Chapter 1

Introduction

"Begin at the beginning and go on till you come to the end; then stop."

-Lewis Carroll (1832-1898)

This chapter serves as the starting point for the thesis and establishes the key concepts and vocabulary used in the rest of the document. We begin with a general introduction to the recommender system in Section 1.1, followed by a brief description of existing methodologies in Section 1.2. General limitations of existing methods in academic recommender systems are illustrated in Section 1.3. In Section 1.4 we provide motivation of our work. Section 1.5 highlights the research goals by defining various research questions. We summarise the main contributions of the thesis in Section 1.6. Finally, Section 1.7 presents the layout of the rest of the thesis.

1.1 Recommender Systems

In recent years, with the rapid proliferation of information technology and ubiquitous computing, a variety of channels and methods to access information have brought great convenience for users [1]. The case of e-commerce, streaming platforms, and social networks, which constitute a substantial part of the web's traffic nowadays, is especially interesting. Typically such services offer a varied and vast amount of content to their customers: more than 200 million products in Amazon.com, 30 million songs in Spotify, 10,000 movies in Netflix, 248 million active users in twitter sending 500 million messages every day, etc. However, the geometric growth of data makes it difficult for users to find information that meets their own needs in time, so “big data” leads to information overload problem and makes a lot of irrelevant, redundant information interfere with users' choices [2].

In the era of big data, recommender systems that aim to suggest items of potential interest for solving information overload have attracted growing amounts of attention from the research community [3]. Recommender systems emerged as an individual field of research in the mid-1990s and derived from different other research areas like cognitive science, approximation theory, information retrieval, forecasting theory, consumer modeling, and also management science [4]. Recommender systems have become a pervasive technology in a wide spectrum of everyday applications and can be said to be familiar to the general public.

A recommender system is a specific type of decision support or advice-giving system that guides users in a personalized way towards exciting options, where a large pool of such options are available [5]. Resnik and Varian [6] describe the recommender system as:

> *“People provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients”.*

While this seems to be a suitable definition for the early recommender system in the late 90s, recommender systems have advanced a lot since then. Nowadays, the term can be described more broadly as defined by Burke et al. [7]:

> *“Any system that produces individualized recommendations as output or has the effects of guiding the user in a personalized way to interesting or useful objects in a large space of possible options”.*

According to Adomavicius and Tuzhilin et al. [4], a recommendation problem is defined as follows. Let U be a set of users, and let J be a set of all possible items that can be recommended, such as books, movies, or restaurants. Let g be a utility function that measures the usefulness of item j to user u, i.e., let g: $U \times J \rightarrow R$, where R is a totally ordered set (e.g., nonnegative integers or real numbers within a certain range). Then for each user, $u \in U$, we want to choose such items j that maximize the users' utility. More formally:

$$\forall u \in U, \quad j'_u = \arg\max_{j \in J} \; g(u, j) \tag{1.1}$$

There are mainly two types of ratings: *explicit* and *implicit*.

An **explicit rating** is a value given directly by a user. Explicit ratings are usually considered a reliable source of information: a user can tell exactly how she feels about a particular item. In contrast, **implicit ratings** are inferred through observation of user behaviors. For example, if a user spends a lot of time reading the description of an item online, we can infer that the user thinks that the item is valuable and assign a high rating for the item.

There has been much work both in industry and academia on developing new approaches to recommender systems over the last decade. The interest in this area is still high and growing because it constitutes a problem-rich research area with abundance of practical applications in many fields such as books [8, 9], e-commerce [10, 11], movies [12, 13], video [14, 15], music [16, 17], e-learning [18, 19], and so on.

It comes as no surprise that in the age of the Internet and streaming content, recommendation systems are also applied to academia. With access to huge collections of scholarly documents, it can be quite challenging to find relevant information and suggest appropriate researcher items (documents, publishers, peers, etc.) based on her need.

1.1.1 Academic Recommender Systems

Recent years have witnessed rapidly growing scholarly information due to the vast research works that are undertaken in academia and industry [20]. Using Microsoft Academic Search and Google Scholar, Williams et al. [21] estimated that there are at least 114 million English-language scholarly documents or their records accessible on the Web. According to them, new scholarly documents are generated at a rate of tens of thousands per day on different topics, with different authors, publication venues – a piece of evidence for the volume, variety, velocity of big scholarly data (BSD) [22]. This rapid rise of scholarly data with various entities, and relationships among the entities, which makes the research publication ecosystem a complex one, brings about new issues and challenges in the domain of data organization, access, and processing (Fig 1.1 is taken from Xia et al. [23]).

