An LDA-Based Approach to Scientific Paper Recommendation

Maha Amami[1,2(✉)], Gabriella Pasi[1], Fabio Stella[1], and Rim Faiz[3]

[1] Department of Informatics, Systems and Communication, University of Milano Bicocca, Milan, Italy
{amami,pasi,stella}@disco.unimib.it
[2] LARODEC, ISG, University of Tunis, Tunis, Tunisia
[3] LARODEC, IHEC, University of Carthage, Tunis, Tunisia
rim.faiz@ihec.rnu.tn

Abstract. Recommendation of scientific papers is a task aimed to support researchers in accessing relevant articles from a large pool of unseen articles. When writing a paper, a researcher focuses on the topics related to her/his scientific domain, by using a technical language.

The core idea of this paper is to exploit the topics related to the researchers scientific production (authored articles) to formally define her/his profile; in particular we propose to employ topic modeling to formally represent the user profile, and language modeling to formally represent each unseen paper. The recommendation technique we propose relies on the assessment of the closeness of the language used in the researchers papers and the one employed in the unseen papers. The proposed approach exploits a reliable knowledge source for building the user profile, and it alleviates the cold-start problem, typical of collaborative filtering techniques. We also present a preliminary evaluation of our approach on the DBLP.

Keywords: Content-based recommendation · Scientific papers recommendation · Researcher profile · Topic modeling · Language modeling

1 Introduction

In the last years a big deal of research has addressed the issue of scientific papers recommendation. This problem has become more and more compelling due to the information overload phenomenon suffered by several categories of users, including the scientific community. Indeed, the increasing number of scientific papers published every day implies that a researcher spends a lot of time to find publications relevant to her/his research interests. In particular, recommender systems serve in this context the purpose of providing the researchers with a direct recommendation of contents that are likely to fit their needs.

Most approaches in the literature have addressed this problem by means of collaborative filtering (CF) techniques, which evaluate items (in this case papers) based on the behavior of other users (researchers), by exploiting the

E. Métais et al. (Eds.): NLDB 2016, LNCS 9612, pp. 200–210, 2016.
DOI: 10.1007/978-3-319-41754-7_17

rates assigned by other users to the considered items. However, generally, CF approaches assume that the number of users is much larger than the number of items [9].

This is verified in applications like movie recommendations, where there are usually few items and several users. For instance, the MovieLens 1M[1] dataset contains 1,000,209 ratings from 6,040 users and 3,706 movies [8]. Moreover the users are clients who are very likely to interact with the system several times, often consuming similar items; therefore ratings are quite easy to obtain. Hence, CF recommendation models can make accurate recommendations for most users in e-commerce domains.

On the contrary, in the domain of scientific papers recommendation, there are usually less users than papers, which results in the data sparsity problem. In fact usually there are few users who select the same papers, and thus finding similar users only based on explicit ratings of papers is a difficult task.

A second general issue that affects CF is the cold start problem, which occurs when a new item to be recommended has not been rated by any user. Furthermore, when a new user is introduced, the system is not able to provide recommendations. A possible strategy is to force the user to rate a minimum number of items before starting to use the system.

Content-based approaches recommend items based on both the items content and a profile that formally represents the user interests [12]. Content-based approaches exploit the items metadata and content to provide recommendations based on users preferences represented in the user profile. However, content-based recommendation approaches need reliable sources of the user interests. More precisely they rely on a user model (the profile) that specifies the user topical preferences; these preferences must be captured either by means of an explicit user involvement or by means of the analysis of various kinds of interactions of the user with the system (implicit feedback). In this way the cold start problem could be shifted from items to users [6,9].

To improve the recommendations some systems apply hybrid approaches [4], which combine content-based and collaborative filtering techniques.

By means of hybrid approaches [11,17,19] the cold start problem for both new users and items can be alleviated. For instance, in [17] the authors suggest to combine ratings and content analysis into a uniform model based on probabilistic topic modeling.

However, in [1] the authors show that producing accurate recommendations depends on the choice of the recommendation algorithm, and mainly it depends on the quality of the information about users. The noise injected in a user profile affects the accuracy of the produced recommendations.

