Semantic clustering-based cross domain recommendation

Anil Kumar and Nitesh Kumar and Muzammil Hussain
Samsung R&D Institute (SRI) Delhi
Noida, India
Email: {anil.k06, nitesh.kumar, muzammil.h} @ samsung.com

Santanu Chaudhury and Sumeet Agarwal
Department of Electrical Engineering, IIT Delhi
Delhi, India
Email: {santanuc, sumeet} @ iitd.ernet.in

***Abstract*—Cross domain recommendation systems exploit tags, textual descriptions or ratings available for items in one domain to recommend items in multiple domains. Handling unstructured/unannotated item information is, however, a challenge. Topic modeling offer a popular method for deducing structure in such data corpora. In this paper, we introduce the concept of a common semantic space, spanning multiple domains, using topic modeling of semantic clustered vocabularies of distinct domains. The intuition here is to use explicitly-determined semantic relationships between non-identical, but possibly semantically equivalent, words in multiple domain vocabularies, in order to capture relationships across information obtained in distinct domains. We used the popular WordNet ontology to measure semantic relatedness between textual words. The experimental results shows that there is a marked improvement in the precision of predicting user preferences for items in one domain when given the preferences in another domain.**

I. Introduction

Recommender systems anticipate user preferences and recommend items to maximize the user experience along with content providers' revenue. These systems analyze users past interaction with items to detect his/her preferences and recommend new items accordingly. Two models are prevalent in literature: *content* and *collaborative*. The first model works by analyses actual information about the items to make predictions. The second model uses collaboration of similar users to decide what a user may like. However, if there does not exist any preference information of a user in one domain, it might be difficult to make any predictions in the first place. Cross-domain recommendation system [5] attempt to solve this problem by trying to link available (preference) information in one domain, to make predictions in another domain. Current cross-domain recommendation approaches exploit common labeled information: ontology [4], tags [17]; rating patterns [9], [6] as well as actual item information [19], [11], [10], in order to establish linkages across domains. These approaches perform well if domains are tightly coupled (e.g., movies and music) and some implicit relation exists between the domains. However, in case the domains are not implicitly related to each other, the resulting precision of the cross-domain recommendations suffers.

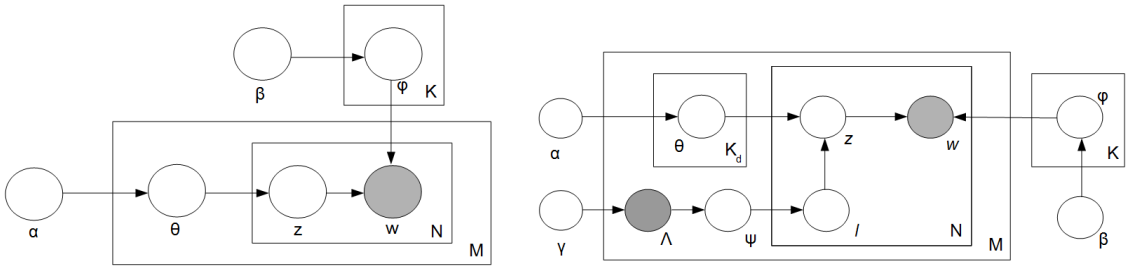
In this paper we propose a novel method to discover and use the common latent characteristics across multiple domains to recommend products across domains. The proposed method relies on *explicitly* obtained semantic information between vocabularies of different domains. Our approach can use both structured (e.g. tags) as well as unstructured textual information together with item ratings to learn a user profile over latent characteristics.

The rest of the paper is organized as follows. We first discuss the most relevant prior work in Section II; we then briefly revise some of the background required for understanding the paper in Section III; Section IV presents our latent shared characteristics based model for cross domain recommendation; Section V evaluates the algorithm for prediction accuracy on real world data; and finally we conclude the paper in Section VI.

II. Related Work

The cross-domain recommendation problem has been addressed by exploiting linkage among domains via common rating pattern sharing [9], [6], ontology [4], tags [17] and topics models [11], [19], [18] to establish linkages across domains. The rating pattern sharing based approaches rely only on implicit rating patterns that can transfer knowledge from one domain into another (no other relationship between the domains are considered). Shi et al. [17] proposed a probabilistic matrix factorization based method to effectively exploit the common tags between different domains for improving the quality of recommendations in each domain. This method assumes that user tag and item tags are available, thus performing poorly in situations where no tag information is available. In case of Fernndez-Tobas et al. [4], they make use of the ontologies (e.g. DBpedia) in order to establish connections between two domains, based on the availability of labeled data. It must be noted that obtaining tagged data is challenging, in the first place.