Academic recommender systems are used to recommend users different objects based on their likings by using various data analysis techniques to resolve the problem of information overload in academia [3,4]. Generally, academic recommender systems mostly provide

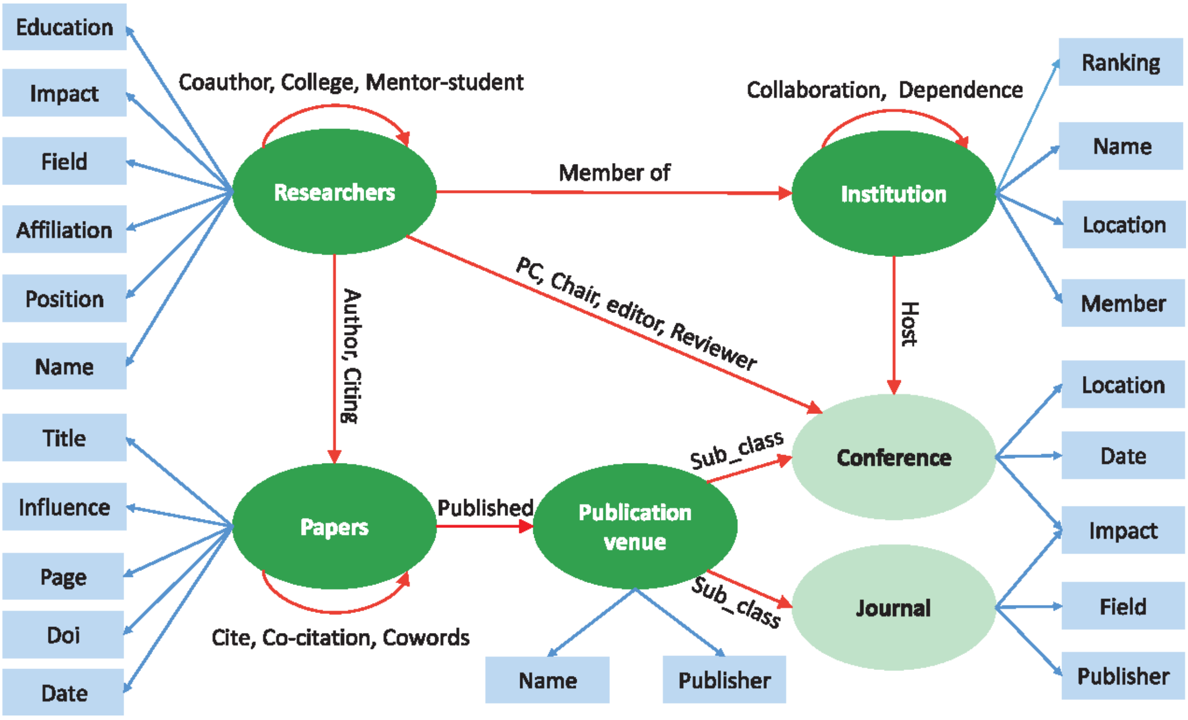

Figure 1.1: Major entities and their relationships in BSD

recommendations for collaborators [24, 25], papers [26, 27], citations [28, 29], and/or academic venues [30,31]. These systems have been useful to academicians as they objectively provide users with personalized information services [23]. Most of the academic recommender systems typically produce two kinds of output: prediction and recommendation. Prediction represents a guess: how a user (researcher) would rate an item (journal). This requires a numerical approach and, as such, the methods that provide the best predictions are statistical approaches. However, for a researcher, a list of top-N suggestions may be of more utility for her item of interest (papers, venues, researchers, etc.).

Example: Table 1.1 represents a simple journal recommender system composed of eight users (researchers) and seven items (journals). Each row is called a researcher rating profile and gives a researcher's rating for an arbitrary subset of the available journals. Researchers' preferences on journals are expressed using discrete numerical values from 1 to 5, where 5 indicates that the researcher likes the corresponding journal very much. Note that ratings are self-judged, and different users may use the same rating scale differently. To make recommendations to Researcher-4, we compute predictions for Researcher-4 on those journals that she has not rated yet (e.g., the entry related to Journal-4) and then recommend journals that have the highest predictions.

Table 1.1: A simple journal recommender system scenario. In order to recommend journals to Researcher-4, predictions of those missing entries corresponding to the row of Researcher-4 are computed first, then journals are recommended to Researcher-4 based on their predictions.

Researcher	**Journal-1**	**Journal-2**	**Journal-3**	**Journal-4**	**Journal-5**	**Journal-6**
Researcher-1	4		5	3		2
Researcher-2	3				2	3
Researcher-3			5	3	1	2
Researcher-4	1			?	4	
Researcher-5	4		2	1		4
Researcher-6	5		5	3	2	
Researcher-7				2		
Researcher-8	2		2	3		5

1.2 Classification of Academic Recommendation Algorithms

There exist several types of recommender systems. Adomavicius and Tuzhilin [4] authored a comprehensive review of recommender systems and suggested mainly following three types of recommender systems based on their working principles.

(i) Content-based Filtering (CBF)

(ii) Collaborative Filtering (CF)

(iii) Hybrid Recommendation

In addition, we also include social filtering [31].

(iv) Social Filtering (SF)

We briefly describe here the techniques in general. Academic recommender systems, being a subset of recommender systems, largely follow the same taxonomy.

1.2.1 Content-based Filtering (CBF)

This filtering technique recommends items for the user based on the description of previously evaluated items for the user. This type of recommender system recommends items that are similar to the one preferred by the user in the past. Items are defined by their associated features. This system recommends items based on the items' content rather than other users' ratings.

In academic recommender systems, both users and items can be either researchers, venues (journals/conferences), papers, topics, institutions, etc. Mainly content-based academic recommendation uses researchers' profiles, the content of their papers as well as that of the papers published at a specific venue [32]. Representative techniques are [33,34].

Example: As shown in Table 1.1, to recommend journals to Researcher-4, we need to compute the content similarity (e.g., topics of previously published papers in a journal) of Journal-1, and Journal-5 with all other available journals (Journal-2 to Journal-4, and Journal-6). Note that abstract, title, keywords, or full-text can be taken into consideration during the similarity.

