Chapter 4

CNAVER: A Fusion-based Scholarly Venue Recommender System

> *"Perhaps there is no single variable which so thoroughly influences interpersonal and group behavior as does trust."*
>
> -Golembiewski and McConkie (1975)

4.1 Introduction

Academic venue recommendation is an emerging field due to rapid increase in the number of scholarly venues coupled with exponential growth in interdisciplinary research areas and research collaborations.

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.

In the last chapter we introduced DISCOVER that provides journal recommendations based on keywords, title, and abstract of a seed paper. The work had a sequential but integrated approach incorporating social network analysis, citation and co-citation analysis, contextual similarity based on topic modeling and main path analysis of a bibliographic citation network. We considered there MAG dataset - a heterogeneous graph comprising over 120 million publication entities and related authors, institutions, venues, and fields of study, as it fitted with our first objective. We evaluated the effectiveness of DISCOVER over two diverse fields of study, such as Computer Science (Engineering) and Biology (Science). The MAG contains 528,682,289 internal citations (citations between the papers in the graph) with each paper in the graph being cited on an average 4.17 times.

In this chapter, we revisit the problem of venue recommendation from the second perspective with a more focused and denser bibliographic network. The MAG dataset, however, does not seem appropriate for this objective. We find DBLP where citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. The tenth version of DBLP contains 3,079,007 papers with 25,166,994 citations, which means each paper in the graph is cited an average 8.17 times. We introduce CNAVER: an integrated framework of Content-based features and Network-based model for Academic VEnue Recommender system that pays special attention to the higher level of connections among papers. Besides, some other issues of venue recommendations are also taken into account. Although the cold start problem of new researchers was reasonably well addressed by DISCOVER, cold start for new venues, sparsity, diversity, and problems of stability were not adequately addressed. Also, DISCOVER clearly did not capture the variation of the venue scope over time (when the journal scope is modified). In CNAVER, these issues are explored in greater detail and depth and addressed.

CNAVER is based on two key components that contribute in parallel, but their contributions are eventually combined to present a coherent venue recommendation. One model is the paper-paper peer network (PPPN) and the other model is the venue-venue peer network (VVPN). While PPPN explores the interaction among papers towards venue recommendation, VVPN actually studies it among publication venues. Meta-path features (common paper, author, citation, co-citation, words, or topics) are considered to provide weights between venues in addition to abstract similarity (paper without citations). Venues with fewer papers and citations also have certain weights in the venue-venue graph (VVG) for ties with other relevant venues. Therefore chances of new venues being included in the recommendation lists are also relatively higher.

CNAVER only requires the title and abstract of a paper to provide venue recommendations, thus assisting researchers even at the earliest stage of paper writing. Another salient point is that, in CNAVER, recommendations are independent of keywords. Within

Table 4.1: Type of vertices used in HIN

No.	Vertices Type
1	P_main={set of papers that belonging to a particular venue}
2	P_ref={set of papers that cited by a P_main paper}
3	P_cite= {set of papers that cites a P_main paper}
4	$A(author)$= {author of any type of paper (P_main, P_cite, P_ref)}
5	$T(term)$={term appearing in titles or abstracts of a P_main paper}
6	$V(venue)$ = {set of any venue where P_main type papers published}

PPPN, we adopt two-stage filtering techniques such as centrality measures based on citation analysis and contextual similarity like LDA on abstract and Doc2Vec on the title. This filtering technique considers all parameters of significance and importance to reduce the bibliographic network size and also to increase the relatedness among papers. An age-discounted weighting method is proposed in CNAVER to capture the change in a venue's scope over time. The topics from recently published papers are prioritized, while topics from older publications are penalized in this method. In a better way than DISCOVER, CNAVER tackles cold start issues such as the presence of an inexperienced researcher and a novel venue, along with the issues of data sparsity and diversity.

4.2 Problem Description

Definition 1 *Heterogeneous Information Network (HIN) [173, 174]. It is defined as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ with a node type mapping function $\delta : \mathcal{N} \rightarrow \mathcal{W}$ and a link type mapping function $\mu : \mathcal{L} \rightarrow \mathcal{Y}$. Each node $n \in \mathcal{N}$ belongs to one particular node type in the node type set $\mathcal{W}$: $\delta(n) \in \mathcal{W}$, and each link $l \in \mathcal{L}$ belongs to a particular link type in the link type set $\mathcal{Y}$: $\mu(l) \in \mathcal{Y}$. Here both type of nodes $\mathcal{W}$ and type of edges $\mathcal{Y}$ depend on the domain in question. Note that both $|\mathcal{W}| > 1$ and $|\mathcal{Y}| > 1$.*

Due to the complexity of HIN and also to understand the node types and link types clearly in the network, meta level (schema-level) description is provided. So the concept of network schema is proposed to describe the meta structure of a network [175].

***Definition 2** (HIN Schema) [173]. The HIN schema denoted as $\mathcal{S} = (\mathcal{W}, \mathcal{Y})$, is a meta template for an information network $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ with a node type mapping function $\delta : \mathcal{N} \rightarrow \mathcal{W}$ and a link type mapping function $\mu : \mathcal{L} \rightarrow \mathcal{Y}$, which is a directed graph defined* *over node types $\mathcal{W}$ and type of edges $\mathcal{Y}$.*

Definition 3 *Scholarly Information Network (SIN) [176]. SIN graph is an instance of HIN. Here both type of nodes $\mathcal{W}$ and type of edges $\mathcal{Y}$ are related to a scholarly network (academia).*

Example. In a SIN, $\mathcal{W}$ can be either authors, papers, publication venues, terms etc. Similarly, type of links $\mathcal{Y}$ can be any type of relations between a pair of members in $\mathcal{W}$ like paper-paper, author-author, paper-author, paper-venue, author-venue, paper-terms, author-terms, venue-terms etc. In Fig. 4.1, we show graphical representation of a SIN with all of its vertices types and their relationship. Here we have six type of nodes $\mathcal{W}$, such that $\mathcal{W} = P\_main \cup P\_ref \cup P\_cite \cup A \cup T \cup V$ and seven type of links $\mathcal{Y}$ (Table 4.2). The meaning of each type of nodes is defined in Table 4.1.

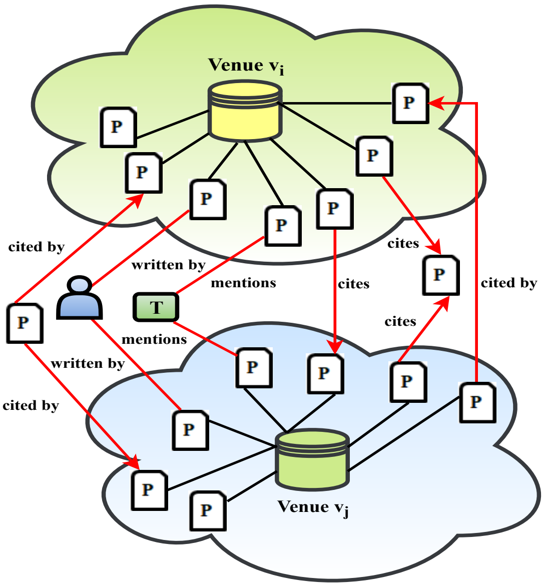


Figure 4.1: Graphical representation of SIN graph

Definition 4 *Venue-Venue Graph (VVG). Let $G' = (V', E')$ be the newly generated venue-venue graph (VVG) from HIN based on the similarity score of abstract and title. $V' = \{v_1, v_2, \cdots, v_l\}$. Each edge $e = (v_i, v_j) \in E'$ represents a currently similar research scope of v_i with v_j based on their past publications. An edge $e = (v_i, v_j) \in E'$ exists if the similarity score between venues v_i and v_j is greater than average similarity score. We weight the edges of the network VVG using content similarity (linear combination of abstract and title) in order to provide a single score as explained in Sec.* 4.6.2.

Definition 5 *Meta-path [177]. A meta-path M is a path defined on the HIN graph. It joins two or more vertices using one or more edges such that $M = n_1 \overset{l_1}{\rightarrow} n_2 \overset{l_2}{\rightarrow} ... \overset{l_t}{\rightarrow} n_{t+1}$, where the starting and ending vertices are of same vertex type P_main, $\delta(n_1) = \delta(n_{t+1})$ and both belong to P_main, P_main $\in W$, $\mu(l_1, l_2, ..., l_t) \in Y$.*

Example. In Fig. 4.1, There will be a meta path between venue v_i and venue v_j via the meta path Venue $v_i \overset{publish}{\rightarrow}$ *P_main* $\overset{citedby}{\rightarrow}$ *P_cite* $\overset{cites}{\rightarrow}$ *P_main* $\overset{publishedby}{\rightarrow}$ Venue v_j.