A number of papers [11], [19], [18] propose dealing with unstructured/unlabeled data by employing topic modeling. In Tang et al. [19], user is considered as a document who has consumed items in multiple domains and it is being used to determine the author-author collaboration while we are trying to recommend items to the user. In Low et al. [11], a hierarchical Bayesian model is proposed to solve multiple domain user personalization. One of the main issue with these work is that they fail to perform well when domains operate over non-identical vocabularies that may be semantically related. Such relationship can only be established through the use of explicit analysis of semantic relationships between words in the disparate vocabularies. Also, none of these paper have looked into exploited tagged information.



(a) Basic LDA (b) PLDA

Fig. 1. Plate notation of original Latent Direchlet Allocation (LDA) [2] and Partial Labeled Dirichlet Allocation (PLDA) [14]. Here, w and Λ are known/observable. We use PLDA in our experiments with the SCD algorithm.

III. Background

In this section, we will briefly discuss some of the underlying concepts of the paper. We first look at collaborative filtering. Thereafter, we look at the idea of topic modeling for a given data corpus and discuss current topic modeling techniques like Latent Direchlet Allocation (LDA).

A. Collaborative Filtering

Collaborative filtering is a method of making predictions about the interests of a user by collating preference information from multiple users. The main underlying assumption is that users with similar preferences/opinions on certain items are more likely to conform with each other on different items than, say, randomly chosen users. Predictions that are pertain to a *single* user are done from preference information from *multiple* users. For example, recommendation systems based on collaborative filtering in an e-commerce portal can make predictions on the product preferences for a user based on already purchased products by other users with matching purchasing trends.

There are a number of approaches to collaborative filtering in current literature. One of the most popular approach is a weighted sum of user ratings, proposed by Resnick et al. [16]. Here, the predicted rating for the user a for an item i is given by the weighted sum of all the ratings on item i by other users:

$$P_{a,i} = \bar{r_a} + \frac{\sum_{u \in U} (r_{u,i} - \bar{r_u}).w_{a,u}}{\sum_{u \in U} |w_{a,u}|} \quad (1)$$

where $\bar{r_a}$ and $\bar{r_u}$ are the average ratings for the user a and u respectively on all other rated items, and $w_{a,u}$ is the weight (typically similarity) between users a and u. The summations are over all users $u \in U$, where U consists of the set of the top k similar users to u (note that all such users may not have rated item i). In this paper, we denote the weight w as the similarity between user profiles.

B. Topic Modeling

Topic modeling offer a statistical model for deducing the abstract *themes* in a collection of documents. The intuition behind topic models is that assuming that documents are based on some inherent topic/theme, specific words related to that topic would appear in the document with a certain characteristic frequency. For example, "spaceship" and "planet" would appear frequently in documents concerning science fiction, while "plant" and "chlorophyll" would appear in documents related to botany with greater regularity. Thus, documents with multiple topics, each with a certain proportion, will have words from the *corresponding vocabularies* also appear in the similar parts.