1.2.2 Collaborative Filtering (CF)

This is one of the most commonly used filtering techniques. The user will be recommended items that people with similar tastes and preferences liked in the past. Generally, these systems create a user profile based on ratings of different objects and then compare these against a wider user group. The system recognizes similarities between users based on their ratings and then creates new recommendations based on the inter-user comparison.

These methods have the interesting property that no item descriptions are needed to provide recommendations since the methods merely exploit information about past ratings [7]. Compared to CBF approaches, CF also has the salient advantage that a user may benefit from other people's experience, thereby being exposed to potentially novel recommendations beyond her own experience [4]. Representative techniques are [35, 36]. Generally, CF approaches are mainly classified into two main categories:

- Model-based

- Memory-based

Model-based Approaches

This type of approaches build statistical models of user/item rating patterns to provide automatic rating predictions. Therefore it uses all available ratings to learn a model, which can then be used to predict the rating of any given item by any given user. These methods are mainly inspired by machine learning techniques such as artificial neural networks, bayesian networks, clustering, and latent factor models.

Memory-based Approaches

This type of approaches make rating predictions by combining ratings of selected users or items that are judged to be relevant. This type of approaches are characterized by their simplicity, since minimal or no learning phase in involved. These approaches are also called neighbors methods, which are further divided in two sub-categories.

- User-based
- Item-based

In the **user-based** methods, similarities between users in their consumption patterns are used to compute recommendations. The idea is that, for a given user, the preferences of similar users, the neighbors, can serve as recommendations.

In the **item-based** methods, similarities between items with common users are exploited. The idea is that items that are similar to those that the user has already rated or consumed are good candidates for recommendations.

Example: To make recommendations to Researcher-4, we need to identify top similar researchers in terms of the ratings they already provided to different journals (user-based approach in CF). Based on the collective preferences or taste information (ratings of journals) of those top similar researchers, the missing entries of all journals for Researcher-4 are predicted. The journals having high ratings are finally recommended to Researcher-4. Note that, item-based approach (journal-journal similarity) can also be adopted to identify the topmost similar journals to those journals which are already rated by Researcher-4. But here, Researcher-4 has rated only one journal (Journal-5), and therefore it should not be a good strategy to apply the item-based approach.

1.2.3 Hybrid Recommendation

All the above-mentioned types have their strengths and weaknesses. This type of recommendation uses a combination of two or more filtering techniques in order to avoid certain limitations of individual techniques. The main assumption of fusion-based approaches can be stated that “hybrid recommendation approaches can provide more accurate recommendation than a single approach and the disadvantages of one approach can be overcome by the other approach [27]”. These methods combine mainly collaborative filtering and content-based methods or any other type of filtering method. Representative techniques are [27, 37].

Burke et al. [7] presented a detailed taxonomy of hybrid recommender systems and existing classified approaches into many types. A few are listed below.

- **Cascade**: Here, the recommendation is performed as a sequential process in such a way that one recommender refines the recommendations of the other.

- **Feature Augmentation**: In this, the output from one recommender is used as an additional input feature for other recommenders.

- **Weighted**: Here, the scores provided by the recommenders are aggregated using a linear combination or a voting scheme.

- **Meta-level**: In this type, the model generated by one of the recommenders is used as the input for other recommenders.

- **Mixed**: Recommendations from several recommenders are available, and are presented together at the same time by means of certain ranking or combination strategy.

- **Switching**: It is a special case of the previous type considering binary weights, in such a way that one recommender is turned on and the others are turned off.

The use of a specific type of hybrid recommendation method depends on the application, but, more importantly, on the type of recommenders being combined. The hybrid techniques are primarily used to solve the new user problem [1]. This thesis mainly focuses on hybrid recommendations (e.g., fusion or ensemble-based approaches).

1.2.4 Social Filtering (SF)

This approach is based on a network representation of the input data and has also gained considerable attention in the recent past [38], mainly to get over the problems of the CF, and CBF approaches. This type of recommender system exploits social information, such as contacts and interactions between users. We shall henceforth refer to this type of recommendation approach as Social Filtering (SF) systems.

Recommendations by SF approaches have the nice property that they are generally easier to explain than user-based CF approaches. Recommendations through friends are indeed easy to interpret by end-users. A user is suggested items that her friends (e.g. in an online social network) liked in the past. This type of recommendation uses a social graph in order to make recommendations. Representative techniques are [31, 39].

Example: As shown in Table 1.1, to recommend journals to Researcher-4, we need to identify those journals which are popular among the friends of the target researcher (Researcher-4). Considering co-authorship as a friendship relation we, need to build a social graph among the co-authors of Researcher-4. An edge exists between two researchers if they co-author at least one paper [40]. The journal that has the highest count among the papers within n-hops from Researcher-4 is recommended. This approach also helps in dealing with new user problem, because it is not feasible to reliably compute their similarity with other users for lack of data.

1.3 Limitations of Academic Recommender Systems

Each type of recommendation approaches has its strengths and weaknesses. We have already provided brief overview of each technique, which are largely dependent on the source of information being used. Personalized recommendations require knowledge about the taste of user, typically in the form of preferences of the user for some items in the recommendation domain. Initially, when a user joins a recommender system, nothing or very little is known about what the user likes or is interested in. This problem is commonly known as cold start problem. There are mainly two kinds of cold-start problems: new item, and new user [3, 4]. In this section we discuss the limitations of each technique.