Table 4.2: Type of edges used in HIN

No.	Edges Type
1	$n_1 \xrightarrow{written\_by} n_2 : \delta(n_1) \in \{P\_main, P\_ref, P\_cite\}, \quad \delta(n_2) = A, \quad n_1, n_2 \in N$
2	$n_1 \xrightarrow{contains} n_2 : \delta(n_1) \in \{P\_main, P\_ref, P\_cite\}, \quad \delta(n_2) = T, \quad n_1, n_2 \in N$
3	$n_1 \xrightarrow{cites} n_2 : \delta(n_1) \in P\_main, \quad \delta(n_2) = P\_ref, \quad n_1, n_2 \in N$
4	$n_1 \xrightarrow{cited\_by} n_2 : \delta(n_1) \in P\_main, \quad \delta(n_2) = P\_cite, \quad n_1, n_2 \in N$
5	$n_1 \xrightarrow{cites} n_2 : \delta(n_1) \in P\_main, \quad \delta(n_2) = P\_main, \quad n_1, n_2 \in N$
6	$n_1 \xrightarrow{cited\_by} n_2 : \delta(n_1) \in P\_main, \quad \delta(n_2) = P\_main, \quad n_1, n_2 \in N$

Definition 6 *Random Walk [178]. A random walk is defined as a node sequence $S_r = \{v_1, v_2, v_3, \cdots, v_l\}$ wherein the i-th node v_i in the walk is randomly selected from the neighbors of its predecessor v_{i-1}.*

Definition 7 *Citation Network. Let $G = (V, E)$ be the citation graph, with n papers. $V = \{p_1, p_2, ..., p_n\}$. In G, each directed edge $e = (p_i, p_j) \in E$ represents a citation from p_i to p_j.*

Definition 8 *Venue Recommendation. Let each paper p_i published in a particular venue v_i. So now we have, $B = \{v_1, v_2, ..., v_l\}$ be a predefined set of publication venues. Given a input paper (seed paper) p_m, the venue recommendation task is to recommend a list of suitable publication venues (v_1, v_2 ,..., v_N) related to the seed paper p_m, where the list is ordered from the most relevant to the least relevant.*

4.3 Architecture of CNAVER

We present an overall architecture of the proposed framework alongside its operational strategies. As the bibliographic dataset is exceptionally massive in size (2,408,010 papers),

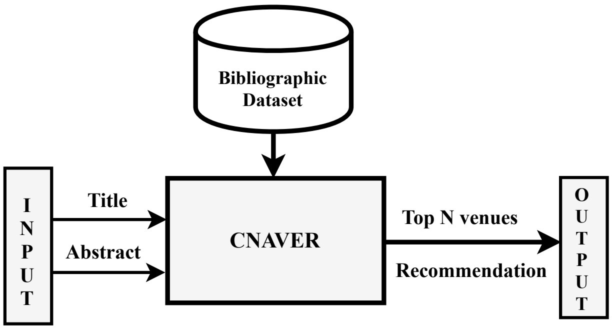


Figure 4.2: The basic block diagram of CNAVER

if we attempt to recognize the topmost similar papers for each seed paper by looking at contextual similarity against the entire dataset, the overall computational overhead will be high. The block diagram is depicted in Fig. 4.2

4.3.1 Framework of CNAVER

We propose a system comprised of two blocks: Block-I and Block-II as depicted in Fig. 4.3. To reduce computational overhead and to make it independent and autonomous of seed papers, particularly Block-I, is developed once for the whole citation network. Later on, we will utilize the seed paper input to interact with Block-II to extract meaningful recommendations from both the PPPN model and VVPN model. Four primary layers are portrayed as given underneath:

(i) Data Preprocessing and Centrality Calculation (**Layer-1**): This layer aims to structure the dataset into a formal model for processing. Mainly it is used for faster extraction of relevant papers and the importance of each candidate papers for further use (**Block I**).

(ii) Contextual Similarity Calculation (**Layer-2**): This layer can also be called the feature extraction layer and is mainly introduced to extract required contextual features needed to compute Paper2Vec in PPPN model and Venue2Vec in VVPN model. It is also used to filter only potentially useful papers from Set-II, based on content similarity (**Block I**).

(iii) Peer-peer Network Model (**Layer-3**): This layer uses a peer-peer network to process the data and to make a recommendation. The objective of this layer is to reduce computational overhead and to make it independent of seed papers (**Block I**).

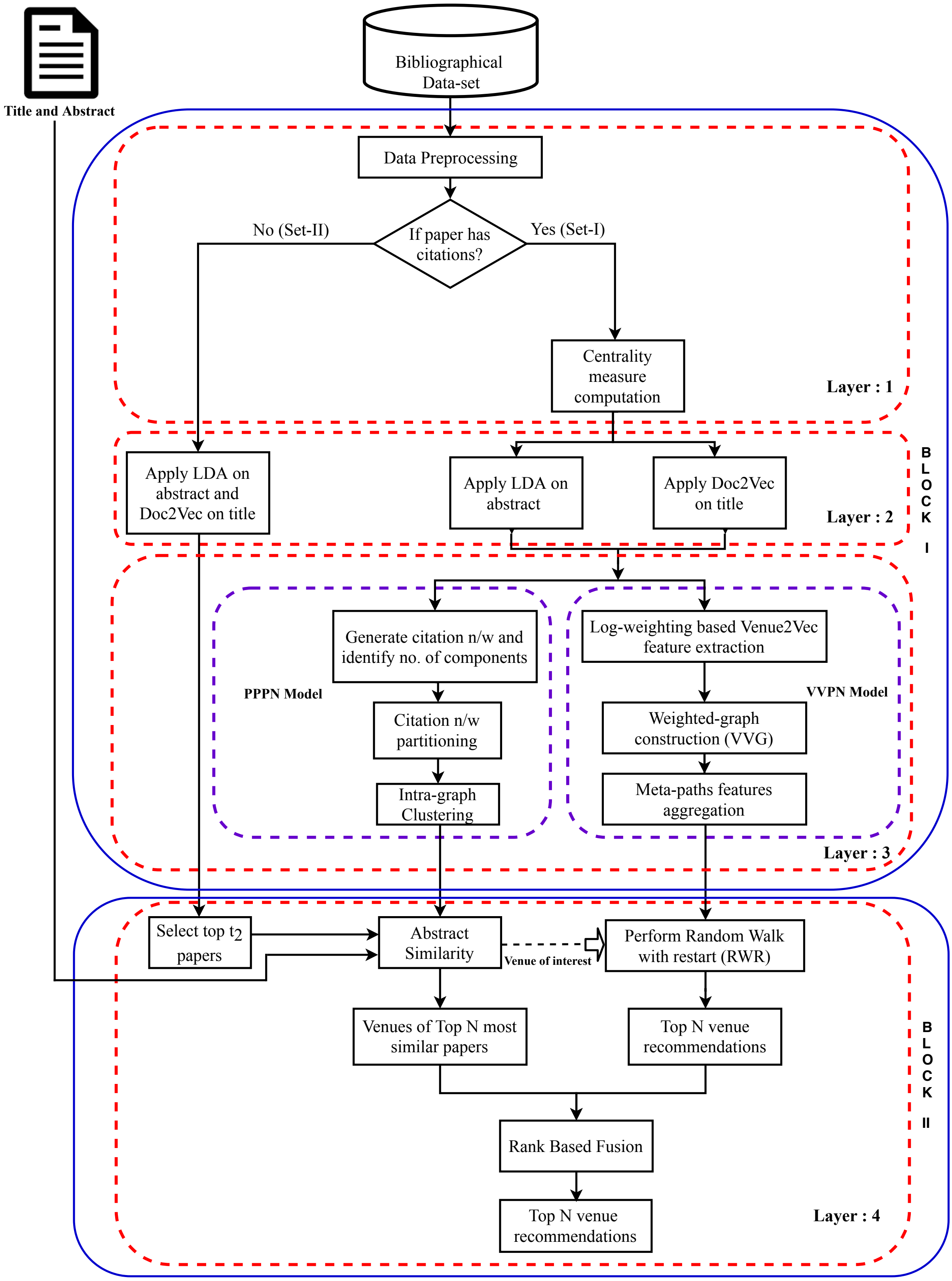

Title and Abstract
Bibliographical Data-set
Data Preprocessing
If paper has citations?
No (Set-II)
Yes (Set-I)
Centrality measure computation
Layer : 1
Apply LDA on abstract and Doc2Vec on title
Apply LDA on abstract
Apply Doc2Vec on title
Layer : 2
BLOCK I
Generate citation n/w and identify no. of components
PPPN Model
Citation n/w partitioning
Intra-graph Clustering
Log-weighting based Venue2Vec feature extraction
VVPN Model
Weighted-graph construction (VVG)
Meta-paths features aggregation
Layer : 3
Select top t_2 papers
Abstract Similarity
Venue of interest
Perform Random Walk with restart (RWR)
Venues of Top N most similar papers
Top N venue recommendations
Rank Based Fusion
Top N venue recommendations
Layer : 4
BLOCK II

Figure 4.3: Architecture of CNAVER

This layer comprises of two distinct models, namely:

(a) PPPN Model: The main objective is to capture the strength of individual papers and their citation relationship with other papers in a citation network to obtain relevant venues to the seed paper.

(b) VVPN Model: The main idea behind this model is to capture the similarity (indirect relationship among venues via meta-path analysis) among venues in a heterogeneous bibliographic network to obtain relevant venues to the seed paper.

(iv) Fusion Model (**Layer-4**): To provide a diversified personalized recommendation, the PPPN, and VVPN models are utilized to make predictions individually and later on a fusion model firstly is applied to integrate the strengths of both the models and to reduce their weaknesses (**Block II**).

4.4 Data Preprocessing and Centrality Calculation (Layer-1)

Initially, we filter duplicate papers, papers with missing fields and also inconsistent entries from the dataset. We also ignore non-textual content from the abstracts of the papers. The detailed statistics of the DBLP data collection are described in Sec. 2.7.2. All such papers are checked for their references section. We separately treat the papers having references or not. The set of papers where references are available are called Set-I and the set of papers without references are called Set-II.

