Chapter 7

Discussions

"It is through science that we prove, but through intuition that we discover."

-Henri Poincare (1854-1912)

Our primary objective in this thesis was to build a multi-objective academic recommender system. So we developed in this thesis a series of principled approaches for both journal and collaborator recommendations. Mainly we were interested in addressing cold-start issues for a new researcher and a new venue along with other issues like data sparsity, diversity, stability, and relevance problems.

7.1 Summary and Contributions

We summarize and highlight here the main findings and contributions of the thesis in light of the research goals stated in Chapter 1. Critical discussion of the entire work will also reveal the underlying connection across the preceding chapters.

7.1.1 RQ1: How to Handle Cold-start Issues in Journal Recommendation?

Most of the existing approaches like CF, CBF, and SF take the researcher's past publication records as input to suggest relevant venues. These approaches fail to provide recommendations in case of new researchers or researchers with fewer publications, or researchers with a fewer number of co-authors (isolated researchers).

We proposed a unified framework DISCOVER to address the problem of cold start issues for new researchers and new venues (Chapter 3). We adopted a cascading approach of social network analysis and contextual similarity in the proposed framework to address the issues mentioned above.

Table 3.13 shows that, even if the seed paper comes from new researchers and new venue, DISCOVER can predict the original venue at the early ranks. It considers only the current area of interest of a researcher along with the title, keywords, and abstract as inputs to recommend the same. Similarly, for a new venue, DISCOVER considers venues with no citation records with abstract similarity and provides equal opportunities in the recommendation (Consideration of Set-II papers, as discussed in Sec. 3.3.8).

DISCOVER also considers those papers which are shortlisted via anyone of centrality measures based filtering, as discussed in Sec. 3.3.4 (degree, betweenness, closeness, eigenvector, HITS score). This way, if a paper lacks in one or more factors in the citation profile, it can qualify through other centrality measures implying a fair chance to all potential papers.

The cold-start issue for new venues is given more careful attention in CNAVER (Chapter 4). It incorporates a VVPN model primarily to focus on venues. In addition to abstract similarity (paper with no citations), meta-paths features (common paper, author, citation, co-citation, terms, or topics) are considered to provide weights among venues. Venues having less number of papers and citations are also getting some weights to links with other related venues in the venue-venue graph (VVG) (see Sec. 4.6.2). Therefore chances of inclusion of new venues in the recommendation lists are relatively higher.

Examination of Table 4.8 and Table 4.9 reveals that, irrespective of seed paper associated with new researchers and new venues, CNAVER can predict the original venue at an early stage of recommendations. It does not require past publication records; rather, it considers only the current area of interest along with the title and abstract as inputs to recommend venues.

DeepRec can address new researcher problems reasonably well but does not solve the issue for new venue. We observe that it requires a minimum of $15 - 20$ papers for each venue to extract and learn contextual features for desired recommendations.

7.1.2 RQ2: How to Address Data Sparsity, and Diversity Issues in Journal Recommendation?

Data Sparsity

To address the issue of data sparsity, our initial proposal was DISCOVER. It used diverse techniques like keyword-based filtering, centrality measures, and citation analysis like Bibliographic coupling (BC) and Co-citation score (CC) at various stages, one after the other. These techniques, at each stage, substantially reduce the number of candidate papers that are potentially related to a given seed paper. 90% reduction occurs after the initial keyword-based search strategy (see Sec. 3.3.3), and there are reductions in other steps as well down the line (see Table 3.14). However, DISCOVER was not able to sufficiently remove irrelevant papers and sometimes unable to reduce the bibliographic network size due to insufficient domain-specific keywords. We observed that a minimum of 3 to 5 keywords are necessary to identify relevant papers (initial filtering) in DISCOVER.

In CNAVER we make recommendations independent of keywords. Here, we adopt two-stage filtering techniques such as centrality measures based citation analysis and contextual similarity like LDA on abstract and Doc2Vec on the title. This filtering strategy considers both importance and relevance parameters to reduce the bibliographic network size and also to increase the relatedness among papers. We observe that this two-stage filtering technique can reduce the bibliographic network size up to more than 97% (3,079,007-32,069 papers).

Although in both DISCOVER and CNAVER, there is substantial reduction in the number of papers, we still need two phases of analysis: citation and contextual similarity analysis. These processes still require some efforts to store and organize those shortlisted papers properly (by storing title, abstract, and citations relationship among papers).

In order to make better space-time utilization, we propose deep learning-based venue recommendation model DeepRec (Chapter 5). It transforms high dimensional and sparse embedding matrix into a lower-dimensional and dense set using CNN based deep learning technique. CNN is specifically designed to process temporal, latent contextual aspects of high dimensional and sparse input. As discussed in Sec. 5.5.7, DeepRec can address the sparsity issue to a great extent.

