Learning-to-Rank in research paper CBF recommendation: Leveraging irrelevant papers

Anas Alzoghbi Victor A. Arrascue Ayala Peter M. Fischer

Georg Lausen

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

ABSTRACT

Suggesting relevant literature to researchers has become an active area of study, typically relying on content-based filtering (CBF) over the rich textual features available. Given the high dimensionality and the sparsity of the training samples inherent to this domain, the focus has so far been on heuristic-based methods. In this paper, we argue for the model-based approach and propose a learning-to-rank method that leverages publicly available publications' meta-data to produce an effective prediction model. The proposed method is systematically evaluated on a scholarly paper recommendation dataset and compared against state-of-the-art model-based approaches as well as current, domain-specific heuristic methods. The results show that our approach clearly outperforms state-of-the-art research paper recommendations utilizing only publicly available meta-data.

CCS Concepts

•**Information systems** → **Learning to rank; Recommender systems;**

Keywords

Research paper recommendation; Learning-to-Rank; Content-based Recommendation; Model-based user profile

1. INTRODUCTION

Scholars and researchers are confronted with an overwhelming number of newly published research papers in their domain of expertise. Although advantageous in restricting the domain, keyword-based search tools typically available in digital libraries offer a limited help to researchers in locating the relevant content. As a result, researchers need to manually search within unspecific search results to identify paper(s) of interest. This is the situation where recommendaer systems have great potential, and indeed plenty of works adopted different techniques to tackle this problem. A recent extensive survey in this domain [3] identified content-based filtering (CBF) as the predominant approach for research paper recommendation because of the rich textual features available. For learning user profile, almost exclusively the focus was on relevance feedback approaches, building on the assumption that papers appearing in user's preference list have an equal (or a presumed extent) share in the underlying user taste. Thus, user profiles are constructed as aggregation of relevant papers' keywords. Based on the classification suggested by Adomavicius et al. in [1], these approaches are referred to as heuristic-based. In contrast, model-based approaches depend on a learning method to fit the underlying user model (profile). This enables constructing a better modeling of researcher-keywords relation in user profiles. But they require a large body of training data which is not intuitively available in this domain. As a result, little work on applying model-based approaches exists for this problem.

In this paper, we employ pairwise learning-to-rank [4] as a model-based technique for learning user profile. We incorporate both relevant and irrelevant "peer" papers -papers published in relevant papers' conferences- to formulate pairwise preferences and enrich the training set. Our main contributions include:

- We investigate and customize learning-to-rank for CBF research paper recommendation.
- We incorporate only a small set of data, restricted to publicly available metadata of papers. This makes our approach suitable for a much larger domain than previous approaches which require papers' full-text.
- We perform an initial, yet systematic study on a real-world datatset in which we show that our approach clearly outperforms existing heuristic- and model-based algorithms.

The rest of this paper is organized as following: the second section provides an overview of existing related work. In section 3 we present our approach and in section 4 we demonstrate experimental setup and results. Finally, we conclude in section 5 by summarizing our findings and situate this work within our future plan.

2. RELATED WORK

A rich amount of related work tackled the problem of research paper recommendation. collaborative filtering (CF) approaches [8, 13, 14] showed a successful application of

CBRecSys 2016, September 16, 2016, Boston, MA, USA.