In this paper we make the assumption that the users (researcher) scientific corpus (publications co-authored by the researcher) is a reliable source of data and information. In fact, in the task of writing articles, a researcher focuses on a set of topics related to her/his scientific investigations, and s/he uses a technical language related to those topics. These core topics play an important role in the

[1] www.grouplens.org/.

selection of unseen papers. The rationale behind the approach we propose in this paper is to make use of the researchers scientific corpus to formally define her/his profile: in this way the user model will exploit the core concepts contained in the articles authored by the researcher.

Formally, the approach we propose relies on topic modeling: the profile of a researcher is a topic model obtained by applying LDA to the abstract of a sample of the articles written by the researcher. Topic models provide the identification of core topics from a provided text collection; each topic is formally represented by means of a probability distribution of n-grams, i.e. sequences of n words. We propose then to formally represent each unseen article (to be recommended) by means of a language model. Then the topics generated by topic modeling from the authors collection and the language models of the unseen papers are compared to assess their similarity, which is employed by the recommendation mechanism.

The outline of this paper is the following. In Sect. 2 we review the related works based on LDA-recommendation models with a brief discussion of their limitations. The proposed content-based recommendation model is presented in Sect. 3. The evaluation of our recommendation model is presented in Sect. 4, with a description of the employed dataset and a discussion of the obtained results. In Sect. 5 we draw some conclusions and outline our future work.

2 Related Work

Several content-based recommender systems formally represent the user profile as a bag of words, represented by a vector. For example in [13] both researchers and unseen papers are represented as vectors in an n-dimensional space of terms, and the cosine similarity measure is applied to determine the relevance of a paper to a user profile.

In [5] both user profiles and unseen papers are represented as trees of concepts from the ACM's Computing Classification System (CCS); the recommender system matches the concepts in the user profile to each concept in the paper representation by means of a tree matching algorithm. Based on this technique the unseen papers in a scientific library are recommended to a user (researcher). A limitation of this approach is that the considered concepts are limited and too general to be able to well distinguish different topics.

Latent Dirichlet allocation (LDA) [3] has been employed as a technique to identify and annotate large text corpora with concepts, to track changes in topics over time, and to assess the similarity between documents. The purpose of this algorithm is the analysis of texts in natural language in order to discover the topics they contain and to represent them by means of probability distributions over words. For real world tasks, LDA has been successfully applied to address several tasks (e.g., analysis of scientific trends [2], information retrieval [18], and scholarly publication search engines[2]).

[2] Rexa.info.

In [10] the authors have proposed an LDA-based method for recommending problem-related papers or solution-related papers to researchers, in order to satisfy user-specific reading purposes. Here the LDA algorithm was used with a fixed number of topics to generate the document-topic distributions from the entire corpus. In this paper any comparative evaluation with an appropriate baseline is provided.

In [17] the authors have proposed an extension of LDA for recommending scientific articles called collaborative topic regression (CTR). This hybrid approach combines collaborative filtering based on latent factor models and content analysis based on topic models. Here matrix factorization and LDA are merged into a single generative process, where item latent factors are obtained by adding an offset latent variable to the document-topic distribution. Like CF approaches, this method is able to predict articles already rated by similar users. However, it performs differently on unseen papers, and it can make predictions to the ones that have similar content to other articles that a user likes.

In [19] the authors have proposed a hybrid recommender which is a latent factor model called CAT. It incorporates content and descriptive attributes of items (e.g., author, venue, publication year) to improve the recommendation accuracy.

However, both the CTR and the CAT systems suffer two limitations. First, they have problems when asked to make accurate predictions to researchers who have only few ratings. Second, they may not effectively support tasks that are specific to a certain research field such as recommending citations. For instance, a biological scientist who is interested in data mining applications to biological science might desire the recommender systems to support tasks such as recommending unseen papers on data mining techniques. While both CAT and CTR are using a rich source of metadata to generate recommendations, they are subject to certain limitations that affect their effectiveness in modeling citation patterns. The citation context, defined as a sequence of words that appear around a particular citation [14] is not highlighted in the learned models.

3 The Proposed Approach

In this section we introduce our LDA-based approach to scientific paper recommendation to address the issues pointed out in Sect. 2.