We generate a citation network only with the Set-I papers. Among the centrality measures, we use degree, betweenness and closeness measures (Sec. 2.4.1) among such papers [152, 153]. We use the above three measures to shortlist a set of candidate papers for Layer-2 (Contextual similarity calculation). The average score of each measure is used as a threshold to filter important papers from each category. Initially, we remove papers with less than average *indeg* as they are not cited by many and hence less influential. After filtering papers with low in-degree, papers with *degree* score greater than or equal to average degree scores are finally shortlisted for further computation. The sets were determined individually and merged as a set-based union to consider just the unique papers. For example, if a very high-quality paper has low in-degree because of its recent publication, the paper may not be considered in degree centrality calculation, but it gets due consideration in Betweenness, and Closeness centrality calculation and, therefore, may qualify based on these measures.

4.5 Contextual Similarity Calculation (Layer-2)

In this module, we mainly extract content-based features to prune the set of papers shortlisted in Layer-1 further. Sometimes it is quite challenging to observe the similarity among papers by looking at only the word level similarities. Even there are cases where the semantic meaning of words is unable to capture the similarity among papers where the use of words in context also needs to be seen. Hence, we need some mechanism specifically to capture the semantic meaning, to discover the hidden patterns and to extract the latent topics other than just identifying words matching. In this work, we applied LDA on abstract and Doc2Vec on the title to address these above issues.

An abstract typically provides a summary containing the main idea of a paper. We use the LDA model on the abstract to generate the feature description [132]. LDA is used to identify topics automatically and to derive hidden patterns exhibited by a text corpus. We have chosen LDA over other methods due to its simplicity, easiness in implementation, fast computation, ability to discover coherent topics and also to handle diverse topics in a text corpus. We set the number of the topic as parameter k while mining a paper's topic distribution to perform LDA. It is used as it generates the probability distribution of words and documents based on the co-occurrence of words and documents, which focus on describing their connotative topics.

We also tried LDA on the title, but due to insufficient terms present in titles, it did not perform well to discover hidden patterns. Hence, Doc2Vec is used to extract the feature description from the title of a paper as Doc2Vec captures contextual information of words occurring in titles [179]. It is mainly used to generate sentence/document embeddings [180]. It is chosen over other methods due to its potential to overcome the weaknesses such as the ordering of words, the semantics of the words, data sparsity, and high dimensionality in bag-of-words models and other approaches. We have used Doc2Vec on the title but not on abstract because it was found a little bit expensive to represent each document by a dense vector that is trained to predict surrounding words in contexts sampled from the document.

4.6 Peer-peer Network Model (Layer 3)

Features so extracted from abstract and title are fused in the next layer to compute:

(i) *Paper2Vec in PPPN model*: In PPPN model, we would like to explore identifying suitable venues through paper-paper peer network by exploiting the concept of Paper2Vec approach without the age-discounted scheme.

(ii) *Venue2Vec in VVPN model*: In VVPN model, we would like to see the quality of the recommendation by incorporating venue-venue peer network through the concept of Venue2Vec approach.

4.6.1 The Architecture of PPPN Model

Due to information overload, it's not practical to check full content similarity to recognize related papers with the seed paper. To address this issue, we are attempting to discover inherent community structures in a bibliographic citation network to understand the network more deeply and reveal interesting relations among the papers.

The process of PPPN model mainly involves four steps:

(i) Paper2Vec feature extraction

(ii) Citation network partitioning

(iii) Topic-oriented intra-graph clustering

(iv) Abstract similarity using Okapi BM25+ algorithm

Paper2Vec Feature Extraction

The results from LDA and Doc2Vec can be considered as two sets of vectors. For each paper p_i, we get a vector A_i for abstract similarity and vector T_i for title similarity. The length of the vector is taken as size k. We are computing the vectors for both abstract and title only once and later on; we will utilize those vectors to calculate the similarity with seed papers. To avoid repetitive computation, a fixed-length vector is considered in this work.

$$\boldsymbol{A}_i^p = [a_{1i}, a_{2i}, \ldots, a_{ki}] \tag{4.1}$$

$$\boldsymbol{T}_i^p = [t_{1i}, t_{2i}, \ldots, t_{ki}] \tag{4.2}$$

Using A_i^p and T_i^p for a paper p_i, we compute cosine similarity with their counterpart from the seed paper (p_j).

$$Sim\_abstract(p_i, p_j) = \frac{\boldsymbol{A}_i^p . \boldsymbol{A}_j^p}{|\boldsymbol{A_i^p}||\boldsymbol{A_j^p}|} = \frac{\sum_{b=1}^{k}(a_{bi} * a_{bj})}{\sqrt{\sum_{b=1}^{k} a_{bi}^2} * \sqrt{\sum_{b=1}^{k} a_{bj}^2}} \tag{4.3}$$

$$Sim\_title(p_i, p_j) = \frac{\boldsymbol{T}_i^p . \boldsymbol{T}_j^p}{|\boldsymbol{T_i^p}||\boldsymbol{T_j^p}|} = \frac{\sum_{b=1}^{k}(t_{bi} * t_{bj})}{\sqrt{\sum_{b=1}^{k} t_{bi}^2} * \sqrt{\sum_{b=1}^{k} t_{bj}^2}} \tag{4.4}$$

The overall similarity between a shortlisted paper (p_i) and the seed paper (p_j) is calculated as a weighted sum of the two similarities.

$$Sim(p_i, p_j) = c * Sim\_abstract(p_i, p_j) + (1 - c) * Sim\_title(p_i, p_j) \tag{4.5}$$

where $c \in [0, 1]$ is a tuning parameter. $Sim(p_i, p_j)$ is used to find similarity with the seed paper (See Algo. 4). Top R papers according to the above similarity are also chosen with the topmost paper being paper of interest (I) for a given seed paper as discussed in Sec. 4.6.1.

Generally, researchers cite conceptually related and relevant papers to their work. But all cited papers are not conceptually related to the citing paper, and their corresponding venues may not be similar to the venue of seed paper.

To capture both the strength of connection as well as semantics such as the related topics shared by papers, we apply a hybrid approach of link analysis and topic-oriented intra-graph clustering in a bibliographic citation network. The reason for employing such technique lies in the state-of-the-art literature [181].

To reduce the time complexity, we perform intra-graph clustering in two stages:

(i) To find sub-graphs for the entire citation network found after centrality measure based on modularity[1] maximization.

[1] Modularity of a partition is a scalar incentive between - 1 and 1 that estimates the density of connections inside sub-graphs when contrasted with joins between sub-graphs.

(ii) Within a sub-graph apply intra-graph clustering based on both link and content information.

There are other reasons for this two-stage intra-graph clustering. We attempt to cluster the entire citation network found after centrality measures using the Jarvis Patrick algorithm. But due to unexpected behavior of citation relationship and non-globular nature of papers, the final clusters are found to have a less intra-cluster similarity. Due to irregular dimensionality or sparseness relationship among papers, the clusters found are either very large or clusters with less number of papers or sometimes results with singleton clusters.

We encountered a couple of issue [2], If we try to cluster the entire citation network without applying intermediate network partitioning. To get dense clusters, clusters with varying shapes, sizes, and densities (either not exactly larger in size nor singleton clusters), and to handle high dimensionality we, therefore, apply network partitioning before applying graph clustering.

Citation Network Partitioning

We use the Louvain algorithm for graph partitioning [182]. The quality of the partitions is ensured by high modularity scores [183,184]. This method is chosen over other community detection approaches due to its simplicity, lesser computational time, and better quality of communities (Modularity).

A weighted citation graph $G = (V, E)$, where $i, j \in V$, an edge $l(i, j) \in E$ has weight $w_{i,j}$. The objective of this step is to partition a citation network G into a set S of mutually exclusive and exhaustive sub-graphs S_i=(V_i, E_i).

$$\bigcup V_i = V; \quad \forall S_i \in S \tag{4.6}$$

$$V_i \bigcap_{i \neq j} V_j = \Phi; \quad \forall S_i, S_j \in S \tag{4.7}$$

The step provides us 293 number of partitions which are almost uniform containing about an equal number of papers.

[2] When Jarvis Patrick algorithm was employed on 32,0,69 papers shortlisted after centrality measures; few clusters were with the average number of papers more than 700, some with less than 3 papers or even a single paper.