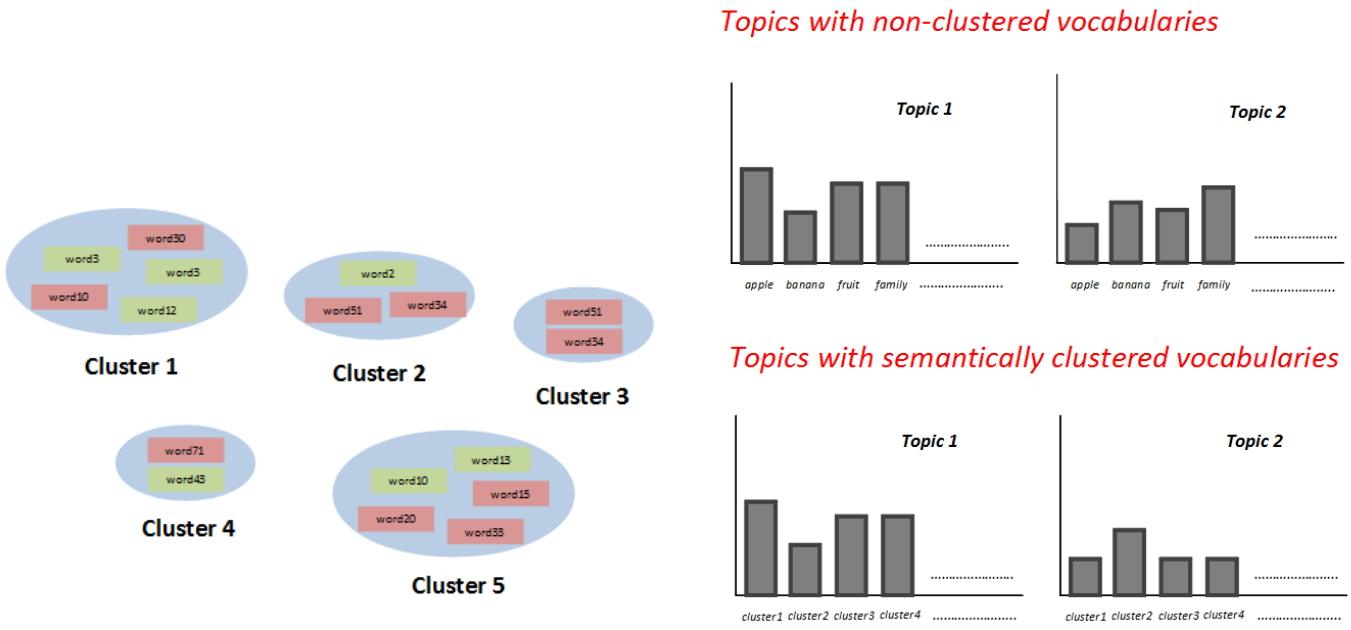
Major topic modeling approaches include latent semantic indexing (LSI) [13], latent semantic analysis (LSA) [3], probabilistic latent semantic analysis (pLSA) [7] and, most recently, Latent Dirichlet Allocation (LDA) [2]. LDA views documents as consisting of words drawn from a mixture of topics, where the topic distributions are assumed to have Dirichlet priors. Fig. III shows the graphical models, represented by *plate notations*, of two types of LDAs. The first type, shown in Fig. 1(a), represents the basic LDA fist proposed by Blei et al. [2]. Here, θ represents the topic distributions for the M documents, while ϕ represents the word distributions for each of the K topics. α and β denote the parameters of the Dirichlet priors for the per-document topic distributions and per-topic word distributions respectively. Fig. 1(b) shows the plate notation of the Partially Labeled Dirichlet Allocation, proposed by Ramage et al. [14], which makes use of tags denoted by Λ. Here, the words w and labels Λ are observable, while the per document label/tag distribution Ψ, topic distributions for the documents θ and the word distribution for each topic β are latent. We use PLDA extensively for our cross-domain recommendation experiments in this paper.

IV. Semantic Clustering based Cross Domain Recommendation

In this section, we will describe the Semantic-clustering based Cross-Domain (SCD) recommendation algorithm in detail. We first describe the concept of the common (latent) semantic space extracted from disparate vocabularies of distinct domains in Section IV-A. Thereafter, we discuss the SCD algorithm in detail in Section IV-B.

A. Common Latent Semantic Space

The SCD algorithm is based on the concept of a *common latent semantic space*. Topic modeling techniques give us *latent* features in a given corpus. These latent features, namely



(a) Clustering of cross-domain vocabularies based on semantic closeness. Words from domain A are shown in green color while words from domain B are shown in red.

(b) Comparison between topics based on basic vocabularies and those based on semantic word clusters. In the latter case, topics are defined as distributions over the semantic word clusters instead of singular words.

Fig. 2. Semantic clustering of vocabularies/corpora of distinct domains. This enables topic modeling, performed on the combined/clustered vocabularies, to capture the relationships between the different domains for unstructured corpora.

in the form of topics, denote those attributes that are implicit to the corpus. However, in case of corpora belonging to multiple domains, the challenge is to find relationships between the respective vocabularies of the domains since it is possible that the vocabularies may not match explicitly. Through the use of ontologies, to derive semantic relationships between words of vocabularies of distinct domains, it may be possible to build a common space that is valid for both domains *at the same time*. For example, Fig. 2(a) shows the clustering of words from different domains based other their semantic closeness. Thus, the definition of topics is also modified to be a distribution of semantic word clusters, instead of single words as shown in Fig. 2(b).

B. Algorithm

We will now describe the SCD algorithm in detail for producing cross domain recommendations from a source domain $\mathcal{S}$ to a target domain $\mathcal{T}$. Fig. 3 shows a graphical representation of the SCD algorithm.