1.3.1 Problems with CBF Approaches

(i) **Limited Content Analysis**. Items to be recommended must have available data related to their features. This data is often unavailable or incomplete [41].

(ii) **New User (Cold Start)**. A user has to show some preference (ratings) for a sufficient number of items before a recommender can build a reliable content-based user profile. For a new user, without publication history, this is not available [3] (Researcher-7 in Table 1.1).

(iii) **Over-specialisation**. CBF recommenders are trained with the content features of the items. All the recommended items are similar to those already rated. Most of the time recommended items are wellknown to the user, poviding little (or no) novelty from the user's perspective [4].

(iv) **Portfolio Effect (Diversity)**. Often the recommended items are very similar among themselves resulting to a set of insufficiently diverse or too redundant item suggestions.

1.3.2 Problems of CF Approaches

(i) **Data Sparsity**. Observed user-item interactions is generally very small compared to the number of all user-item pairs.

(ii) **Grey Sheep**. Since CF approaches rely on the tastes of similar people to suggest new items, when a user has very specific or unusual preferences, she may not rceive useful suggestions.

(iii) **New User (Cold Start)**. A user has to show some preference (ratings) for a sufficient number of items before a recommender can build a reliable content-based user profile (Researcher-7 in Table 1.1).

(iv) **New Item (Cold Start)**. Until a new item has been rated by a substantial number of users, a recommender system may not be able to recommend it (Journal-2 in Table 1.1).

(v) **Stability**. It is defined as the consistency between the original recommendations and the recommendations using the combination of the historical data and some of the original recommendations [42]. The stability in the predictions and recommendations influences on the users' trust towards the recommender systems. The predictions provided by the recommender system should not change strongly over a short period of time.

(vi) **Scalability**. It is defined as the capability of a system to handle a growing amount of work, and its potential to be enlarged to accomodate that growth. Recommender systems should be in a position to provide recommendations in real time for millions of users in terms of data storage and algorithmic computation.

1.3.3 Problems with SF Approaches

(i) **Social Sparsity**. Every user has to be connected through at least one contact in the social network to be able to get recommendations.

(ii) **New Social Connection**. Recommendations may get biased if a user has a very small social network, up to the point that if she has only one connection, every social recommendation would be generated based on the activity of just one user.

(iii) **Soial Similarity**. Similarity based on social connections is used in SF recommenders. Two users socially connected may or may not have interests in common.

1.4 Motivation

Although there have been quite a few works on different academic recommendations, very little body of work exists in the literature on academic venue recommendations, albeit started quite early [36]. The researchers, in general, intend to publish in academic venues that acknowledge high-quality papers and participate in academic conferences or workshops that are relevant to their area of research [43]. Among various problems that researchers confront, an important task is to identify relevant publication venues. The task is nowadays being increasingly difficult due to the continuous increase in the number of research areas and dynamic change in the scope of journals [36].

1.4.1 Journal Recommendation

More and more collaborations are taking place among different disciplines of research leading to reduced compartmentalization at the coarse level but continuous increase in the number of venues in interdisciplinary areas [36]. For example, Microsoft Academic Graph (MAG) [1] dataset, a heterogeneous graph of information relating to a collection of scientific publication records and their relationship within that collection had 23,404 journals and 1,283 conferences in 2016 [44] and is increased to 48,668 journals and 4,344 conferences as of 2019 [2][blue].

As research horizon expands, researchers find it challenging to remain up to date with new findings, even within their own disciplines [45]. With the passage of time, researchers' own interests also expand, evolve, or adapt in rapidly changing subject areas needing for information on appropriate venues in the changed scenario [46]. Moreover, increase in interdisciplinary research areas poses great challenges to research institutes and their libraries as they strive to understand information-seeking behaviors and dynamic information needs of the users [36]. Information specialists need prompt and seamless information on researchers' reading priorities in order to make decisions on venue subscriptions instead of relying only on the venues' impact factor or on users' explicit requests.

On the other hand, researchers also need to know about new venues to remain updated. They usually get updates from colleagues/supervisors, friends, internet, and books but often the information is not sufficiently comprehensive and/or appropriate as their research actually demands. The researchers, therefore, sometimes end up approaching inappropriate venues resulting rejections, delays in publication and/or compromise in the quality of the publication. Venue recommendation for either journals or conferences, in particular, has, therefore, become an important area of research in recent times [47]. Out of many reasons for its increasing importance, some are given below as emerging scenarios [31].

(a) A researcher from industry has made a breakthrough in her research area. To collaborate with her peers from academia, she may want to find a suitable academic venue (conference) that she is not very aware of.

[1] http://research.microsoft.com/en-us/projects/mag/

[2] https://academic.microsoft.com (as of 08.01.2019)

(b) A junior researcher, i.e., a researcher who is at the initial stage of her research and has no or very few publications, intends to extend her research area. But lack of knowledge on appropriate academic venues becomes a challenge for her to explore newer areas.

(c) A veteran researcher knows her research area very well, but when she ventures into a new field or works in an interdisciplinary area, she may look for cross-field venue recommendations.

