the researcher r and the keyword k_j, or in other words the importance of k_j for r. Consequently, the user profile is modeled by means of $\overrightarrow{\theta}$. Meaning, that in order to find the user profile $\overrightarrow{V_r}$, we should solve the previously mentioned regression problem and find the vector $\overrightarrow{\theta}$. This problem is solved by minimizing the cost function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\overrightarrow{\theta} \cdot \overrightarrow{V_{p_i}} - IE_{r,p_i})^2$$

This is a well known optimization problem and there exist a couple of algorithms such as gradient descent or Normal equation to solve it [1]. We use an algorithm known for its efficiency, namely the L-BFGS algorithm [4].

5 Experiments

We conducted experiments to validate our approach and compared it against some state-of-the-art approaches. In the following we describe the used dataset along with the used evaluation metrics. Finally, we show and discuss the results.

5.1 Dataset

To evaluate the presented approach, we used the Scholarly publication Recommendation dataset[1]. It covers information about 50 anonymous researchers, enclosing their publication set, in addition to a set of publications of interest for each researcher. The interest lists are subsets of a larger collection of 100,531 publications called the candidate publications which is also provided.

[1] `https://www.comp.nus.edu.sg/~sugiyama/dataset2.html`

To the best of our knowledge, this is the only available dataset which provides the interest list for such a number of researchers. However, we had to resolve a major obstacle before we could use the dataset. That is, publications in the dataset are named by unique IDs without titles or author names, hence they cannot be identified and no meta-data was provided.
In order to make the dataset usable for our evaluation, we needed to identify the publications to be able to retrieve their meta-data. This was achieved by the following steps: (a) requesting and obtaining original pdf files from the dataset authors; (b) extracting publications' titles from the pdf files and using them to find publication identities within the DBLP[2] register; and finally (c) having the electronic edition pointer (ee) from DBLP publication's attributes, we retrieved needed information from corresponding publisher web site[3]. The result is a rich dataset that contains meta-data for 69,762 candidate publications, and more importantly the full publications and interest sets for 49 researchers. Lastly, for all publications in this dataset we applied the keyword extraction.
Keywords extraction and weighting. We use *Topia's Term Extractor*[4] because of its efficiency and usability. It is a tool that uses Parts-Of-Speech (POS) and statistical analysis to determine the terms and their strength in a given text. Yet we extended this tool in order to extract keywords with higher quality and make the best use out of the limited available resources. Our extensions to *Topia* are: (a) we apply post filtering on the resulting terms by choosing only those terms which appear in a white list of computer science terms; (b) the weights of extracted terms is calculated based on the normalized TF-IDF weighting scheme.

5.2 Evaluation metrics

We report the quality of our method with two important and widely adopted metrics for evaluating ranking algorithms in information retrieval. For the following metrics r is a researcher from the set of researchers R:
Mean Reciprocal Rank (MRR). MRR measures the method's quality by checking the first correct answer's position in the ranked result. For each researcher r, let p_r be the position of the first interesting publication from the recommended list, then MRR is calculated as $MRR = \frac{1}{|R|} \sum_{r \in R} \frac{1}{p_r}$
Normalized Discounted Cumulative Gain (nDCG)[3]. DCG@k indicates how good are the top k results of the ranked list. Typically in recommender systems DCG is measured for $k \in \{5, 10\}$ as users don't usually check recommended items beyond the 10^{th} position. The DCG for a researcher r is calculated as $DCG_r@k = \sum_{i=1}^{k} \frac{2^{rel(i)}-1}{log_{10}(1+i)}$ where $rel(i)$ indicates the relevance of the item at position i: $rel(i) = 1$ if the i^{th} item is relevant and $rel(i) = 0$ otherwise. nDCG is the normalized score which takes values between 0 and 1, it is calculated as: $nDCG@k = \frac{DCG@k}{IDCG@k}$, where $IDCG@k$ is the DCG@k score of the ideal ranking, in which the top k items are relevant. In our case we report on the average $nDCG@k$ over all researchers for $k \in 5, 10$

[2] `http://dblp.uni-trier.de/`
[3] We received the ACM publications' meta-data from ACM as XML.
[4] `http://pypi.python.org/pypi/topia.termextract`