Step 1: *Get item-based topic distributions*. Perform semantic clustering on the combined corpora of domain $\mathcal{S}$ and $\mathcal{T}$ to obtain clusters of semantically equivalent words. These clusters will replace actual words in the vocabulary in the corpora of domains $\mathcal{S}$ and $\mathcal{T}$. We then perform Latent Direchlet Allocation (LDA) analysis on individual corpus of domains $\mathcal{S}$ and $\mathcal{T}$ *separately*. Thus we get separate *item-based* topic distributions $\theta_{\mathcal{S}}$ $\{\theta^1, \theta^2, ... \theta^{K_{\mathcal{S}}}\}$ and $\theta_{\mathcal{T}}$ $\{\theta^1, \theta^2, ... \theta^{K_{\mathcal{T}}}\}$, where $K_{\mathcal{S}}$ and $K_{\mathcal{T}}$ are the number of topics for domains $\mathcal{S}$ and $\mathcal{T}$.

Step 2: *Get user-based topic distributions*. Obtain the topic distribution for users as the weighted sum of topic distributions of items rated by the user. Let $\mathcal{X}_u = \{\mathcal{X}_u^1, \mathcal{X}_u^2, ..., \mathcal{X}_u^{K_{\mathcal{S}}}\}$ denote the user profile for user u in source domain $\mathcal{S}$. $\mathcal{X}_u$ is obtained as:

$$\mathcal{X}_u^j = \frac{\sum_{i \in I} \theta_i^j \times r_u^i}{\sum_{i \in I} r_u^i} \tag{2}$$

where I is the set of items rated by user u, θ_i^j is the jth component of the topic distribution for item i and r_u^i is the rating given by user u on item i.

Step 3: *Find equivalent user profiles in the target domain* $\mathcal{T}$. This step involves making the user profile $\mathcal{X}_u$ equivalent to user profiles in domain $\mathcal{T}$, i.e., into $\hat{\mathcal{X}}_u$. First, we build a topic matching matrix between the topics in domain $\mathcal{S}$ and $\mathcal{T}$. Since topics in both domains, $\mathcal{S}$ and $\mathcal{T}$, are probability distributions over *same* vocabularies, namely the semantic word clusters obtained in Step 1, we use cosine similarity to measure their closeness:

$$sim(\psi_{\mathcal{S}}^i, \psi_{\mathcal{T}}^j) = \frac{\psi_{\mathcal{S}}^i . \psi_{\mathcal{T}}^j}{|\psi_{\mathcal{S}}^i||\psi_{\mathcal{T}}^j|} \tag{3}$$

where $\psi_{\mathcal{S}}^i$ is the ith topic in domain $\mathcal{S}$ and $\psi_{\mathcal{T}}^j$ is the jth topic in domain $\mathcal{T}$, where $1 \leq i \leq K_{\mathcal{S}}$ and $1 \leq j \leq K_{\mathcal{T}}$.

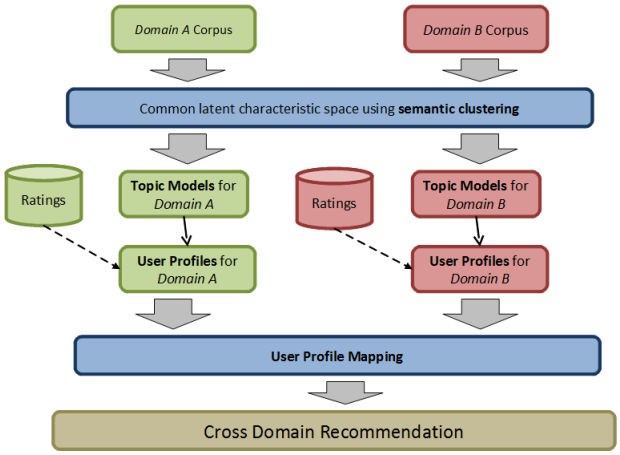


Fig. 3. Graphical representation of the SCD algorithm.