(d) A journal may merge with some other related journals with modified scopes and objectives. The researchers may not be aware of such developments.

In order to recommend relevant venues of high quality, we need to focus from the perspective of a researcher's needs and development of the particular research area in question on the following issues.

(i) What are the most relevant venues of publications for a researcher in question?

(ii) How can a researcher find high-quality venues?

(iii) What are the most suitable conferences/workshops a researcher should participate in, for a given area?

Most of the existing techniques depend on co-authors' past publications and/or ratings of the venues provided by other researchers for such venue recommendations. A few approaches also use random walk model, topic-based similarity for the same.

Problems with CBF Approaches

(i) CBF approaches suffer from limited content analysis, which can significantly reduce the quality of recommendation [48]. Most of the time, they require the full text of the paper and, thus, are not usable at the early stage of paper-writing [49]. Usually, the abstract is not sufficient to extract the necessary reliable and relevant information.

(ii) New venues are less likely to be recommended as the models prefer venues with a high number of papers published therein.

(iii) The models provide a poor recommendation to a new researcher who lacks publication records.

(iv) The recommendations are heavily biased towards the past area of research of a researcher and, therefore, not suitable when one changes her area of interest or works in an interdisciplinary field.

Problems with CF Approaches

(i) CF approaches are less effective when there are not enough ratings present in the researcher-venue matrix. The recommendations may not be useful in case of a new researcher who lacks publication history.

(ii) The techniques are not likely to recommend a new venue or a less popular venue as the venue lacks in publication statistics. Therefore, some relevant venues may be missed because of the new entrance or lack of publication statistics.

(iii) Computational cost is high because of an extensive number of articles, venues, and researchers involved are taken into consideration during processing, and thus, scalability is a challenge.

(iv) Researcher-venue matrix that is at the core of the techniques is exceptionally sparse as most of the researchers publish and cite very few articles and are involved with a limited number of academic venues.

Problems with SF Approaches

(i) Irrespective of actual content, each paper authored by the same set of authors will receive the same recommendation.

(ii) Recommendations are very poor for a new researcher who does not have any past publication records.

(iii) It cannot recommend a new venue as the model is based on the publication history of venues.

(iv) Venues with less popularity among the co-authors of a given author are seldom recommended, although content-wise, they may be appropriate.

Problems with Freely Available Online Services

Elsevier Journal Finder (EJF) and Springer Journal Suggester (SJS) are two popular freely available online journal recommender systems based on CBF that came of late. Some limitations of these systems are as follows.

(i) Both the systems restrict suggestions to their own publications only (either Elsevier or Springer) - which is a severe limitation from users' point of view. Suggested journals are often found not matching with the topic of the paper.

(ii) EJF suggests a maximum of ten journals, whereas SJS, provides a maximum of twenty. Sometimes they provide less than ten journals due to the constraints of unavailability of similar contextual papers.

(iii) The ranking algorithm used in both the systems only works well if there are enough sample papers (at least 100) in each journal. However, for some new journals, there may not be that many published papers, and the existing systems fail to recommend the relevant journals [50].

(iv) There is no provision for the recommendation of high-quality conferences or workshops (in few domains, conferences/workshops have high visibility, e.g., information retrieval, machine learning, etc.).

1.4.2 Collaborator Recommendation

Studies suggest that the more the collaboration, the higher is the popularity in the research community. A large number of collaborations are taking place among researchers and productive researchers to be more collaborative [51]. There are various motivations for a researcher to search for collaborators.

(a) Generally, researchers collaborate to improve their research exposure and profile. It also helps them to get involved in more influential research and achieve higher productivity.

(b) Junior researchers who are not acquainted with new research areas and look for collaborators to exchange ideas, acquire expertise and resources for high-quality research.

(c) A veteran researcher may know her research domain exceptionally well; however, when she ventures into another field or works in an interdisciplinary area, she needs to look for potential collaborators.

Of late, tremendous growth in big scholarly data has made it difficult to search for and find out appropriate information [52]. For example, DBLP [3] dataset, a collection of scientific publication records and their relationship within that collection has 4,419,797 publications from more than 2,205,561 researchers in 9,585 computer science conferences [4] and more than 4152 [5] journals [33]. This phenomenon has created huge opportunities for researchers in the area of academic collaboration [53]. But on the other hand, with the growing number of research areas and dynamic changes in researchers' interests, finding suitable collaborators is getting increasingly challenging. To recommend appropriate collaborators, we need to focus on the following research issues.

(i) What are the most academic influence-aware features in a big-scholarly network which can affect the co-authorship explicitly or implicitly of target researcher?

(ii) How to capture the rapidly changing nature of research domains and research interest of individual researchers?

Mainly, there are two types of collaboration recommendation as described below [39, 54].

(a) Recommendation of most potential collaborators (MPCs) who have never worked with a target researcher (i.e., to build new collaborations).

(b) To recommend the most valuable collaborators (MVCs) among researchers who have earlier collaborated with the target researcher before (to reinforce old collaborations).

Numerous collaborator recommendation approaches have been proposed in the past. Initially keyword-based [55], or topic-based [56] techniques recommended collaborators based on similarity of research area. Most of them failed to capture the essence of researchers'

[3] http://dblp.uni-trier.de/db/

[4] http://dblp.uni-trier.de/db/conf/

[5] http://dblp.uni-trier.de/db/journals/

collaboration patterns or social proximity leading to inappropriate collaborator recommendations. In content-based approaches, authors' co-authorship profile is not taken into account.