model-based methods incorporating knowledge from other "similar" users. However, we restrict our search to content-based scenarios considering only information from the active user. In this domain, the main focus in learning user profile has been on heuristic-based approaches with a wide adoption of relevance feedback and cosine similarity [3]. Papers are recommended which are most similar to one or more of previously published or liked papers. In [10], De Nart et al. used extracted terms (keyphrases) from user's liked papers in constructing user profile. The profile has a graph representation, and the focus here was on the keyphrases extraction method and the graph structure. The approach of Lee et al. [6] proposed a memory based CBF, where users' papers are clustered based on their similarity, and candidate papers are ranked based on the distance from user's clusters. Sugiyama et al. in [11, 12] applied a relevance feedback approach utilizing all terms from the fulltext of the researcher's publications in addition to terms from the citing and the referenced papers in order to build profiles. All of these works are heuristic-based, where weights in user profile are set by aggregating individual keywords' scores of relevant papers. On the contrary, model-based approaches depend on machine learning techniques to learn user affinity towards keywords, promising a more representative user profile. In a previous work [2], we showed the superiority of a model-based method over relevance feedback methods for CBF research paper recommendations. We applied multivariate linear regression to learn researchers' profiles from their previous publications. Yet, the work was tailored to researchers with previous publications and didn't consider irrelevant papers. In [9], Minkov et al. presented a collaborative ranking approach for events recommendation. They compared it with a content-based baseline that applies pairwise learning-to-rank on pairs of relevant and irrelevant events. In our work, we follow similar approach in applying learning-to-rank on pairs of relevant an irrelevant papers. However, we push it further and investigate the quality of these pairs and their effect on the model performance.

3. PROPOSED APPROACH

This work targets users who have previously interacted with scientific papers and identified some as papers of interest (relevant papers). Having a set of relevant papers for a user, the recommendation process can start and a machine learning method is applied to fit a user profile (model). The learned model is used to rank a set of candidate papers and recommend the top ranked papers to the user. Our approach is to employ the pairwise learning-to-rank technique in building the user profile. We chose this method because of its desirable properties: It was proven to be successful in solving ranking tasks in similar problem domains like online advertising [7]. It also shows a good performance on problems with sparse data. The main idea of pairwise learning-to-rank is to build pairs of preferences out of the training set. Each pair consists of a positive and a negative instance. Afterwards, the pairs are fed as training instances to a learning algorithm, which in turn learns the desirable model. In the underlying problem, papers marked as interesting by users are the positive instances. However, the negative instances or the irrelevant papers are usually not explicitly provided by the users. This makes pairwise learning-to-rank not directly applicable on this setup. In our contribution, we seek implicit information about the irrelevant papers. For this, we start from the following hypothesis: when users identify relevant papers, they, to some extent, implicitly rate other papers published at the same conference (we call them *peer papers*) as irrelevant[1]. Based on this hypothesis, we utilize peer papers as irrelevant papers as follows: for each user, we build pairs of preferences out of relevant and peer papers. Such pairs are called pairwise preferences or for simplicity pairs, we will use these terms interchangeably along the paper. Afterward, we feed these pairs as training examples to a learning algorithm in order to fit the user's model. This model is used later to rank candidate papers and recommend top ranked ones to the user. Before delving deeper in the method details, we first introduce some notation. The function $peer(.)$ is defined over the interest set P^r_{int} of a user r. It delivers for a paper $p \in P^r_{int}$ the set of p's peer papers. In practice, this can be retrieved via digital libraries like DBLP registry[2]. For the paper modeling, we adopt a vector space model representation. Having the domain related keywords extracted from paper's title, abstract and keyword list as features, each paper p is a vector: $p = \langle s_{p,v_1}, ..., s_{p,v_{|V|}} \rangle$, with $v_i \in V$ is a domain-related vocabulary and s_{p,v_i} is a score reflecting the importance of v_i in p. We adopt the TF-IDF score as the weighting scheme. Based on this representation, the similarity between two papers is calculated by the cosine similarity between the papers' vectors.

3.1 Method Steps

An overview of the proposed approach is depicted in Figure 1. For the experimental setup only, we split user's r interest set P^r_{int} into training and test sets P^r_{train}, P^r_{test} respectively. However, this step is dropped out in the non-experimental recommendation scenario and the first step receives, in this case, the complete interest set P^r_{int}.