Algorithm 4: Modified Jarvis-Patrick clustering

```
Input: Observed citation sub-graphs S with paper-paper connectivity
Output: The algorithm partition input papers into non-hierarchical clusters
Initialization: Let
P = $\{p_1, p_2, \ldots, p_n\}$ be the set of candidate papers present in sub-graphs S
T = User-defined threshold for similarity
F = minimum required number of neighbors in common
for $i \leftarrow 1$ to $|P|$ do
    for $j \leftarrow 1$ to $|P|$ do
        if $(p_i \neq p_j)$ then
            for $k \leftarrow 1$ to $11$ do
                c ← (k-1)*0.1    /* param values c = {0, 0.1, ..., 1} */
                $sim_k(p_i, p_j) \leftarrow \frac{c*S_1+(1-c)*S_2}{dist(p_i,p_j)}$
                where,
                $S_1 \leftarrow abstract\_similarity(p_i, p_j)$ using Eqn. 4.3
                $S_2 \leftarrow title\_similarity(p_i, p_j)$ using Eqn. 4.4
                $dist(p_i, p_j) \leftarrow$ the minimum hop length between $p_i$ and $p_j$
            end
        end
    end
end
for $i \leftarrow 1$ to $|P|$ do
    resultant-set($p_i$)= set of neighbors of $p_i$
    = $\{p_j : sim_k(p_i, p_j) >= T$ for any k$\}$
end
for $i \leftarrow 1$ to $|P|$ do
    for $j \leftarrow 1$ to $|P|$ do
        if $|resultant\text{-}set(p_i) \cap resultant\text{-}set(p_j)\ | >= F$ then
            cluster($p_i$ and $p_j$)
        end
    end
end
return identified clusters along with their non-overlapping papers
```

Topic-oriented Intra-graph Clustering

We consider each partition for further clustering based on link and contextual similarity. A weighted sub-graph $S_i = (V_i, E_i)$ is divided here into n_i clusters using Jarvis Patrick algorithm [185]. The reason behind the selection of Jarvis Patrick to cluster each sub-graphs found after Louvain algorithm are: It will find tight clusters embedded in loose one. It is mainly good for detecting chain-like or non-globular clusters. The clustering steps are very fast, and the overhead requirement is very low. The capability to find clusters of different shapes, sizes, and densities in high dimensional data.

Algorithm 5: Sub-clusters merging algorithm

```
Input: Identified sub-clusters along with non-overlapping set of papers
Output: Merging clusters to collect relevant candidate set of papers
Initialization: Let
C = $\bigcup_i \{c_{i1}, c_{i2}, ..., c_{in_i}\}$ be the set of sub-clusters for all the partitions taken together
  (found after applying Jarvis Patrick algorithm)
R= $\{r_1, r_2, ..., r_r\}$ be the set of topmost $r$ similar papers by using Eqn. 4.5
$candidate\_set = \phi$
for $i \leftarrow 1$ to $|R|$ do
    for $j \leftarrow 1$ to $|C|$ do
        if $(r_i \in c_{ij})$ then
            $candidate\_set = candidate\_set \bigcup c_{ij}$    /*All papers in $c_{ij}$ */
        end
        $j \leftarrow j + 1$
    end
    $i \leftarrow i + 1$
end
collect the set of identified sub-clusters and merge them
return final candidate_set
```

The objective of this step is to make from each partition coherent clusters of papers that are closely related to each other.

Let C_i be a set of n_i number of such clusters for partition S_i.

$$\bigcup_{j \in \{1,2,\ldots,n_i\}} c_{ij} = C_i \tag{4.8}$$

$$c_{ij} \bigcap_{j \neq k} c_{ik} = \Phi \tag{4.9}$$

Although Jarvis Patrick works well in graph clustering it suffers from a problem [3]. It utilizes two parameters: the minimum number of common neighbors (F) and the size of the neighbor list (T) between a pair of nodes. But these parameters are predefined before applying Jarvis Patrick and are not generally modified dynamically. Due to these hand-coded or fixed size of the neighbor list (T) in citation networks, we are not guaranteed to get clusters with consistent quality. The reason is the non-globular or irregular dimensionality among papers in a citation network.

To address the above issue and to catch a gathering of more similar objects in one cluster, we alter the original Jarvis- Patrick algorithm. A variable-length nearest neighbor list, a proximity threshold is utilized to decide a variable number of neighbors for each paper. All neighbors that pass the similarity threshold are considered as neighbors to this work. By this alteration, outliers are prevented from joining a cluster while preventing the arbitrary splitting of large clusters is emerging from the limitations imposed by the fixed-length threshold. The detailed steps are given in Algo. 4. This step provides us 387 number of clusters.

Abstract Similarity Using Okapi BM25+ Algorithm

Keeping in mind the overall goal to retrieve only conceptually related papers with the seed paper, merging of clusters need to be done before applying abstract similarity. The complete steps are quoted in Algo. 5. To perform such merging, we need to take after the accompanying rules as given below:

(i) Select top R papers considering the cumulative score of abstract and title similarity with seed paper as discussed and examined in Sec. 4.5 and Sec. 4.6.1.

[3] Between any two papers A and B; A may have a high number of neighbors while B having very few due to the fixed size of neighbor lists. Now for the minimum number of common neighbor (F) and size of the neighbor list (T), A and B cannot come to a cluster although they are semantically quite close and related papers in a bibliographic citation network.

(ii) Select the topmost similar paper as paper the of interest (I) and extract its associated venue as the venue of interest (Z).

(iii) Take individually selected papers (R) and identify their corresponding clusters found by the Jarvis Patrick algorithm.

(iv) Extract all papers present in those identified clusters (assume t_1) and merge them with the selected top papers from set-II (assume t_2).

So after getting top R similar paper, merging of clusters is done by using Algo. 5. In our experiment, we have generally considered 80-120 ($t_1 + t_2$) papers to check the abstract similarity with the seed paper. It has been experimentally observed that there are 65-105 papers (t_1) present after the merging of clusters.

To address the deficiency of Okapi BM25 in its term frequency (TF) normalization component, i.e., the TF normalization is not lower bounded properly, in this work, we adapted Okapi BM25+ (a variant of Okapi BM25) to compute the abstract similarity of P_{seed}, and P_{test} papers. It is specifically applied to retrieve only conceptually related papers with seed paper. Okapi BM25+ is based on the probabilistic retrieval framework [129], whose weighting scheme is defined in Eqn. 2.8 (Sec. 2.4.2).

The papers are sorted and ranked in decreasing order of their similarity score with the seed paper. The ranked papers are used to fetch the venues in the same order and suggest user-specified top N (usually $N \neq t_1$ or t_2) unique venues.

4.6.2 The Architecture of VVPN Model

We are attempting to discover inherent community structures in a venue-venue graph (VVG) to understand the network more profoundly and reveal interesting relationships shared among venues. To measure the topic distribution of venues to capture their respective current scope, age-discounted based Venue2Vec is proposed.

The process of VVPN model mainly involves six steps:

(i) Venues scope variation with time

(ii) Venue2Vec edge weighting

(iii) Generation of the venue-venue graph (VVG)

(iv) Combining meta-path features

(v) Computing meta-path edge weights as features

(vi) Recommendation of biased RWR model

Venues Scope Variation with Time

Researchers usually desire to contact those venues which are currently publishing similar research papers. Hence, topic distribution and title embeddings in recent years can describe the current scope of a venue more accurately. Table 4.3 displays the topic distribution of venue v_i. To quantify a venue's scope, we initially categorize their publications year-wise to capture the topic distribution of venue using their published papers as depicted in Table 4.3.

To capture the variation of the scope of venues, we apply LDA based topic modeling on abstract and Doc2Vec on the title of papers published in venues. LDA gives the year wise topic distribution of the venues and Doc2Vec returns a vector for each year based on contextual information from venues published titles. The results from LDA and Doc2Vec can be considered as two sets of vectors. L_i^v represents the vector of year-wise topic distribution vectors and D_i^v represents the vector of year-wise title embeddings vectors as depicted in Eqn. 4.10 and in Eqn. 4.11 respectively. The years considered are 2000, 2001, ..., 2012. Each year-wise vector is again a vector of k different topics as given in Eqn. 4.15 and Eqn. 4.16.

$$\boldsymbol{L}_i^v = [L_{2000i}^v, L_{2001i}^v, \cdots, L_{2012i}^v] \tag{4.10}$$

$$\boldsymbol{D}_i^v = [D_{2000i}^v, D_{2001i}^v, \cdots, D_{2012i}^v] \tag{4.11}$$

Now, we employ a weighted addition of vectors from each set to get one vector for abstract similarity and one vector for title similarity. We use age-discounted scheme (inverse log-weighting scheme) to give more weight to the current year vectors, and the weight reduces in the decreasing order of years. For each venue v_i, we get a vector A_i^v for abstract similarity and vector T_i^v for title similarity as depicted in Eqn. 4.12 and in Eqn. 4.13 respectively.

$$\boldsymbol{A}_i^v = \sum_{y_i \in Y} \frac{L_{y_i}^v}{log_2(y_o - y_i + 2)}, \text{ and} \tag{4.12}$$

$$\boldsymbol{T}_i^v = \sum_{y_i \in Y} \frac{D_{y_i}^v}{log_2(y_o - y_i + 2)} \text{ where} \tag{4.13}$$

$$Y = \{2000, \cdots, 2012\} \text{ and } y_o \text{ is the latest year in } Y. \tag{4.14}$$

$$L_{y_i}^v = [a_{1i}, a_{2i}, \ldots, a_{ki}] \tag{4.15}$$

$$D_{y_i}^v = [a_{1i}, a_{2i}, \ldots, a_{ki}] \tag{4.16}$$

Using A_i^v and T_i^v for a venue v_i, we compute cosine similarity with their counterpart from the seed paper as discussed in next section Venue2Vec edge weighting.