The rationale behind our approach is that the generation of the researcher profile should rely on the content generated by the researcher her/himself, as it exposes the topics of interests of the researcher, as well as the technical language s/he uses to generate her/his articles. The researcher profile is then conceived as a mixture of topics extracted by the LDA algorithm from the researcher past publications. To estimate if an unseen article could be of interests to the researcher, a formal representation of the unseen article by a language model is provided, which is then compared with each topic characterizing the researcher profile (we remind that a topic is formally represented as a probability distribution over the considered vocabulary, as also a language model is).

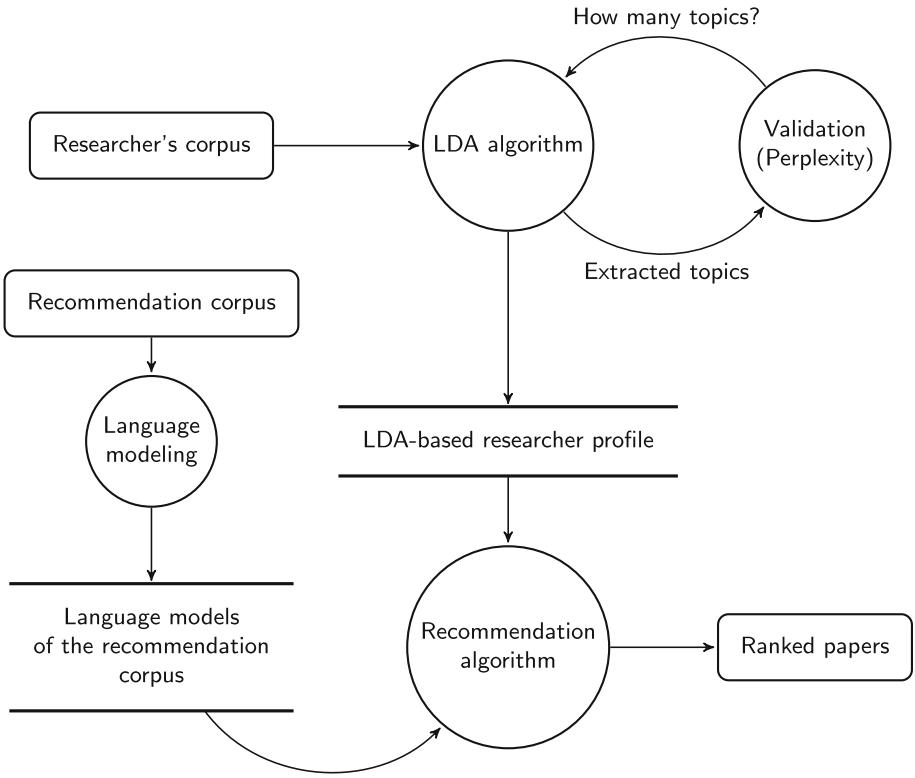

Fig. 1. Process flow of our recommendation model

3.1 Notation and Approach

We denote by Q_i the set of articles (co-)authored by a researcher A_i, and by $D = \{d_1, d_2, \ldots, d_M\}$ the set (consisting of M papers) that contains the articles unseen by researcher A_i and that could be potentially interesting to her/him. We refer to Q_i as to the researcher's corpus and to D as to the recommendation corpus. Furthermore, let W_i be the vocabulary employed in the extended set of articles, i.e., the set including both the researcher's corpus and the recommendation corpus ($Q_i \cup D$). W_i contains then the set of words w occurring in the corpus Q_i of researcher A_i and/or in the recommendation corpus D. The researcher profile is formally represented as a topic model obtained by applying the LDA algorithm to the texts constituted by the abstracts of the considered articles authored by the considered researcher (as it will be better explained in Sect. 3.2). We remind that formally each topic is represented by a probability distribution over the considered vocabulary $p_i(w|k)$ where w ranges over the vocabulary W_i and k is a topic among the K_i topics. In particular, we denote by $p_i(w|1 : K_i)$ the topic model associated with researcher A_i.

The objective of the content-based filtering algorithm is to recommend to researcher A_i the top m papers from the recommendation corpus D. We address this task by means of the following sub-tasks (see Fig. 1); *(i)* to define the topic model representing the researcher's interests: as previously outlined this is done by applying the LDA algorithm to the researcher's corpus Q_i, *(ii)* to validate the topics extracted by the LDA algorithm form the researcher's corpus, *(iii)* to evaluate if a given article d_j from corpus D has to be recommended to researcher A_i; to this purpose the distance between the validated topics and the language model of the article to be recommended is computed, and *(iv)* to rank papers belonging to D in descending order of similarity and recommend, to researcher A_i, the first m papers in the provided ranking. In Sect. 3.2 the above sub-tasks will be detailed.