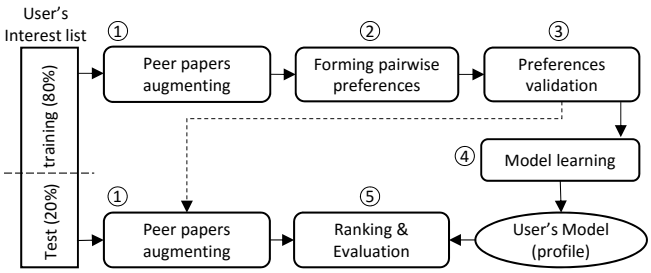

Figure 1: Overview the proposed approach steps

1. Peer papers augmenting: in this step, the peer papers are retrieved for all relevant papers. Retrieved peer papers serve as potential negative classes and are important for empowering the learning algorithm to construct a better understanding of user's taste.
2. Forming pairwise preferences: here we apply the concept of pairwise learning from learning-to-rank. The training set in this step is reformulated as a set of pairs $\mathcal{P}$, where each pair consists of two components: a relevant paper and an irrelevant paper. That is, each relevant paper $p \in P^r_{train}$ is paired with all papers from $peer(p)$:

$$\mathcal{P} = \{(p, p') | \forall p \in P^r_{train} \wedge \forall p' \in peer(p)\}$$

[1]Later, we introduce a validation process that checks the correctness of this hypothesis for each pair.
[2]http://dblp.uni-trier.de

A pair $(p, p') \in \mathcal{P}$ depicts a preference in user's taste and implies that p has a higher relevance to user r than p'.

3. Preferences validation: In the first step, we introduced the peer papers as negative classes based on the hypothesis mentioned earlier in this section. Yet, this can't be adopted as a ground truth due to: (a) it is not explicitly affirmed by users that they are not interested in peer papers; and (b) some peer papers might be of interest to the user but might have been overlooked. Having this in mind, not all pairwise preferences formulated in the previous step have the same level of correctness. Therefore, this step examines pairwise preferences and makes sure to pass valid ones to model learning. We propose two different mechanisms to accomplish this validation: pruning based validation and weighting based validation. We explain these techniques in the next section.
4. Model learning: In this step, we apply a pairwise learning-to-rank method (Ranking SVM [5]) to train a user model $\hat{w}_r$. Using validated pairwise preference from the previous step, we seek $\hat{w}_r$ that minimizes the objective function:

$$\hat{w}_r = \arg\max_{w_r} \frac{1}{2}||w_r||^2 + C.\mathcal{L}(w_r)$$

With $C \geq 0$ is a penalty parameter and $\mathcal{L}(w_r)$ is the pairwise hinge loss function:

$$\mathcal{L}(w_r) = \sum_{(p,p') \in \mathcal{P}} max(0, 1 - w_r^T(p - p'))^2 \qquad (*)$$

5. Ranking & Evaluation: Given the user's model as a result of the previous step, here we apply the prediction on candidate papers. For the experimental setup, this is the the test set, which is constructed out of relevant papers P^r_{test} (the positive instances), in addition to their peer papers as irrelevant papers (the negative instances).

3.2 Preferences Validation Methods

As pairwise learning-to-rank expects pairs that show contrast between negative and positive classes, pairs with "wrongly assigned" peers pose a potential noise to the learning process. After all, the validity of a pairwise preference (p, p') depends on the correctness of considering its peer paper p' irrelevant. The pair's relevant paper p forms the ground truth and hence, it can be considered as the reference point to decide whether p' is irrelevant or not. For each pair $(p, p') \in \mathcal{P}$ we measure the similarity between p and p', and adopt two methods to validate the pair based on this similarity:

Weighting Based Validation (WBV). This strategy is based on giving pairwise preferences different weights based on the dissimilarity between the pairs components. This boosts the importance for pairs with dissimilar components and assures that the more similar the pair's components are, the less important the pair for model learning is. Therefore, we weight the importance of each pair according to the distance (1-similarity) between the relevant paper and the peer paper. Then, we redefine the loss function from (*) to consider pairs' weights as following:

$$\mathcal{L}(w_r) = \sum_{(p,p') \in \mathcal{P}} max(0, 1 - w_r^T(1 - similarity(p, p'))(p - p'))^2$$

Pruning Based Validation (PBV). Here we filter out invalid pairwise preferences. Validity is judged based on the dissimilarity between the pair's components. If they prove to be similar, then we don't consider p' as an irrelevant paper and consequently, the pair (p, p') is not eligible for model learning. A similarity threshold τ is applied and a pair (p, p') is pruned if $similarity(p, p') > \tau$. In our experiments, we empirically test a range of values for τ and discuss the corresponding effect on the model.