5.3 Experimental results

Using the previously described dataset and evaluation metrics, we conducted quality evaluations for our method with the following setup: given a set of candidate publications, and a set of researchers with their publications set, the system should correctly predict the interesting publications for each researcher. The results are demonstrated in the first row of Table 1. It shows that PubRec manages to achieve a high MRR score of 0.717. Looking deeper into the details of this metric by examining results for individual researchers gives more insights: for 29 out of 49 researchers the first relevant publication appeared at the first position of the recommended list, and at the second position for 7 researchers.
We compared our approach with two state-of-the-art publication recommender systems [5, 9]. The work presented in [9] models each publication p using terms from p, from publications referenced by p and from publications that cite p. Additionally, the authors extended the set of citing publications by predicting potential citing publications. As our key contribution lies in utilizing only publicly available data, we implemented their core method[5] (Sogiyama) for modeling scientific publications considering only the terms which appear in the underlying publication. We compared PubRec against Sogiyama on two different setups: (a) Sogiyama using all terms appear in the full text of the publication[6]; (b) Sogiyama using our domain-related keywords. The results are shown in second and third row of the Table 1 respectively. In both setups PubRec outperforms Sogiyama in the three measured metrics. Furthermore, comparing our results with the results of Sogiyama as appeared in [9] where they assume the availability of the full text of the citing and referenced publications (5^{th} row in Table 1) in addition to the potentially citing publications (4^{th} row in Table 1), we find that our approach with such a limited available information is competitive and exhibits a reasonable trade-off between data-availability and recommendation quality. The last row in the table shows the scores of [5][7], where publications are modeled using N-grams extracted from titles and abstracts. Each user identifies a representative publication and the recommendation process turns into finding similar publications to the representative one by means of the cosine similarity.

6 Conclusion

We have proposed a novel approach on recommending scientific publications. By exploiting only publicly available meta-data from digital libraries the quality of the predictions is superior to state-of-the-art approaches, which require access to the full text of the paper. The focus is primarily on the user profiling, where a strategy to determine the trend of interests of a user in her own publications over time is integrated into a multivariate linear regression problem. The efficacy

[5] This method applies a light-weight variation of the Rocchio algorithm [7]
[6] The dataset provided by the authors of [9] contains all terms (not only domain-related terms) that appear in the full text of the publication.
[7] Values are taken from [9]

	MRR	nDCG@5	nDCG@10
PubRec	0.717	0.445	0.382
Sogiyama On their dataset	0.550	0.395	0.358
Sogiyama On PubRec dataset	0.577	0.345	0.285
Sugiyama and Kan [9]	0.793	0.579	0.577
Sugiyama and Kan [8]	0.751	0.525	0.479
Nascimento et al. [5]	0.438	0.336	0.308

Table 1: Recommendation accuracy comparison with other methods

of our approach is demonstrated by experiments on the Scholarly Paper Recommendation dataset. As future work, we plan to investigate the relationship between the *anti-aging* factor λ and researchers. Furthermore, we are interested in investigating the effects of enriching our modeling with meta-data from citing and referenced publications.

Acknowledgments. Work by Anas Alzoghbi was partially supported by the German Federal Ministry of Economics and Technology (BMWi) (KF2067905BZ). We thank Kazunari Sugiyama for his efforts in providing us with the complete Scholarly dataset. We also thank ACM for providing meta-data for our dataset.

References

1. Alpaydin, E.: Introduction to Machine Learning. Adaptive Computation and Machine Learning Series, MIT Press (2014)
2. Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., Nürnberger, A.: Research paper recommender system evaluation: A quantitative literature survey. In: Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation. RepSys '13 (2013)
3. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (Oct 2002)
4. Liu, D., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Mathematical Programming 45(1-3), 503–528 (1989)
5. Nascimento, C., Laender, A.H., da Silva, A.S., Gonçalves, M.A.: A source independent framework for research paper recommendation. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. JCDL '11, ACM (2011)
6. Rencher, A., Christensen, W.: Methods of Multivariate Analysis. Wiley Series in Probability and Statistics, Wiley (2012)
7. Rocchio, J.J.: Relevance feedback in information retrieval (1971)
8. Sugiyama, K., Kan, M.Y.: Scholarly paper recommendation via user's recent research interests. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries. JCDL '10, ACM (2010)
9. Sugiyama, K., Kan, M.Y.: Exploiting potential citation papers in scholarly paper recommendation. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM (2013)
10. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11 (2011)