3.2 Generation of the Researcher Profile and Topic Validation

As explained in Sect. 3.1, sub-tasks *(i)* and *(ii)* are performed by applying LDA to the texts extracted from the researcher's corpus Q_i for each researcher A_i. In particular, to select in a distinct way the optimal number of topics K_i for each researcher A_i the researcher's corpus Q_i is cross validated. More specifically, the optimal number of topics K_i for a given researcher A_i is selected by optimizing the cross validated perplexity, where perplexity [16] measures the uncertainty in predicting the occurrence of a word when using a given model. In topic modeling, the perplexity measures how well the topics extracted by LDA using the training papers (in-sample papers, i.e., a portion of Q_i), allows to predict the occurrence of words belonging to validation papers (out-of-sample papers, i.e., the papers belonging to Q_i which are not used by LDA to extract topics). Perplexity is defined as follows [7]:

$$Perplexity(Q_i^{out-of-sample}) = exp\{-\log \frac{p(Q_i^{out-of-sample}|p_i(w|1:K_i))}{|Q_i^{out-of-sample}|}\} \quad (1)$$

where $Q_i^{out-of-sample}$ is the set of out-of-sample papers belonging to Q_i.

3.3 The Recommendation Algorithm

Step (iii) of the proposed procedure consists of computing, for each researcher, the similarity between her/his K_i validated topics and the language model computed for each article in the recommendation corpus D. Formally, we propose to define the similarity between the profile of the researcher A_i and the article $d_j \in D$ as the maximum value among the K_i similarity values between the language model of article d_j and the K_i topics (probability distributions over words, as illustrated in Sect. 3.2) associated with the profile of author A_i. As each topic is represented as a probability distribution over words (as produced by the LDA algorithm), the similarity between a topic in the LDA-based researcher profile and the language model representing the article to be recommended is defined

by exploiting the *symmetrized Kullback Leibler divergence* between the above probability distributions. The language model associated with the unseen article $d_j \in D$ is computed as follows:

$$p(w|d_j) = \frac{nocc(w, d_j) + \frac{\mu nocc(w,Q)}{\sum_w nocc(w,Q)}}{\sum_w nocc(w, d_j) + \mu} \tag{2}$$

where $nocc(w, Q_i)$ is the number of occurrences of word w in Q_i, and μ is the hyperparameter of the Dirichlet distribution. Indeed, Formula (2), which is known as *Bayesian smoothing using Dirichlet priors*, does not incur the black swan paradox; i.e. a word w not occurring in the researcher's corpus Q_i is assigned a null probability value.

Given the topic distribution $p_i(w|k)$, i.e., the probability distribution over words w associated with topic k, extracted by LDA using the corpus of papers (co-)authored by researcher A_i, and the language model $p(w|d_j)$ associated with paper $d_j \in D$, we compute the symmetrized Kullback Leibler divergence between the topic k and the paper d_j as follows:

$$SKL(k, j) = \frac{1}{2} \sum_{w \in W_i} p(w|d_j) \log \frac{p(w|d_j)}{p(w|k)} + \frac{1}{2} \sum_{w \in W_i} p(w|k) \log \frac{p(w|k)}{p(w|d_j)} \tag{3}$$

Then, for each researcher A_i and paper $d_j \in D$, we find the topic k^* which minimizes the symmetrized Kullback Leibler divergence (3) across all the K_i validated topics associated with researcher A_i. Then, the similarity between researcher A_i and paper d_j is defined as follows:

$$Similarity(i, j) = \frac{1}{SKL(k^*, j)} \tag{4}$$

where we assume $SKL(k^*, j) \neq 0$ for each paper $d_j \in D$. Formula (4) corresponds to an optimistic computation of the similarity between researcher A_i and paper $d_j \in D$. Indeed, we are assuming that each researcher is summarized by a single topic k^*, i.e. the topic which is the most similar to the language model associated with the considered paper $d_j \in D$. Once we have computed for each paper $d_j \in D$ the similarity (4), we rank papers of the recommendation corpus D in descending order of similarity and use the ranking to recommend papers to researcher A_i. We then apply the same procedure to all researchers to implement step *(iv)* of the proposed procedure.