Diversity

DISCOVER adopts diverse techniques of contextual analysis and social network analysis in a cascading manner (sequential). Thus it recommends relevant venues that are diverse. Table 3.11 scores show good diversity they are. However, due to diverse venues existing in the dataset (MAG is a collection of the heterogeneous domain). Although diversity is reasonably addressed, we need venues with more diversity in scenarios when a sufficient number of venues are not available or venues recommendations in a subdomain (IR, NLP, ML, etc.).

In CNAVER venues are given special importance in VVPN model that contribute in parallel to the content of the paper (PPPN model). Both PPPN and VVPN bring in more venues through a network analysis that DISCOVER could not. We can not straightway compare DISCOVER and CNAVER as different datasets are used in. Although DBLP dataset is more biased to major computer science journals and proceedings such as computing, database, and logic programming still CNAVER shows a higher diversity (Table 4.18).

Diversity is not a major objective in DeepRec. A stacking ensemble (stacked generalization) model is used for training LSTM and CNN based architecture together. Here both CNN and LSTM captures contextual information (local, global hidden features) from texts. Diversity is not taken into consideration here, and thus recommended venues are of lower diversity scores as compared to that of CNAVER (Table 5.16).

7.1.3 RQ3: How to Improve Relevance and Stability in Journal Recommendation?

Relevance

To improve the relevance of recommendations, we presented a content-aware system DISCOVER with the title, keywords, and abstract of a paper as input. Chapter 3 demonstrates the effectiveness of DISCOVER in terms of relevance (e.g. P@15= 0.684) over other state-of-the-art techniques. However, due to dependency on domain-specific keywords, this approach is not so effective in providing relevant recommendations in all scenarios. Sometimes, a journal may change its scope and objectives. DISCOVER did not have any mechanism to capture variation in scope of a venue with time. Thus it

showed average performance in terms of precision@k in few sub-domains like SE, MM, and DM as depicted in Fig. 3.12 and Fig. 3.13 (see Sec. 3.5.7).

To capture the shift in a venue's scope with time, an age-discounted weighting scheme is proposed in CNAVER. The topics from recently published papers are prioritized, while topics from older publications are penalized in the age-discounted weighting scheme. The overall results of CNAVER in Tables 4.13, 4.14, 4.15, and 4.16 demonstrate better recommendations in terms of relevance than the state-of-the-art methods (P@15=0.762). However, when the topmost R papers similar to a given seed paper are loosely coupled in the bibliographic citation network, Jarvis Patrick may create clusters with less number of related papers (Sec. 4.9.11). Hence, the proposed model CNAVER may fail to capture the relevant papers resulting in possibly irrelevant venue recommendations.

To provide a solution for the above issue and to make the recommendations independent of the bibliographic network, we proposed an ensemble learning-based model, DeepRec. To enhance the recommendation quality in terms of relevance, we extracted latent features from abstract and title with CNN and LSTM models and combined them by training a meta-model employing stacked generalization technique. The analysis in Tables 5.10, 5.11, 5.12, and 5.13 demonstrates the efficacy of DeepRec in terms of relevance (P@15=0.903).

Stability

To maintain stability, the abstract similarity is computed using LDA, Okapi BM25+, and non-negative matrix factorization (NMF) techniques in DISCOVER. Later on, score-based fusion techniques such as CombMNZ is applied to fuse the similarity score of both LDA and NMF in order to maintain the stability of overall recommendations. But after addition of new papers into the model, we observed some changes in ranking order of venues recommended by DISCOVER (Table 3.11). This is because of the involvement of citation analysis (centrality measures calculations for Set-I papers) in the framework and identification of ranking of venues based on abstract similarity (mainly Set-II papers).

To alleviate this ordering issue in DISCOVER and also to make a stable system, the fusion model CNAVER is proposed incorporating both PPPN and VVPN models. Any network-based approach is known to cause instability in the ranks with time as the introduction of new nodes or edges change the topology and thereby change recommenda-

tions [42]. We therefore also took into account content-based approaches at several stages within both PPPN and VVPN pipeline. CNAVER shows a lower MAS than all other standard approaches (Table 4.18) and demonstrates the effectiveness in terms of stability.

In DeepRec, we adopted a stacked generalized ensemble leaning technique of both CNN and LSTM models in order to capture the relevance and to extract low dimensional latent factors of high dimensional input aiming to cope with stability issue. It also leveraged the demerits of CNN and LSTM by ensembling them to recommend the most suitable venues. To fully exploit contextual similarity, both abstract and title are considered. DeepRec shows a lower MAS than all other standard approaches (Table 5.16) and thus demonstrates the effectiveness in terms of stability of recommended venues.

7.1.4 RQ4: How the Overall Quality (Popularity) of Recommendation is Enhanced in Journal Recommender System?

In this thesis, popularity is used as a measure to denote the quality of recommended journals in terms of Google's H5-Index. We adopt three-stage layered approach: social network analysis, citation analysis and main path analysis where each layer performs a specific task in DISCOVER. We employed various social network analysis measures (centrality measures) such as degree, betweenness, closeness, eigenvector, and HITS in order to identify the important papers in a bibliographic citation network.