To incorporate social proximity, a few research works considered network structure [57]. Network analysis based techniques do not consider content-based similarity and mainly focus on a single type of relation (co-author, citation, or co-citation). Also, for a prolific researcher, research interests may change with time and diversify. Present techniques are yet to capture the dynamic research interest of a researcher while identifying the current active researchers in a given area.

Recently few papers used hybrid approaches consisting of both content and network structure into account [54]. Although the hybrid approaches capture the connection and compatibility of collaboration among researchers, there remain a lot of implicit factors like physical distance between their affiliations, age or pedigree, personality that affect collaboration in real life. We do not explicitly consider these features and/or feed enough data to learn their weights in machine learning-based models.

Most of the existing approaches mainly focus on recommending possible collaborators without highlighting the most influential collaborators (MICs). Moreover, various meta-path features such as citations, co-citations, and academic activity among researchers are not yet adequately exploited in collaborator recommendation. The advantage of deep learning is that it uses unsupervised or semi-supervised feature learning and efficient feature extraction algorithms instead of manually acquiring features, and we presume these implicit factors can be taken into account in deep learning-based methods [58].

1.5 Research Goals

The main objective of the research presented here is to adopt various techniques to develop an academic recommender system[6] with improved performance. During the thesis work, we focus on journal and collaborator recommendations. However, as explained, there are several issues with the existing approaches. We, therefore, attempted to find answers to the following research questions (RQs) as part of our study.

[6] Academic venue recommendation is also known as scholarly venue recommendation. In the present thesis we use terms scholarly and academic interchangeably. The same applies to terms venue and journal.

RQ1: How to handle cold-start issues in journal recommendation?

RQ2: How to address data sparsity, and diversity issues in journal recommendation?

RQ3: How to improve relevance and stability in journal recommendation?

RQ4: How the overall quality (popularity) of recommendation is enhanced in journal recommender system?

RQ5: How is to improve overall relevance and also to handle cold-start issue in collaborator recommendation?

1.6 Contributions

This thesis is devoted to the development of an effective journal and collaborator recommender systems. To find answers to the research questions (RQs) as stated in Sec 1.5, we, therefore, attempt to investigate the problem of venue recommendation mainly from the following two opposite aspects: i) when we have a *large heterogeneous* bibliographic networked data comprising diverse fields of research and are *sparse* in nature ii) when the dataset is comparatively *smaller* and *densely connected* in a *narrow* and/or *focused* field of study.

The main contributions of this thesis are related to the design, implementation, and evaluation of the performance of the recommender system, where we have addressed several existing issues using novel models and methods. We have proposed three novel models for journal recommendation and one model for an effective collaborator recommendation, as given below.

1.6.1 DISCOVER: A Journal Recommendation System

We propose a Diversified yet Integrated Social network analysis and COntextual similarity-based scholarly VEenue Recommender system (DISCOVER). Key contributions of this work are the following.

- To deal with the **cold-start** issues like new researchers and new venues in venue recommendation, an integrated approach of social network analysis, contextual similarity, citation, and co-citation analysis are taken into consideration. DISCOVER works irrespective of researchers' past publication records and co-authorship network, rather focuses only on the work in hand. New venues with no citation records available are also considered for abstract similarity and provided equal opportunities in the recommendation.

- **Data sparsity**[7] is a major issue in an academic venue recommender systems. Handling any bibliographic network[8] with few inter-node connections is a severe challenge. We 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.

- To address the issue of **stability**[9], a topic modeling based contextual similarity is incorporated into the proposed system. We develop a content-aware system based on the title, keywords, and abstract similarities with a given seed paper. The abstract similarity is computed using LDA, Okapi BM25+, and non-negative matrix factorization (NMF) techniques. Later on, score-based fusion technique such as CombMNZ is applied to fuse the similarity score of both LDA and NMF in order to maintain the stability of contextual similarity.

- In venue recommender systems, **scalability**[10] is one of the major issues. In the proposed system, the most important papers in a citation network are identified through a number of stages. But at the very first step, we very selectively choose the potential candidates for the subsequent steps. This keyword-based search strategy is computationally linear in database size and does not increase the output size, even with an increase in the input size. Therefore, there is no considerable increase in overall computation when input data grows.

[7] Sparsity, in this context, denotes the number of empty, or zero-value entries (often useless) in a given researcher-venue matrix. The less the amount of sparsity, the better is the space-time utilization.

[8] Any bibliographic network is a huge graph with relatively few edges where the number of edges is close to the minimal number of edges.

[9] Stability denotes the resilience to change in ranked recommendations with the introduction of new papers

[10] Scalability denotes the ability of a system to accommodate new researchers and/or papers.

- Extensive experiments were conducted using MAG dataset on DISCOVER. The proposed system is seen to outperform several other state-of-the-art venue recommendation systems with substantial improvements in precision, nDCG, accuracy, MRR, and average venue-quality (ave-quality).