The equivalent user-profile of user u in target domain $\mathcal{T}$, $\hat{\mathcal{X}}_u$, is obtained using

$$\hat{\mathcal{X}}_u^j = \frac{\sum_{i=1}^{K_\mathcal{T}} \mathcal{X}_u^j \times sim(\psi_\mathcal{S}^j, \psi_\mathcal{T}^i)}{\sum_{i=1}^{K_\mathcal{T}} sim(\psi_\mathcal{S}^j, \psi_\mathcal{T}^i)} \quad (4)$$

Thereafter, we get a representation of the user profile in domain $\mathcal{S}$ in an equivalent form for domain $\mathcal{T}$. Next, we find the *similarity* of the translated user profile with other user profiles in domain $\mathcal{T}$, typically using the cosine similarity measure, though the Kullback-Leiblar Divergence can also be used. Finally, we are able to gather the best matching user profiles in target domain $\mathcal{T}$ to the current user profile in domain $\mathcal{S}$.

Step 4: *Generate cross-domain recommendations in target domain* $\mathcal{T}$. First, we take the top M matching profiles in domain $\mathcal{T}$ with respect to the user profile $\hat{\mathcal{X}}_u$. Then, we use collaborative filtering to obtain rating of all items in domain $\mathcal{T}$ (for all the user profiles in domain $\mathcal{T}$). Finally we recommend the top N items in the target domain $\mathcal{T}$, based on the rating of items in domain $\mathcal{T}$, to user u.

V. Evaluation

In this section, we will discuss the experimental evaluation of the proposed SCD algorithm. We will first describe the dataset. Then, we will describe the setup for the experiments. Finally, we will look at the results.

A. Dataset

We use the MovieLens [15] and the BookCrossing [20] datasets for our experiments in the cross-domain recommendation. We downloaded the MovieLens dataset with 100,000 user ratings. The MovieLens dataset information about 943 users and 1581 movies with a rating ratio of 5.23%.. We also crawled movie and book description information from the imdb.com and amazon.com websites respectively. The original BookCrossing dataset comprised of more than 1.1 million ratings (with rating scale varying from 1 to 10) provided by 278,858 users on 271,379 books. We carefully selected frequently reviewed books (7472 books rated by at least 20 users) and active users (3156 users with more than 20 ratings), with a rating ratio of 0.86%. We finally normalize rating of both domains to range from 1 to 5. We preprocessed the crawled data, removing stop words, and performed stemming. We then used TF-IDF to rank the importance of all words in their respective documents, and considered only the top 25%.

B. Experimental Setup

We implemented the SCD algorithm in Java. In order to perform the laborious task of semantic clustering among millions of words, we make use of the Hadoop distributed computing framework. We create an offline map of semantic relationships between all words, and thereafter use the data for experiments involving the SCD algorithm. We make use of the popular WordNet ontology [12] (via the ws4j library [1]) to establish semantic relatedness between textual words. We use the Lesk algorithm [8] for establishing semantic relatedness between words. We use community clustering algorithms (JUNG2, Java Universal Network/Graph framework) for forming the word clusters, using a threshold of 3 for the Lesk algorithm. We mark 80% of the dataset as training data and the remaining as test data. Since we do not have a dataset where a single user has simultaneous movie and book rating, we use the following method for evaluation: we first predict a items in a target domain from a given source domain. Next we take those predictions as input, and make predictions for the original source domain and calculate the precision.