4. EXPERIMENTS

4.1 Dataset & Setup

We evaluated the proposed approach on the Scholarly publication recommendation dataset from [12], including the extensions applied in our previous work [2]: Papers are identified and enriched with meta-data from the DBLP register, namely titles, abstracts, keywords and the publishing conference. The dataset contains 69,762 candidate papers, as well as the lists of relevant papers for 48 researchers. The number of relevant papers ranges from 8 to 208 with an average of 71 papers. After augmenting peer papers, we got a skewed distribution as the ratio of relevant papers to peer paper ranges from 0.45% to 3% with an average of 1.2%. We performed offline experiments with 5-folds cross validation following the steps outlined in Figure 1. For each researcher we randomly split the interest list into training and test sets; then, we learn researchers' models as described in section 3; finally, we evaluate the learned models on the test set. The test set consists of: (a) positive instances, the test relevant papers (20% of the researchers interest list) and (b) negative instances, the peer papers of the positive instances. This applies for all of our experiments, except for experiments on the pruning based validation method (PBV). In PBV, we filter out those pairs which components have a similarity higher than τ from the training set. Therefore, we apply the same rule on the test set and we filter out peer papers based on their similarity to the corresponding relevant paper. For example, given a similarity threshold τ and a relevant paper p from the test set, a peer paper $p' \in peer(p)$ is added as an irrelevant paper to the test set if and only if $similarity(p, p') \leq \tau$.

4.2 Metrics

We measured the following metrics to determine the performance for top k ranking and also overall classification. We show the averages over all researchers for each metric:

Mean Reciprocal Rank (MRR): evaluates the position of the first relevant paper in the ranked result.

Normalized Discounted Cumulative Gain (nDCG): nDCG@k indicates how good the top k results of the ranked list are. We look at nDCG for $k \in \{5, 10\}$

AUC and Recall: used to study the behavior of validation strategies PBV, WBV and the baseline algorithms: Logistic Regression and SVM.

4.3 Results & Discussion

In total, we performed three different experiments. The first experiment (with the results shown in Table 1) shows a superior performance for our weighting based validation method (WBV) over the state-of-the-art heuristic-based work (Sugiyama [12]) and model-based (PubRec [2]) approach.

The experiments were performed using the same features and datasets present in these works and show a clear lead over all metrics.

	MRR	nDCG@5	nDCG@10
WBV	0.728	0.471	0.391
PubRec	0.717	0.445	0.382
Sugiyama[12] via [2]	0.577	0.345	0.285

Table 1: WBV compared to state-of-the-art model-based and heuristic-based approaches

The second experiment compares the performance of our approach over other, baseline classification algorithms like SVM and logistic regression to provide a more general understanding of its capabilities. As shown in Figure 2, logistic regression showed a weak performance on all metrics, particularly on Recall. It didn't succeed in identifying relevant papers even when it is fed with a balanced training set. However, SVM showed a better ability to recognize the relevant papers with a better recall value, but produced a lot of false positives and this is clear from its lower MRR and nDCG values. In contrast, all variants of our method showed a superior performance in all metrics. Finally, we compare between the suggested pair validation techniques WBV and PBV, including tuning the latter by varying the similarity threshold τ from 1 (where no pairs are filtered, this case represents the CBF approach of [9]), down to $4 * 10^{-4}$ (where a lot of "noisy" pairs are pruned from the training set). WBV showed in general a very good performance, beating PBV for higher values of τ on all metrics except recall. There, PBV gives a slightly better recall even without filtering any pairs (when $\tau = 1$). This refers to the fact that weighting the pairs in WBV causes the model to miss some relevant papers, while PBV made models more capable of recognizing the relevant papers by eliminating the noisy pairs from the training set. When decreasing τ, PBV shows very good scores, but these results need additional investigation before leading to a clear conclusion. As mentioned earlier in this section, reducing τ also leads to a smaller number of irrelevant papers in the test set. This reduces the underlying bias in the test set which has an (additional) positive impact on the metrics, even though there is still a clear bias (the relevant/peer ratio is on average 11.2%) present at the lowest τ values.