Example: Table 4.3 shows the initial topic distributions for five topics of venue v_i and Table 4.4 shows the topic distribution after age-discounted weighting scheme being applied. Eqn. 4.17 shows the topic distribution vector of venue v_i in year 2010. The age-discounted vector is given by Eqn. 4.18 (latest year= 2012).

$$A_{2010}^v = [0.1, 0, 0.6, 0.2, 0.1] \tag{4.17}$$

$$\frac{A_{2010}^v}{log_2(4)} = [0.05, 0, 0.3, 0.1, 0.05] \tag{4.18}$$

Furthermore, we adopt a weighted addition of vectors to obtain the final vector, as given in Table 4.4. The final vector A_i for venue v_i after weighted addition will be:

$$A_i^v = [0.81, 0.65, 0.57, 0.47, 0.38] \tag{4.19}$$

If we had applied a simple vector addition without any weights, we would have got a vector $A_i^{v\prime}$ as:

$$A_i^{v\prime} = [1.3, 1.1, 1.2, 0.8, 0.6] \tag{4.20}$$

We can clearly see the difference between A_i^v and $A_i^{v\prime}$. It clearly indicates the influence of topic distribution vector of recent year 2012 in the calculation of A_i^v where as in $A_i^{v\prime}$, all the year wise vectors contribute equally. Furthermore, venue-venue similarity is done among venues exploiting their corresponding weighted vector A_i^v and T_i^v respectively.

Venue2Vec Edge Weighting

Using A_i and T_i for a venue v_i, we compute cosine similarity between any two venues. We get two cosine similarities, $Sim_a(v_i, v_j)$ and $Sim_t(v_i, v_j)$, for a pair of venues, v_i and v_j, using (A_i, A_j) and (T_i, T_j) respectively.

$$Sim_a(v_i, v_j) = \frac{\boldsymbol{A_i^v . A_j^v}}{|\boldsymbol{A_i^v}||\boldsymbol{A_j^v}|} = \frac{\sum_{b=1}^{k}(a_{b,i} * a_{b,j})}{\sqrt{\sum_{b=1}^{k} a_{b,i}^2} * \sqrt{\sum_{b=1}^{k} a_{b,j}^2}} \tag{4.21}$$

Table 4.3: Research topic distribution of venue v_i

Year	$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
2008	0.4	0.3	0.2	0	0.1
2009	0	0.3	0.2	0.4	0.1
2010	0.1	0	0.6	0.2	0.1
2011	0.5	0.2	0.2	0	0.1
2012	0.3	0.3	0	0.2	0.2

Table 4.4: Weighted score of topic distribution of venue v_i

Year	$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
2008	0.15	0.11	0.07	0	0.03
2009	0	0.12	0.08	0.17	0.04
2010	0.05	0	0.3	0.1	0.05
2011	0.31	0.12	0.12	0	0.06
2012	0.3	0.3	0	0.2	0.2

$$Sim_t(v_i, v_j) = \frac{\boldsymbol{T_i^v}.\boldsymbol{T_j^v}}{|\boldsymbol{T_i^v}||\boldsymbol{T_j^v}|} = \frac{\sum_{b=1}^{k}(t_{b,i} * t_{b,j})}{\sqrt{\sum_{b=1}^{k} t_{b,i}^2} * \sqrt{\sum_{b=1}^{k} t_{b,j}^2}} \tag{4.22}$$

Now we utilize these two similarity metrics to get one final metric, $Sim(v_i, v_j)$ with the help of an adjustment parameter m as:

$$Sim(v_i, v_j) = m * Sim_a(v_i, v_j) + (1 - m) * Sim_t(v_i, v_j) \tag{4.23}$$

where $m \in [0, 1]$.

We consider this similarity score as contextual similarity features (CSF). We are using this CSF score in Sec. 4.6.2 to generate a weighted VVG (venue-venue) graph and also to compute the edge-weight among venues.

Generation of Venue-Venue Graph (VVG)

In this section, we will create a homogeneous undirected venue-venue graph (VVG) from the HIN graph to recommend relevant venues to the input seed paper. We define this graph as an undirected graph, VVG=(B, D) with a vertex type mapping function ω: B $\rightarrow B$ and an edge type mapping function $\pi : D \rightarrow D$. Here, we have one type of vertex B for each venue.

$$B = \{\text{set of venues where only } P\_main \text{ papers published}\} \tag{4.24}$$

The type of edge D is defined as
$b_1 \xrightarrow{connects} b_2 : \omega(b_1) \in \{P\_main\}, \omega(b_2) \in \{P\_main\}, b_1, b_2 \in B$.

It joins two venues using only one type of edge such that $b_1 \overset{d_1}{\to} b_2$, where $\pi(d_1) \in D$. Table 4.5 lists all types of meta-paths defined in our model. We are extracting the venue of P_main and considering as a core venue to maintain a homogeneous VVG graph. Initially, the CSF score as computed in Sec. 4.6.2 among venues is used to create the VVG graph. The average CSF score is used as a threshold to create the edge between venues. No edge exists with less than average CSF score found among venues.

Table 4.5: Meta-paths used in VVPN model

No.	Meta-path	Description
1.	*common_author*	Core venues share an author
2.	*common_term*	Core venues share a term
3.	*direct_cites*	Core venue cites core venue
4.	*direct_cited_by*	Core venues cited by core venues
5.	*citation_paper*	Core venues share a reference (ref)
6.	*co_citation_paper*	Core venues co-cited together (cite)

Combining Meta-path Features into VVG

Since meta-paths are mostly composite relations of various edge types in a HIN graph, they can capture the distinct relationship between a pair of HIN vertices [177]. We assume that a meta-path connects two different P_main papers x, y that belong to two disjoint core venues v_i, and v_j respectively.

We observed that meta-path features with more than two degree [4] are not much meaningful in our work and even not able to create much difference to compute the similarity among venues. To reduce the time complexity and to obtain a tightly coupled relationship among venues, only one-degree and two-degree meta-path features are incorporated into this VVPN model, and a homogeneous VVG graph is exploited to recommend academic venues. We believe that research papers that share many similar references may use a common set of background knowledge. By using this hypothesis, this information could be used to compute the possible associations among papers.

[4] The degree of a meta-path indicates its length and the distance between two main papers.

Computing Meta-path Edge Weights as Features

To discover the latent association between venues, we have divided the above six meta-paths as depicted in Table. 4.5 into 3 categories of edge weighting.

(i) *Common_Features* (CF): Common author and common term meta-path belong to this category. Common author similarity and common term similarity between two venues v_i and v_j are represented by $Sim_A(v_i, v_j)$ and $Sim_T(v_i, v_j)$ respectively. Term appearing in titles or abstracts of a *P_main* paper after stop word removal and stemming are consider for similarity computation. We use snowball stemmer to get the root words [151]. Jaccard similarity coefficient is used to calculate both $Sim_A(v_i, v_j)$ and $Sim_T(v_i, v_j)$ (Eqn. 4.25). In case of computation of $Sim_A(v_i, v_j)$, sets E and F denote list of authors associated with venue v_i and and v_j respectively.

$$J(E, F) = \frac{|E \cap F|}{|E \cup F|} \tag{4.25}$$

where $0 \leq J(E, F) \leq 1$.

Similarly during $Sim_T(v_i, v_j)$ computation, sets E and F denote sample terms occur in venue v_i and and v_j respectively [159]. Then we are combining the above two similarity scores to obtain CF score (Common_Features) between two venues v_i and v_j respectively. The computation of CF edge weighting between v_i and v_j is defined below.

$$CF(v_i, v_j) = Sim_A(v_i, v_j) + Sim_T(v_i, v_j) \tag{4.26}$$

Generally, none of the CF similarity scores among two venues will get a perfect score of 1, and also random walk is sensitive to a higher probability score. Normalization of data within a uniform range (e.g., (0-1)) is essential to prevent larger applies to the output variables. This representation numbers from overriding smaller ones. One way is to scale input and output variables (z) in the interval $[\rho_1, \rho_1]$ corresponding to the range of the transfer function [186]. Before adding this meta-path CF score into the model, we are individually applying the normalization to be in the range of [0.1-0.9] as shown in Eqn. 4.27.

$$z_i = \rho_1 + (\rho_2 - \rho_1)\frac{(x_i - x_i^{min})}{(x_i^{max} - x_i^{min})} \tag{4.27}$$

After applying this normalization, we will get a normalized CF score $CF^{'}(v_i, v_j)$ among two venues v_i and v_j.

(ii) $Direct-Citation\_Features$ (DCF): The meta-paths such as direct_cites and direct-cited-by are included in this group. The computation of edge weighting of DCF is defined below.

$$DCF(v_i, v_j) = |P_{ij}| + |P_{ji}| \tag{4.28}$$

Where P_{ij} denotes set of papers published at venue v_i and refering to papers published at venue v_j. After applying the normalization defined in Eqn. 4.27, we will get a normalized DCF score $DCF^{'}(v_i, v_j)$ among two venues v_i and v_j.

(iii) $Co-Citation\_Features$ (CCF): The remaining meta-paths such as citation_paper and co-citation_paper are within this group. The computation of edge weighting of CCF is defined below.

$$CCF(v_i, v_j) = \sum_{\substack{k \neq i, \\ k \neq j}} |P_{ik} \bigcap P_{jk}| + \sum_{\substack{k \neq i, \\ k \neq j}} |P_{ki} \bigcap P_{kj}| \tag{4.29}$$

where P_{ik} is the set of papers published at venue v_i and referring to papers published at venue v_k. After applying the normalization defined in Eqn. 4.27, we will get a normalized CCF score $CCF^{'}(v_i, v_j)$ among two venues v_i and v_j.