1.6.2 CNAVER: A Journal Recommendation system

In DISCOVER we combined both social network analysis and contextual similarity in a sequential manner (cascading technique). This model can address the problem of cold start for new researchers reasonably well. However, cold start issues for new venues, sparsity, diversity, and stability issues are not adequately addressed. To bridge this gap, we propose CNAVER: An integrated framework of Content-based features and Network-based model for Academic VEnue Recommender system. CNAVER is built on two major components that contribute in parallel, but finally, their contributions are fused together to present a coherent venue recommendation. One is the paper-paper peer network (PPPN) model and the other venue-venue peer network (VVPN) model. While PPPN explores the interaction among papers towards venue recommendation, VVPN actually studies it among publication venues. Key contributions of this work are the followings:

- To deal with **cold-start** issue like a new venue, mainly VVPN model is proposed. In addition to abstract similarity (paper with no citations), meta-paths features (common paper, author, citation, co-citation, terms) 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. Therefore chances of inclusion of new venues in the recommendation lists are relatively higher.

- To address **data sparsity**, we adopted two-stage filtering techniques such as centrality measures based citation analysis and contextual similarity such as 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.

- To resolve the issue of **diversity**[11] a fusion model incorporating both PPPN and VVPN model are proposed. Specifically, age-discounted(age-discounted) based Venue2Vec,

[11] Diversity means how many different venues are recommended.

meta-path features, and biased random walk are incorporated into the model to recommend venues from diverse publishers.

- To address the issue of **stability**, a fusion model CNAVER incorporating both PPPN and VVPN models are proposed. Any network-based approach is known to cause instability in the ranks with time as the introduction of new nodes or edges to change the topology and thereby change recommendations [42]. We therefore also took into account content-based approaches at several stages within both PPPN and VVPN pipeline.

- Comprehensive experiments were conducted using a real-world dataset, i.e., DBLP, to evaluate the performance of the proposed system CNAVER. The proposed system outperforms several other state-of-the-art venue recommendation models with substantial improvements in precision@k, nDCG@k, accuracy, MRR, average venue-quality (ave-quality), and diversity.

1.6.3 DeepRec: A Journal Recommendation System

Both DISCOVER and CNAVER approaches addresses cold start issues, diversity, and scalability issues to some great extent. However, relevance (accuracy in recommending relevant venues), stability, and sparsity issues are not adequately addressed. To bridge these gaps, we propose DeepRec: A stacked generalized ensemble learning-based scholarly venue recommender system. It is mainly designed based on Convolution Neural Network (CNN), and Long Short-Term Memory (LSTM) techniques. It also leverages the demerits of CNN and LSTM by ensembling them to provide an integrated framework to recommend the most suitable venues. Key contributions of this work are the followings:

- To enhance the recommendation quality in terms of **relevance**[12] we extracted latent features from abstract and title with CNN and LSTM model and combined them into the proposed model DeepRec.

- To address **data sparsity** issue, we transformed 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

[12] Here relevance means semantic closeness between our need and the recommendations. Here, we measured the quality of relevance in terms of accuracy, precision, nDCG, and MRR

aspects of high dimensional and sparse input. Due to the capacity of extracting hidden contextual features that might be relevant deep learning approach is highly preferred across domains.

- To resolve the issue of **diversity**, a stacking ensemble (stacked generalization) model is used for training LSTM and CNN based architecture together in our model. In this stacking, the algorithm takes the output of both CNN and LSTM models as inputs and combine them by training a meta-model to output predictions based on multiple predictions returned by both CNN and LSTM models.

- The stacked generalized ensemble learning model also helps to maintain a **stability** by capturing the relevance of papers. Here for contextual similarity, both abstract and title are considered.

- Comprehensive experiments were conducted using a real-world dataset, i.e., DBLP, to evaluate the performance of the proposed system DeepRec. The system outperforms several state-of-the-art venue recommendation models by substantial margins in precision@k, nDCG@k, accuracy, MRR, average venue-quality (ave-quality), diversity, and stability.

1.6.4 DRACoR: A Collaborator Recommendation System

Collaborator recommendation is an important requirement for any researcher. While for a new researcher, it is of utmost value, it is often necessary for an old researcher as well when she ventures into a new field. We propose a multi-level fusion based system DRACoR (Deep learning and Random walk based Academic Collaborator Recommendation). It fuses Meta-path aggregated Random walk based Collaborator Recommendation (MRCR) that finds out MPCs with Deep learning-Boosted Collaborator Recommendation (DBCR) models that find MVCs so that their combination (MICs) can be recommended. The key contributions of this work are sixfold, as described below.

- DRACoR works irrespective of a researcher's past publication records and is entirely based on her current works. Isolated researchers, researchers with less number of co-authors, or researchers with fewer publication records also get an equal chance