We integrated bibliographic coupling and co-citation components along with hop-distance to obtain candidate score (C-score) of individual papers in citation network. This C-score can reflect both semantic similarity and citation strength. The top C-scoring papers are shortlisted for further main path analysis. We employed key-route identification based main path analysis to identify most significant path or structural backbone in the evolution of a scientific field. We observe that the three-stage approach is able to identify the most important papers and resulting in quality journals (Sec. 3.5.6).

In CNAVER we adopted social network analysis to determine the importance of papers in a citation network. CNAVER mainly focused to address cold-start issue for new venue and diversity. CNAVER adopted a fusion approach integrating both PPPN and VVPN models to provide optimum recommendations. However, there is no such filtering technique other than centrality measures to filter quality venues. Due to which

Table 7.1: Comparative study among all proposed venue recommender systems

Parameters	**DISCOVER**	**CNAVER**	**DeepRec**
Inputs	Keywords, Title, Abstract	Title, Abstract	Title, Abstract
Methods	SNA, Content similarity	Content, Citation analysis	Deep learning
Strategies	Cascade (sequential)	Mixed (fusion)	Ensemble (stacking)
Datasets	MAG	DBLP	DBLP
Cold-start	NR, NV	NR, NV	NR
Relevance	Low	Moderate	High
Sparsity	Moderate	Moderate	High
Diversity	Moderate	High	Low
Stability	Low	Moderate	High
Popularity	High	Moderate	Moderate

NR and NV denote new researcher and new venue

the recommended venues are of lower quaity as compared to DISCOVER. Although direct comparison is not possible as they employed different datasets (Fig. 4.8).

In DeepRec both CNN and LSTM models are combined through stacked generalization to capture contextual information (local/global hidden features) from texts. CNN is mainly adopted to extract local structure of the data, while LSTM can capture the temporal correlation and dependencies in the text snippet. We considered the dataset prepared for online evaluation (journals having atleast 500 papers) to measure the venue quality. Although due to this biasing, chances of capturing better quality journals in terms of H5-Index are relatively higher. But still recommended venues are of moderate quality in terms of H5-Index as compared to DISCOVER Fig. 5.6b).

We include a comparative study among all proposed venue recommender systems to get an overall idea of their individual strengths and weaknesses (Table 7.1). During this comparative study various issues solved by journal recommender systems are taken into consideration. We employed three scales (low, moderate, and high) to measure the degree of solving a particular issue in journal recommendation.

7.1.5 RQ5: How is to Improve Overall Relevance and also to Handle Cold-start Issues in Collaborator Recommendation?

Relevance

DRACoR is mainly designed to recommend MICs incorporating Meta-path aggregated Random walk based Collaborator Recommendation (MRCR) that finds out MPCs with Deep learning-Boosted Collaborator Recommendation (DBCR) models that find MVCs (Chapter 6).

To capture the shift in an author's research interest with time, a time-inverse logarithmic weighting scheme is proposed. The recent research area is prioritized, while old areas are penalized in the time-aware weighting scheme. This mechanism can capture active researchers that share similar research interests at the peer level in MRCR model.

Deep learning incorporating Word2Vec and Long Short Term Memory (LSTM) techniques are employed in DBCR. This framework is able to identify the association and collaboration compatibility among researchers. The fusion of both MRCR and DBCR can provide relevant collaborators for a target researcher.

The analysis in Tables 6.9, 6.10, 6.11, and 6.12 demonstrates the effectiveness of DRACoR over state-of-the-art collaborator recommendation models with substantial improvements in F1-score, MRR, and nDCG on both hep-th and DBLP datasets (Sec. 6.10.3).

Cold-start Issues

State-of-the-art techniques mainly apply their models on the co-authorship network (CN) to provide recommendations. They fail to provide meaningful recommendations for researchers with less number of co-authors, or researchers with fewer publication records. Sometimes they are unable to even initiate recommendations, especially for isolated researchers (single author). In order to address such cold start issues, Author-Author Graph (AAG) has been used instead of the co-authorship network (CN) (Sec. 6.6.1).

We run a topic model on abstracts and Doc2Vec on titles on year-wise publications to capture dynamic research interests of researchers in MRCR model. Author-author cosine similarity is computed from the feature vectors extracted from abstracts and titles

and is then used to weigh edges in the author-author graph (AAG). This mechanism not only captures the current research interest but also can provide some weight among researchers, including those who prefer to work alone (isolated researcher). The addition of meta-path features and profile-aware features can also help to link researchers having less number of co-authors in AAG. Therefore isolated researchers, researchers with less number of co-authors, or researchers with fewer publication records are also getting an equal chance for inclusion in the final recommendation.

Experiments with varying academic level (n_c) and varying target researcher's degree (n_d) as described in Sec. 6.9.5 show the effectiveness of DRACoR. Analysis in Figs. 6.8a, 6.9b, 6.11a, and 6.12b demonstrate that DRACoR can recommend both new and old collaborators irrespective of degree and academic level of a target researcher.