We add each normalized meta-path scores into the model to analyze their effect on the recommendation quality. We already have initial edge weighting score CSF, which is computed based on the age-discounted scheme (inverse log weighting scheme) based abstract and title similarity as calculated in Sec. 4.6.2. It was purely based on the contextual similarity to be in the range of (0-1). So after applying normalization defined in Eqn. 4.27, we will get a normalized CSF score $CSF^{'}(v_i, v_j)$ among two venues v_i and v_j.

$$CSF^{'}(v_i, v_j) = Sim(v_i, v_j) \tag{4.30}$$

Initially, the recommendation will be provided based on the normalized $CSF^{'}$ matching score.

$$CWS(v_i, v_j) = CSF^{'}(v_i, v_j) \tag{4.31}$$

We need to combine individual normalized meta-path scores into the model, and we call it a combined weighted score (CWS). In addition to normalized CSF score all normalized scores obtained from Eqs. (4.26), (4.28) and (4.29) are added to obtain the

CWS(v_i, v_j) to increase the probability of recommending relevant venues during recommendation. The CWS score can be used as a probability score between venues in VVG graph as computed using Eqn. 4.34 to apply random walk with restart (RWR).

$$CWS(v_i, v_j) = CSF^{'}(v_i, v_j) + CF^{'}(v_i, v_j) + DCF^{'}(v_i, v_j) + CCF^{'}(v_i, v_j) \tag{4.32}$$

4.7 Fusion Model: CNAVER (Layer-4)

To be more specific, the predictions resulting from the PPPN model and VVPN model are first produced separately, allowing us to leverage the individual strengths of both approaches since there is no interdependency between them.

4.7.1 Top Venues Recommendation (PPPN Model)

We apply LDA on abstract and Doc2Vec on the title for Set-II papers (Sec. 4.4) and the top t_2 similar papers are chosen. Abstract and title similarity is computed as discussed in Sec. 4.6.1. We have four assumptions regarding the inclusion of these t_2 papers obtained from Set-II paper for abstract similarity.

(a) There may be few papers which are recently got published without having any citations (Set-II), may be involved with many reputed venues.

(b) The seed paper's title and keywords are matching with some papers in Set-II so there is a possibility that the seed paper may get accepted at similar venues as that of Set-II papers.

(c) Generally the papers published in reputed venues get a high number of citations. Chances of getting acceptance in a new venue are relatively easier than reputed venues.

(d) New venues should get an equal chance of inclusion in the final recommendation to reasonable address the new venue cold-start issue.

4.7.2 Top Venues Recommendation (VVPN Model)

To exploit collaboration network information along with publication content, we employ a popular network-based approach known as a random walk with restart (RWR). RWR provides an excellent way to measure how closely related two nodes are in a graph [187]. The core equation of the RWR model is shown in Eqn. 4.33.

$$R^{(t+1)} = \alpha \mathbf{S} R^{(t)} + (1-\alpha)Q \tag{4.33}$$

where $\mathbf{S}$ is the transfer matrix, representing the probability for each node to jump to other nodes. $R^{(t)}$ is the rank score vector at step t and Q is the initial vector of the form $(0,,\cdots,1,\cdots,0)$. Initially, the rank score of the target node is 1, while others are 0. α is the damping coefficient. With probability $(1-\alpha)$, walker restarts from the start node. We use the transfer matrix S to bias our walker's behavior.

We use the weighted combined score (CWS) found after aggregating various meta-paths features in Eqn. 4.32, to bias the walker towards nodes with a higher content as well as semantical similarity. Edge weight w_{v_i,v_j} for an edge from v_i to v_j is given by the equation below:

$$w_{v_i,v_j} = \frac{CWS(v_i,v_j)}{\sum_{x \in N(v_i)} CWS(v_i,x)} \tag{4.34}$$

where $N(v_i)$ is set of nodes which have incoming links from v_i. RWR is an iterative process. After certain iterations, $R^{(t)}$ converges to a steady-state probability vector. We use $R^{(t+1)}$ venue-rank score vector to give our final top N recommendation.

4.7.3 Final Venues Recommendation (Fusion Model)

Although the social network analysis (SNA), content-based filtering (CBF) and random walk with restart (RWR) are widely used for making venue recommendations, they may not provide the best recommendation results due to their limitations. After getting the individual top N recommendations from both the PPPN model and VVPN model, we need to apply some rank-based fusion because the fusion can provide better recommendation than a single approach and the disadvantages of one approach can be overcome by the other.

Fusion has been widely investigated in the recommendation community. They were often divided into two categories: score-based and ranking-based. Score-based combination methods require similarity information to conduct ranking list aggregation, such as CombSum, CombMNZ, and weight combination [169, 188]. Ranking-based combination methods need rank or position information to integrate different candidate's ranking lists, such as Borda fusion, Condorcet fusion, and MAPFuse [189]. In this research, the Borda

Algorithm 6: Fusion of PPPN and VVPN models

```
Input: shortlisted papers after Okapi BM25+ ($t_1$) and shortlisted Set-II papers ($t_2$),
        Venue of interest ($Z$) for a given seed paper $p_m$
Output: Top N recommended list of venues for $p_m$
Initialization: Let
T= $t_1 + t_2$ be the set of candidate papers
$\mathcal{L}$ = Ordered list of unique venues from top-ranked papers based on abstract
  similarity scores (Sec. 4.6.1)
$= \{a_1, a_2, \ldots, a_N\}$
Borda Count $B_c(a_i) \leftarrow N - i + 1$
$\mathcal{M}$ = Ordered list of unique venues in decreasing order (Sec. 4.7.2)
$= \{b_1, b_2, \ldots, b_N\}$
Borda Count $B_c(b_i) \leftarrow N - i + 1$
$\mathcal{N}$ = Final list of unique venues
$= \{v_1, v_2, \ldots, v_{|\mathcal{N}|}\}$ where $|\mathcal{N}| \leq 2N$
for $i \leftarrow 0$ to $|\mathcal{L}|$-1 do
    for $j \leftarrow 0$ to $|\mathcal{M}|$-1 do
        if ($a_i == b_j$) then
            Borda Count $B_c(v_i) \leftarrow B_c(a_i) + B_c(b_j)$   /*they are same venue*/
        else
            individually consider Borda Count $B_c(v_i) \leftarrow B_c(a_i)$ and $B_c(v_j) \leftarrow B_c(b_j)$
        end
    end
end
Sort venues in the decreasing order of Borda Count ($B_c(v_i)$)
Prepare the final list of top N venues recommendation
```

fusion technique is applied to incorporate the existing prediction lists generated by the PPPN model, and the VVPN model as PPPN provides scores for each venue while VVPN provides ranks of them [190]. The complete steps are quoted in Algo. 6.

4.8 Experiments

In this section, we present the experiments of the proposed fusion model "CNAVER" to evaluate the effectiveness of it. In this section, we present the experimental datasets, evaluation strategy, evaluation metrics, experimental setting, parameter tuning, and baseline methods. All experiments are performed on a laptop with 64-bit Windows 10 operating system, Intel i7-3540M, CPU@3.00 GHz, and 8 GB memory. All the programs are implemented in python.

4.8.1 Dataset

We use a real-world dataset DBLP-citation-network V10 (Sec. 2.7.2) to demonstrate the effectiveness of the proposed system CNAVER against other state-of-the-art techniques.

4.8.2 Evaluation Strategy

We adopt two kinds of evaluations such as Coarse-level or offline evaluation and Fine-level or online evaluation to measure the performances of CNAVER against other state-of-the-art methods (Sec. 2.5).

4.8.3 Evaluation Metrics

We employed various evaluation metrics such as accuracy, MRR, $F-measure_{macro}$, precision@k, nDCG@k, average venue-quality (Ave-quality), diversity, and stability to evaluate the performance of CNAVER (Sec. 2.6).

4.8.4 Experimental Setting

While preparing the test dataset, we consider two scenarios. Firstly, due to operational constraints, 20 sub-domains of computer science were selected as a testing dataset in our experiment. A total of 120 seed papers (6 from each sub-domains) are chosen manually from 20 sub-domains: information retrieval (IR), image processing (IP), security (SC), wireless sensor network (WSN), machine learning (ML), software engineering (SE), computer vision (CV), artificial intelligence (AI), data mining (DM), theory of computation (TC), databases (DB), human-computer interaction (HCI), algorithms and theory (AT), natural language processing (NLP), parallel and distributed systems (PDS), world Wide Web (WWW), web semantics (WS), computer architecture (CO), compiler design (CD) and multimedia (MM).

Secondly, while identifying seed papers following conditions are taken into consideration to measure the effectiveness of CNAVER to handle cold start issues like a new venue and new researcher.

(i) *Category 1 ($2 \leq v_c < 8$)*: Select papers whose associated venues have publications greater than or equal to 2 but less than 8.

(ii) *Category 2 ($8 \leq v_c < 15$)*: Select papers whose associated venues have publications greater than or equal to 8 but less than 15.

(iii) *Category 3 ($15 \leq v_c$)*: Select papers whose associated venues have publications greater than or equal to 15.

(iv) *Category 4 ($2 \leq p_c < 8$)*: Select papers whose associated authors have publications greater than or equal to 2 but less than 8.

(v) *Category 5 ($8 \leq p_c < 15$)*: Select papers whose associated authors have publications greater than or equal to 8 but less than 15.