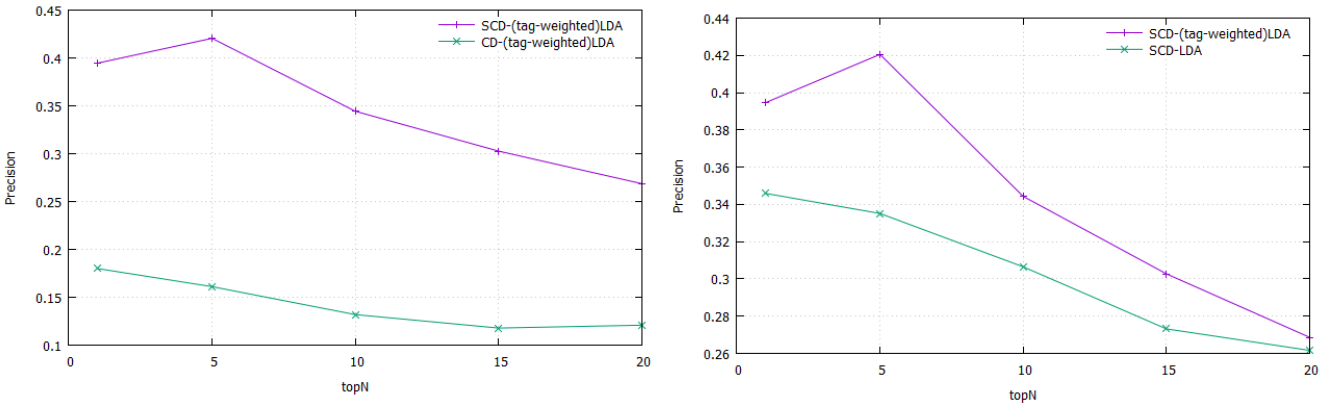
C. Results

The results for our experimental analysis for the SCD algorithm are presented in Fig. V. In Fig. 4(a), we see that using semantic clustering for topic-modeling based cross-domain recommendation increases the precision by more than twofold. This indicates that the semantic clustering is successful in bridging the two domains as a common space for knowledge transfer. One example word cluster we found was {wavelet, ruffle, flounce, furbelow, undulate, ripple, cockle}. A single word-cluster then replaced all these words in the modified vocabulary of the corpora of both domains.

Next, we evaluate the effect of using partially tagged data when semantic clustering is employed. We used the Partial Labeled Dirichlet Allocation (PLDA) [14] for taking advantage of both the tagged data, while the basic LDA cannot do the same. Thus, while for LDA we use only the text corpora generated from movie and book domains dataset, in case of PLDA, we also used tags along with text corpora. In Fig. 4(b), we see that using PLDA performs better than basic LDA since the tags are able to impart a strengthened accuracy for the cross-domain recommendation algorithm.

VI. Conclusion

In this paper, we introduced a new approach to do topic-modeling based cross-domain recommendation - that of using semantic clustering to form a common latent space between disparate corpora of different domains. Here, we redefine topics as distributions over word-clusters instead of single



(a) Comparison of performance of tag-weighted LDA-based cross-domain recommendation and semantic clustering enhanced cross domain recommendation.

(b) Comparison between topics based on basic vocabularies and those based on semantic word clusters. In the latter case, topics are defined as distributions over the semantic word clusters instead of singular words.

Fig. 4. Evaluation of semantic-clustering based cross-domain recommendations.

words, in order to capture the inherent similarities between vocabularies of multiple domains. We devised a cross-domain recommendation system that is based on collaborative filtering and uses these modified version of topics. We show that our method provides significant improvements over the precision of cross-domain recommendations compared to the scenario when only the combined (but non-clustered) vocabularies are used. In future, we would want to explore cross-domain recommendation systems based on topic modeling and semantic clustering over different types of media, such as image and video together with textual information about the item.

REFERENCES

[1] Ws4j (wordnet similarity for java). https://code.google.com/p/ws4j/.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[4] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. A generic semantic-based framework for cross-domain recommendation. In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pages 25–32. ACM, 2011.

[5] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. Cross-domain recommender systems: A survey of the state of the art. In *Proceedings of the 2nd Spanish Conference on Information Retrieval. CERI*, 2012.

[6] Sheng Gao, Hao Luo, Da Chen, Shantao Li, Patrick Gallinari, and Jun Guo. Cross-domain recommendation via cluster-level latent factor model. In *Machine Learning and Knowledge Discovery in Databases*, pages 161–176. Springer, 2013.

[7] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[8] Michael E Lesk. Word-word associations in document retrieval systems. *American documentation*, 20(1):27–38, 1969.

[9] Bin Li, Qiang Yang, and Xiangyang Xue. Transfer learning for collaborative filtering via a rating-matrix generative model. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 617–624. ACM, 2009.

[10] Shuangyin Li, Guan Huang, Ruiyang Tan, and Rong Pan. Tag-weighted dirichlet allocation. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 438–447. IEEE, 2013.

[11] Yucheng Low, Deepak Agarwal, and Alexander J Smola. Multiple domain user personalization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–131. ACM, 2011.

[12] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM, 1998.

[14] Daniel Ramage, Christopher D Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465. ACM, 2011.

[15] GroupLens Research. Movielens dataset. http://grouplens.org/datasets/movielens/.

[16] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM, 1994.

[17] Yue Shi, Martha Larson, and Alan Hanjalic. Generalized tag-induced cross-domain collaborative filtering. *arXiv preprint arXiv:1302.4888*, 2013.

[18] Shulong Tan, Jiajun Bu, Xuzhen Qin, Chun Chen, and Deng Cai. Cross domain recommendation based on multi-type media fusion. *Neurocomput.*, 127:124–134, March 2014.

[19] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1293. ACM, 2012.

[20] Cai-Nicolas Ziegler. Bookcrossing dataset. http://www2.informatik.uni-freiburg.de/ cziegler/BX/.