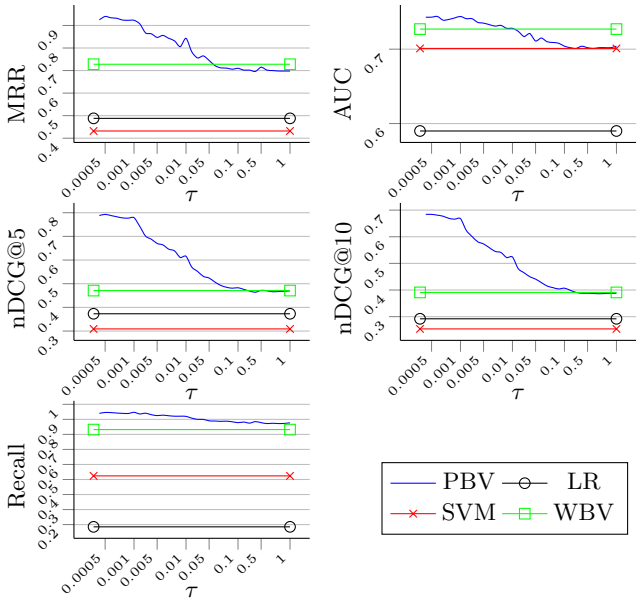

Figure 2: WBV and PBV compared with Logistic regression and Support Vector Machine

5. CONCLUSION

In this paper, we investigated the application of learning-to-rank in research paper recommendation. We proposed a novel approach that leverages irrelevant papers to produce more accurate user models. Offline experiments showed that our method outperforms state-of-the-art CBF research paper recommendations utilizing only publicly available metadata. Our future steps will focus on further understanding the effect of the similarity threshold in pruning based validation (PBV) on the model quality and study the suitability of pairwise learning-to-rank algorithms other than Ranking SVM for this problem.

6. REFERENCES

[1] G. Adomavicius, Z. Huang, and A. Tuzhilin. *Personalization and Recommender Systems*. 2014.

[2] A. Alzoghbi, V. A. A. Ayala, P. M. Fischer, and G. Lausen. Pubrec: Recommending publications based on publicly available meta-data. In *LWA*, 2015.

[3] J. Beel, B. Gipp, S. Langer, and C. Breitinger. Research-paper recommender systems: a literature survey. *IJDL*, 2015.

[4] L. Hang. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 2011.

[5] R. Herbrich, T. Graepel, and K. Obermayer. *Large Margin Rank Boundaries for Ordinal Regression*. 2000.

[6] J. Lee, K. Lee, J. G. Kim, and S. Kim. Personalized academic paper recommendation system. In *SRS*, 2015.

[7] C. Li, Y. Lu, Q. Mei, D. Wang, and S. Pandey. Click-through prediction for advertising in twitter timeline. In *KDD*, 2015.

[8] S. M. McNee and et al. On the recommending of citations for research papers. In *CSCW*, 2002.

[9] E. Minkov, B. Charrow, J. Ledlie, S. Teller, and T. Jaakkola. Collaborative future event recommendation. CIKM, 2010.

[10] D. D. Nart and C. Tasso. A personalized concept-driven recommender system for scientific libraries. *Procedia Computer Science*, 2014.

[11] K. Sugiyama and M.-Y. Kan. Scholarly paper recommendation via user's recent research interests. In *JCDL*, 2010.

[12] K. Sugiyama and M.-Y. Kan. Exploiting potential citation papers in scholarly paper recommendation. In *JCDL*, 2013.

[13] A. Vellino. A comparison between usage-based and citation-based methods for recommending scholarly research articles. In *ASIS&T*, 2010.

[14] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD*, 2011.