4 Evaluation of the Effectiveness of the Proposed Recommendation Approach

In this section we describe the experimental evaluations that we have conducted to verify the effectiveness of our approach. We first present the dataset and pre-processing step followed by details of the experimental procedure.

4.1 Dataset

We used the dataset of ArnetMiner[3], which contains 1.5 million papers from DBLP and 700 thousand authors. We have preprocessed this dataset to select only papers with complete titles and abstracts [15]; we denote this reduced set by $\mathcal{L}$ with $|\mathcal{L}| = 236,012$. To the purpose of our evaluations we have randomly selected 1,600 authors. We denote the set of the considered authors as $U = \{A_1, \ldots, A_{1,600}\}$ with $|U| = 1,600$ and we denote by $Q_A = \{q_1, \ldots, q_{N_A}\}$ the set of papers written by author A_i with $N_A \geq 10$.

We build the profile for each author based on her/his scientific production, namely the papers s/he (co-)authored by using the MALLET topic model API[4]. We have applied the following pre-processing steps to the titles and abstracts of the author's scientific production. First, we eliminated any words occurring in a standard stop list. Then, we converted the abstracts to a sequence of unigrams. To the purpose of defining a feasible test set, we have assumed that citations in papers written by a researcher A_i represent her/his preferences. We denote the test set by C, and its cardinality is defined as:

$$|C| = \sum_{A \in U} |C_A| = 24,539$$

where $C = \{d_j \in \mathcal{L} | \exists q_i \in Q, (q_i \rightarrow d_j)$ and $d_j \notin Q\}$ where $q_i \rightarrow d_j$ means q_i cites d_j.

4.2 Metrics

Two possible metrics to quantitatively assess the effectiveness are precision and recall. However, as the unrated papers (false positives) in the test set are unlabeled. It is not possible to establish if they are known by the user. This makes it difficult to accurately compute precision.

Hence, the measure we have used to assess the effectiveness of the proposed algorithm is recall. In particular, we have performed a comparative study to evaluate our approach with respect to the CAT and CTR algorithms. The recall quantifies the fraction of rated papers that are in the top-m of the ranking list sorted by their estimated ratings from among all rated papers in the test set. For each researcher A_i:

$$Recall@m = \frac{|N(m; A_i)|}{|N(A_i)|} \tag{5}$$

where $|.|$ denotes the cardinality of a set, $N(A_i)$ is the set of items rated by A_i in the test set and $N(m; A_i)$ is the subset of $N(A_i)$ contained in the top-m list of all papers sorted by their estimated relevance to the user model.

The recall for the entire system can be summarized using the average recall from all researchers.

[3] aminer.org/citation.
[4] mallet.cs.umass.edu/index.php.

4.3 Parameters

To the purpose of our experiments, we have selected the optimal number of topics K_i for each researcher by optimizing the cross validated perplexity as described in Sect. 3.2 with 5-cross validations. We used the left-to-right method defined in [16] to compute the perplexity.

The value of μ in the language model presented in Sect. 3.3 is a value determined empirically, and it is set to $\mu = 0.000001$.

4.4 Results

We compare the average Recall@m results for 1,600 researchers produced by our approach to the results produced by the state-of-the-art LDA-based recommender systems CTR and CAT. We report the averaged results of 5 repeated experiments to measure the performance of the different methods. Figure 2 shows the comparison of the average Recall@m values for 1,600 researchers with $m \leq 100$. Our approach achieves better Recall@m values than CTR and CAT systems. For $m = 40$, our approach performs better with a 60.4 % improvement over CTR and CAT.

As explained in Sect. 2, CTR and CAT systems are not able to make accurate recommendations to researchers who use few metadata (rates) to create the user profile.