(vi) *Category 6 ($15 \leq p_c$)*: Select papers whose associated authors have publications greater than or equal to 15.

There are two major categories, i.e., venue count (v_c) and publication count (p_c). Generally v_c denotes the number of published papers of individual venue and p_c denotes the number of publications of a researcher. It is ensured that each category is well represented in the seed papers.

Procedure of Online Evaluation

For this evaluation, we did not have the ready annotation, but we need one. The annotation or relevance assessment is collected from the volunteers through crowdsourcing in the best effort basis. There are 57 researchers with expertise in the subjects of the papers provided with input and output of our recommender system where for each paper, 15 venues recommended. Out of 57 researchers, 23 evaluated 3 papers each, 17 researchers evaluated 2 each and the rest 17 were evaluated by 17 researchers.

All the experts were identified from academia with a minimum of 3 years of research experience. Most were having a Ph.D. except few research students and research assistants who were pursuing a Ph.D. with bachelors' or masters' degree in science or technology. The experts or researchers were so chosen that their active areas of research perfectly match with the topics of seed papers. Among 57 researchers, there were 8 professors, 11 associate professors, 19 assistant professors, 12 senior research students, and the remaining 7 were research assistants.

The experts check the titles, abstracts, authors, year of publication, and venue recommendations of the papers and determine the relevance-level of the recommendations. In this experiment the relevance value r is ternary, i.e., r ∈ {0, 1, 2}.

$$\text{Relevance } (r) = \begin{cases} 2 & \text{perfectly matching} \\ 1 & \text{partial matching} \\ 0 & \text{otherwise} \end{cases} \tag{4.35}$$

It is set to 2 if the expert agrees that the research paper is completely matching with the scope of the journal, set to 1 if there is a partial matching or set to 0 otherwise. But while computing precision, we have assumed the partial relevance as not relevant, i.e., relevance 1 is substituted with a relevance value of 0.

To comprehensively evaluate our proposed method and more specifically, to address the broad research questions (RQs) discussed in Sec. 1.5, we prefer to examine the following sub-queries (SQs):

SQ1: How effective is CNAVER in comparison to other state-of-the-art methods?

SQ2: How is the quality of venues recommended by CNAVER as compared to other state-of-the-art methods?

SQ3: How does CNAVER handle cold-start issues and other issues like data sparsity, diversity, and stability?

Table 4.6: Experimental parameter settings

Parameter	Range	Default
Vector dimension (k)	(10-200)	100
Adjustment parameter (c and m)	(0.1-0.95)	0.7
Similarity threshold (T)	(0.2-0.55)	0.35
Number of neighbor (F)	(5-50)	10
Top similar paper (R)	(5-20)	10
Number of Set-II papers (t_2)	(5-45)	15
Damping constant (α)	(0.1-0.95)	0.65
Number of recommended nodes	(5-50)	15

4.8.5 Parameter Tuning and Optimization

In this section, we demonstrate the impact of various experimental parameter settings including dimensions of vectors (k) for A_i and T_i calculations, adjustment parameter (c), threshold (T), minimum number of neighbor (F), top similar papers (R), number of Set-II papers (t_2) to perform PPPN recommendation and adjustment parameter (m), and damping constant (α) to perform VVPN model respectively.

The ranges and default values of the parameters are depicted in Table 4.6. When the effect of the parameter is under examination, the other parameters are set to default values. These experimentations are performed in the training phase contains known output, and the model learns on this data in order to be generalized to other data later on. The ranges of values of various parameters for which the model achieves higher performance are identified as optimal parameters. During this experimentations, the best results are marked by the 'bold-face' in each position.

Influence of Vector Dimension (k)

In order to find the ideal dimension (k=no. of topics) for LDA, we conduct experiments on four values for vector dimension, i.e., {10,50,100,200}. To find the ideal dimension for vectors A_i and T_i, the value of the adjustment parameter is set to be 0.7, and α is set to be 0.65. We extracted the venues of identified seed papers and selected them as a target node to run the VVPN model respectively, then, observed the average performance of the VVPN model in terms of MRR upon various categories. We repetitively performed such experiments with varying recommendation lists in length to evaluate the influence of the vector dimension on the results. We conduct experiments on four values for vector dimension (k), i.e. {10, 50, 100, 200}. It is observed that the model performs best when the value of the vector dimension is 100.

Table 4.7: Influence of vector dimension (k) on MRR

Topic	MRR					
dimension	2<=v_c<8	8<=v_c<15	15<=v_c	2<=p_c<8	8<=p_c<15	15<=v_c
10	0.0573	0.0881	0.0923	0.0495	0.0761	0.0982
50	0.0619	0.0893	0.1044	0.0562	0.0892	0.1016
100	0.0932	**0.0993**	0.1097	0.0838	0.1038	**0.1134**
200	**0.0946**	0.0989	**0.1104**	**0.0866**	**0.1039**	0.1128

Table 4.7 represents the performance of the VVPN model on various vector dimensions. As we can see, MRR score keeps on increasing and behaving a consistent performance while incrementing of vector dimension. From the whole point of view, the model performs a little better with vector dimension 100. Although with vector dimension 200 its results with the best performance ever. There is no significant improvement in MRR while changing the size of k from 100 to 200. As we know, it's computationally costly as compared to vector dimension 100. So we have considered the vector dimension (k) as 100 in our experiment.

Influence of Adjustment Parameter (c and m)

In order to find the ideal value of m to get the efficient combined score of vectors A_i and T_i, we conduct experiments on 10 possible values for adjustment parameter, i.e. {0.5, 0.45, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05 }. The value of the vector dimension is set to 100, and α is set to be 0.65. We have followed a similar procedure, as explained for vector dimension (k) in Sec. 4.8.5.

In Table 4.8, we can observe that the variation tendency of MRR score performs roughly consistent. We can see that the MRR shows a downward trend with the decreasing value of adjustment parameter $1 - m$. The model performs the best while the value of the adjustment parameter is 0.3. This is due to the case that, in most of the cases, the abstract is giving a better clarity of topic similarity while in some instances, the title resulting better. So considering a similar nature, in this experiment, the value of (1-m) and (1-c) has been taken as 0.3.

Table 4.8: Influence of adjustment parameter (m) on MRR

Adjustment	MRR					
prob.(1-m)	2<=v_c<8	8<=v_c<15	15<=v_c	2<=p_c<8	8<=p_c<15	15<=p_c
0.5	0.0793	0.0849	0.0853	0.0854	0.0861	0.0864
0.45	0.0798	0.0879	0.0851	0.0853	0.0893	0.0847
0.4	0.0867	0.0905	0.0893	0.0915	0.0841	0.0859
0.35	0.0972	0.0895	0.0949	0.0858	0.0903	0.0885
0.3	**0.1093**	**0.1197**	**0.1167**	**0.1134**	**0.1127**	**0.1185**
0.25	0.0972	0.1014	0.1298	0.1016	0.1039	0.1073
0.2	0.0668	0.0848	0.1132	0.0894	0.0917	0.0995
0.15	0.0526	0.0773	0.0866	0.0739	0.0877	0.0914
0.1	0.0473	0.0637	0.0725	0.0683	0.0746	0.0828
0.05	0.0437	0.0591	0.0683	0.0565	0.0677	0.0769

Influence of T, and F in Intra-graph Clustering

Similarly, during Jarvis Patrick, by varying the value of T, we observed the effect on sub-clusters size and similarity among papers belong to each sub-clusters. We investigate the performance of sub-clusters found after varying threshold T as 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5 and 0.55, and F as 5, 7, 10, 12, 15, 20, 30, and 50 while performing intra-graph clustering. We observed that, while considering T as 0.5 or more than that, it results in a larger number of clusters with less number of papers in each sub-clusters. We also saw that due to high T, chances of forming singleton clusters are more. Due to the low value of T, there is a high chance of forming less number of clusters with a larger number of papers in each cluster.

Table 4.9: Influence of restart probability on MRR

Restart	MRR					
prob.(1-α)	2<=v_c<8	8<=v_c<15	15<=v_c	2<=p_c	8<=p_c<15	15<=p_c
0.5	0.0633	0.0729	0.1134	0.0635	0.0862	0.1067
0.45	0.0765	0.0914	0.1048	0.0714	0.0975	0.1095
0.4	0.0933	0.0975	0.1086	0.0837	0.1037	0.1135
0.35	**0.1357**	**0.1432**	**0.1791**	**0.1137**	**0.1174**	**0.1248**
0.3	0.1141	0.1265	0.1464	0.1089	0.1146	0.1195
0.25	0.0973	0.1012	0.1296	0.1019	0.1034	0.1078
0.2	0.0763	0.0847	0.1134	0.0896	0.0917	0.0995
0.15	0.0525	0.0772	0.0865	0.0734	0.0877	0.0918
0.1	0.0472	0.0636	0.0724	0.0687	0.0745	0.0823
0.05	0.0432	0.0594	0.0688	0.0563	0.0671	0.0766

The average similarity among papers in each cluster is less due to their loosely coupled relationship. A similar pattern is shown by the value F during intra-graph clustering. We observed that with the value of T as 0.35, resulting in desired clusters with high average similarity among papers and also showing a tightly coupled relationship. The modularity value observed as 0.69. The best result is found with a value of F as 10 during the experimentation.