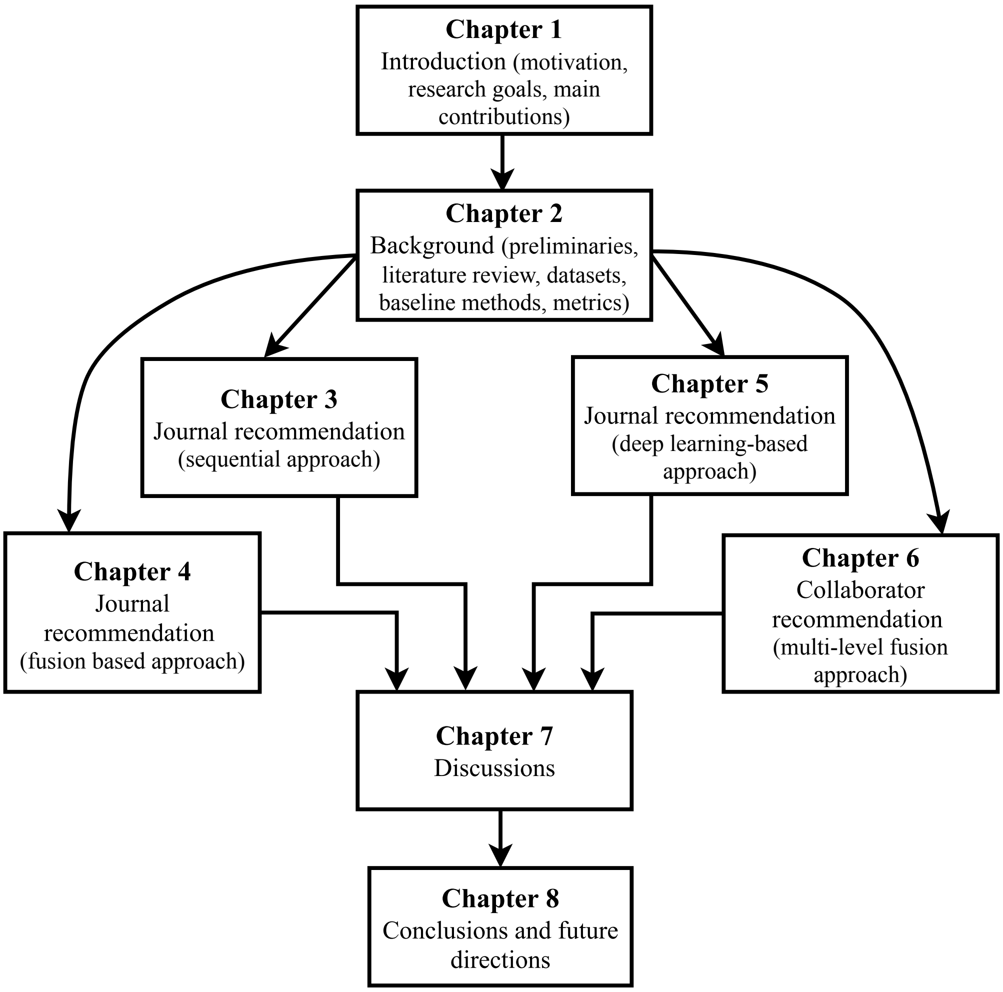
Chapter 1
Introduction (motivation, research goals, main contributions)
Chapter 2
Background (preliminaries, literature review, datasets, baseline methods, metrics)
Chapter 3
Journal recommendation (sequential approach)
Chapter 5
Journal recommendation (deep learning-based approach)
Chapter 4
Journal recommendation (fusion based approach)
Chapter 6
Collaborator recommendation (multi-level fusion approach)
Chapter 7
Discussions
Chapter 8
Conclusions and future directions

Figure 1.2: Illustration of thesis structure

for inclusion in the final recommendation. Hence it deals with **cold-start**[13] for a new researcher very well.

- To capture the shift in an author's research interest with time, a time-aware 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 improve the **relevance**[14] of the recommendations.

- An improved random walk with restart (RWR) based recommender model MRCR is proposed that aggregates various meta-path features along with H-index based level similarity and percentage of collaboration to suggest MPCs.

- To identify the association and collaboration compatibility among researchers, we adopt

[13] Cold start issues mainly indicate the new researchers in academia.

[14] Here we measured the quality of relevance in terms of F1, nDCG and MRR

a deep learning-based collaborator recommendation model DBCR considering Word2Vec and Long Short Term Memory (LSTM) techniques. The technique provides MVCs.

- Comprehensive experiments are conducted using two real-world datasets DBLP and hep-th, to evaluate the performance of the proposed system DRACoR. Our model outperforms several state-of-the-art collaborator recommendation models with respect to precision, recall, F1-score, MRR, and nDCG.

1.7 Structure of the Thesis

The thesis consists of seven different chapters. Fig. 1.2 provides an overview of the thesis and interconenction among chapters.

Chapter 1 starts with a brief introduction of recommender systems and their types. The motivation, research goals, contributions of our work are also discussed.

Chapter 2 sets the background by providing an overview of the state-of-the-art in journal and collaborator recommender systems. Necessary theoretical background including the principal evaluation metrics, and methodologies used are discussed.

Chapter 3 proposes DISCOVER: our first model in journal recommendation. The functional architecture, a pipeline of different similarity calculation techniques along with performance evaluations are discussed.

Chapter 4 presents our second journal recommender system CNAVER: a parallel pipeline based model that takes into account both paper-paper and venue-venue relationship. System design, experimental set-up, comparison with different state-of-the-art techniques are described in detail.

Chapter 5 presents a deep learning based journal recommendation model DeepRec. How deep learning architecture can capture more subtle and finer features hidden within the context is described along with the performance scores.

Chapter 6 describes our attempt to explore another dimension of academic recommendation: collaborator recommendation. We propose a multi-level fusion model DRACoR here. System design, experiments, and results are reported.

Chapter 7 summarize the entire thesis work highlighting the key insights obtained along with discussions and limitations.

Chapter 8 finally concludes with directions of future work.