The advantage of our approach does not only consist in being as accurate as or better than other hybrid approaches, but in employing a researcher profile that is only based on past publications and therefore using few metadata (only content item) and alleviate the cold start problem for new items. Furthermore,

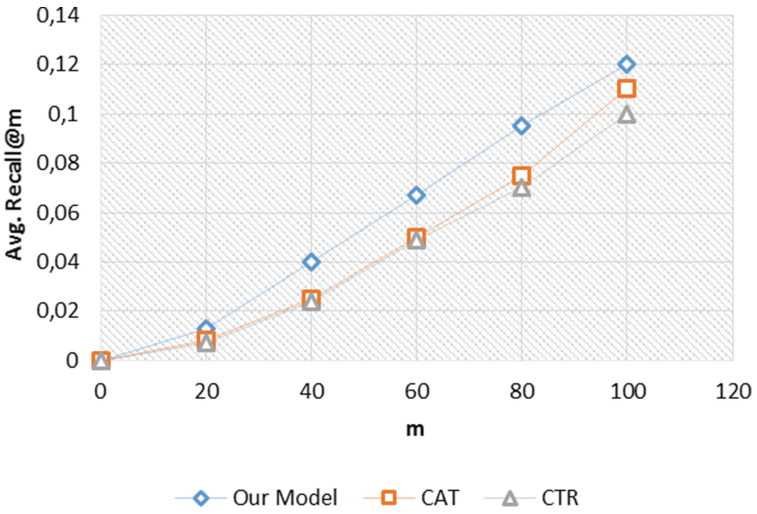

Fig. 2. Comparison of average Recall@m results for 1600 researchers

our approach offers ways to better explain researchers why a specific paper is recommended.

5 Conclusion and Future Work

In this paper we have proposed a fully content-based approach to the recommendation of scientific papers based on the researchers corpus. The researcher profile is built upon the topics generated by LDA algorithm on the researchers publications corpus. The profile built by this technique is easily interpretable, and it can explain the recommendation results. Our preliminary experiments show that our approach is performing well compared to the LDA-state-of-the art models, which make use of several metadata.

As a future work we aim to extend our work to include various attributes from the citation graph such as unseen article's recency, and the author's impact factor to improve the recommendation results.

References

1. Bellogín, A., Said, A., de Vries, A.P.: The magic barrier of recommender systems – no magic, just ratings. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) UMAP 2014. LNCS, vol. 8538, pp. 25–36. Springer, Heidelberg (2014)
2. Blei, D., Lafferty., J.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 113–120 (2006)
3. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**, 77–84 (2012)
4. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adap. Inter. **12**(4), 331–370 (2002)
5. Chandrasekaran, K., Gauch, S., Lakkaraju, P., Luong, H.P.: Concept-based document recommendations for CiteSeer authors. In: Nejdl, W., Kay, J., Pu, P., Herder, E. (eds.) AH 2008. LNCS, vol. 5149, pp. 83–92. Springer, Heidelberg (2008)
6. De Nart, D., Tasso, C.: A personalized concept-driven recommender system for scientific libraries. In: The 10th Italian Research Conference on Digital Libraries, pp. 84–91 (2014)
7. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**, 5228–5235 (2004)
8. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. (TOIS) **22**(1), 5–53 (2004)
9. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. Cambridge University Press, New York (2010)
10. Jiang, Y., Jia, A., Feng, Y., Zhao, D.: Recommending academic papers via users' reading purposes. In Proceedings of the 6th ACM International Conference on Recommender Systems, Dublin, Ireland, pp. 241–244 (2012)
11. Liu, Q., Chen, E., Xiong, H., Ding, C.H., Chen, J.: Enhancing collaborative filtering by user interest expansion via personalized ranking. IEEE Trans. Syst. **42**(1), 218–233 (2012)
12. Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: state of the art and trends. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 77–105. Springer, Heidelberg (2010)

13. Philip, S., John, A.O.: Application of content-based approach in research paper recommendation system for a digital library. Int. J. Adv. Comput. Sci. Appl. **5**(10), 37–40 (2014)
14. Sugiyama, K., Kan, M.A.: Exploiting potential citation papers in scholarly paper recommendation. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital libraries, pp. 153–162 (2013)
15. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 990–998 (2008)
16. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1105–1112 (2009)
17. 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, pp. 448–456 (2011)
18. Wei, X., Croft, B.C.: LDA-based document models for Ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178–185 (2006)
19. Zhang, C., Zhao, X., Wang, K., Sun, J.: Content + attributes: a latent factor model for recommending scientific papers in heterogeneous academic networks. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C.X., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 39–50. Springer, Heidelberg (2014)