Influence of R in Recommendation of PPPN

We first examine whether increasing the number of papers can produce desired recommendation performance. We gradually changed the value of R as 5, 8, 10, 12, 15, and 20, respectively. We observed that, after 10 papers, no such changes are occurring in the recommendation order. This is because most similar papers occur in the list of top 10 and papers that are strongly coupled to those 10 papers are also exhibit a high contextual similarity as well as tightly coupled semantic relationship. After applying Jarvis Patrick, we observed that 10 papers are sufficient to capture most similar papers to recommend appropriate venues as discussed in Algo. 5. We found that there are no such changes on the final recommendation after increasing the value of R. So finally, the value of R is considered as 10 during the experimentation.

Influence of t_2 in Recommendation of PPPN

We have also experimentally tested the effect of the number of papers (t_2) selected from Set-II to perform abstract similarity. We first examine whether increasing the number of papers can produce desired recommendation performance. We gradually changed the value of t_2 from 5 to 45 and noticed that, after 15 papers, there are no such changes occurring in the recommendation order. The upper limit is taken as 45 to offer equal opportunity to Set-II path along with with Set-I path (papers found after intra-graph clustering) as on an average the intra-graph clustering results in 45-85 number of papers. Our proposed model is assumed to recommend a maximum of 15 venues.

Influence of Damping Constant (α)

Also, we measured the performance of CNAVER on damping constant α. The damping constant α is an important parameter in RWR. In order to find the ideal value of α

to perform the random walk on venues venues graphs, we conduct experiments on ten possible values for damping constant, i.e. {0.5, 0.45, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05 }. The value of the vector dimension is set to 100, and the adjustment parameter is set to be 0.35. With higher values of α, the probability of random walker reaching far away as the number of nodes increases. Hence, the chances of getting new venues will be more, but it may result in irrelevant venues.

It is evident from Table 4.9 that there is a drastic increase in MRR while decreasing the probability (1-α) till 0.35. Afterward, it exhibits a downtrend with the decreasing value of damping constant. The MRR score performs the upper convex curve, rapidly rising with the value of (1-α) as 0.35 and then shows a decline and downtrend in performance. So based on the above statistics, we have considered the value of damping constant (1-α) as 0.35 in the rest of the experiment.

4.8.6 Baseline Methods

To measure the effectiveness of the proposed venue recommendation, we compare our results with eight state-of-the-art methods such as FB, CF, CN, CBF, RWR, PRS, PVR, and PAVE (Sec. 2.8.1). Among these eight methods CF and PVR are based on collaborative filtering approach, PAVE and RWR are based on random walk with restart algorithm (RWR), CN and FB are based on co-authorship network, and CBF and PRS are based on content-based filtering method.

4.9 Results and Discussion

In this section, we evaluate the effectiveness of CNAVER against existing state-of-the-art methods. Before assessing the performance of the proposed fusion model CNAVER individual performance analysis of PPPN model and VVPN model are analyzed in two phases such as Offline or Coarse-level evaluation and Online or Finer-level evaluation. During the assessment, best results and the second-best are marked by 'bold-face' and '+' symbol respectively.

Table 4.10: PPPN and VVPN recommendation performance in terms of accuracy and MRR

Approach	Acc@3	Acc@6	Acc@9	Acc@12	Acc@15	MRR
FB	0.0555	0.0972	0.1250	0.1666	0.1944	0.0338
CF	0.0972	0.1111	0.1527	0.1805	0.2361	0.0451
CN	0.1111	0.1388	0.1805	0.2222	0.2500	0.0516
CBF	0.1527	0.1805	0.2083	0.2361	0.2916	0.0648
RWR	0.1944	0.2222	0.2500	0.2916	0.3194	0.0775
PVR	0.2083	0.2361	0.2368	0.3194	0.3472	0.0863
PRS	0.2063	0.2291	0.2486	0.2793	0.3419	0.0875
PAVE	0.2500^{+}	0.2916^{+}	0.3055^{+}	0.3611^{+}	0.4305^{+}	0.0906^{+}
PPPN	**0.3334**	**0.3611**	**0.4027**	0.4722	0.6805	0.1150
VVPN	0.3055	0.3457	0.3888	**0.5138**	**0.7361**	**0.1169**

Best results are highlighted in bold, and 2nd best are marked by (‘+’)

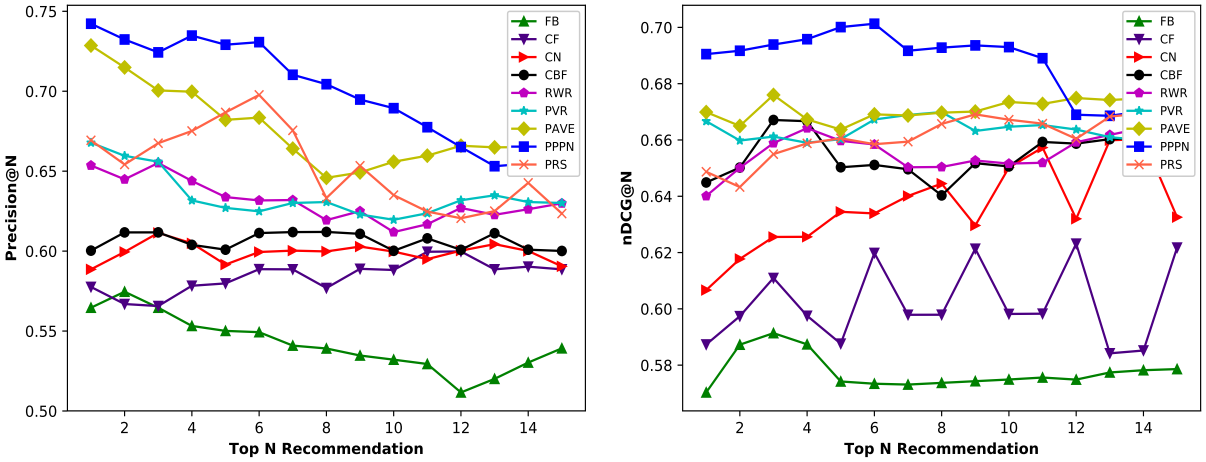
4.9.1 Offline Evaluation of PPPN Model

The complete results of accuracy and MRR are presented in Table 4.10 during the position 3, 6, 9, 12, and 15 respectively. We can see that the PPPN model reveals to a consistent accuracy over all other state-of-the-art strategies. More than 33% of the time (Acc@3=0.3334), it can predict the original venue of the seed paper within top 3 recommendations. The PPPN approach shows an accuracy of 0.6805 while recommending top 15 recommendations. FB strategy exhibits bad performance with an accuracy of 0.1944 while recommending 15 recommendations. More than 68% time, PPPN model can predict the original venue of the seed paper within top 15 recommendation.

For MRR, PPPN performs excellent behavior (MRR 0.1150). The proposed approach can predict the original venue at early ranks compared to all other methods. In the case of MRR also, the least performance is demonstrated by the FB method.

4.9.2 Online Evaluation of PPPN Model

In this section, we analyze the performance of the PPPN model against other state-of-the-art methods. The evaluation metrics, including precision, nDCG, and average venue quality (H5-Index), are taken into consideration throughout this assessment.



(a)

(b)

Figure 4.4: (a) precision@k of PPPN (b) nDCG@k of PPPN

Precision@k

In Fig. 4.4, we can see the significance of the PPPN model as far as precision@k and nDCG@k over all other standard approaches. The PPPN model exhibits the highest precision of 0.7348 at position 4, and after that, it marginally downgrades and furthermore achieves a precision about 0.6775 at position 11 as depicted in Fig. 4.4a. The PPPN model performs superior until the initial 11 recommendations. Afterward, it shows a descending pattern because of which it is unable to maintain consistency as depicted in Table 4.11. This model demonstrates a lower precision of 0.6531 at position 13.

For the first venue, PPPN accomplishes the highest precision among all other methods. Later on, those precision continues diminishing and furthermore achieves a precision about 0.6509 at position 15. PAVE method indicates higher performance over PPPN model at position 12, 13, 14, and 15 respectively. The worst performance among all methods is demonstrated by the FB method.

nDCG@k

The overall nDCG@k of all methods are shown in Table 4.12. Throughout nDCG@k evaluation, PPPN model demonstrates superior scores over all other state-of-the-art methods. The PPPN model performs an upward trend and furthermore achieves the most astounding nDCG 0.7013 during position 6, also subsequently again, it reveals to a descending

Table 4.11: PPPN and VVPN performance in terms of precision

Methods	P@3	P@6	P@9	P@12	P@15
FB	0.5646	0.5493	0.5347	0.5116	0.5392
CF	0.5656	0.5887	0.5889	0.5998	0.5885
CN	0.6114	0.5994	0.6028	0.6003	0.5904
CBF	0.6117	0.6113	0.6108	0.6008	0.6001
RWR	0.6551	0.6317	0.6254	0.6273	0.6299
PRS	0.6675	0.6976^{+}	0.6533^{+}	0.6205	0.6234
PVR	0.6559	0.6248	0.6229	0.6318	0.6301
PAVE	0.7005^{+}	0.6835	0.6492	**0.6659**	**0.6678**
PPPN	**0.7243**	**0.7307**	0.6948	0.6651	0.6509
VVPN	0.6992	0.6998	**0.7207**	**0.7114**	**0.7219**

Best results are highlighted in bold, and 2nd best are marked by ('+')

Table 4.12: PPPN and VVPN performance in terms of nDCG

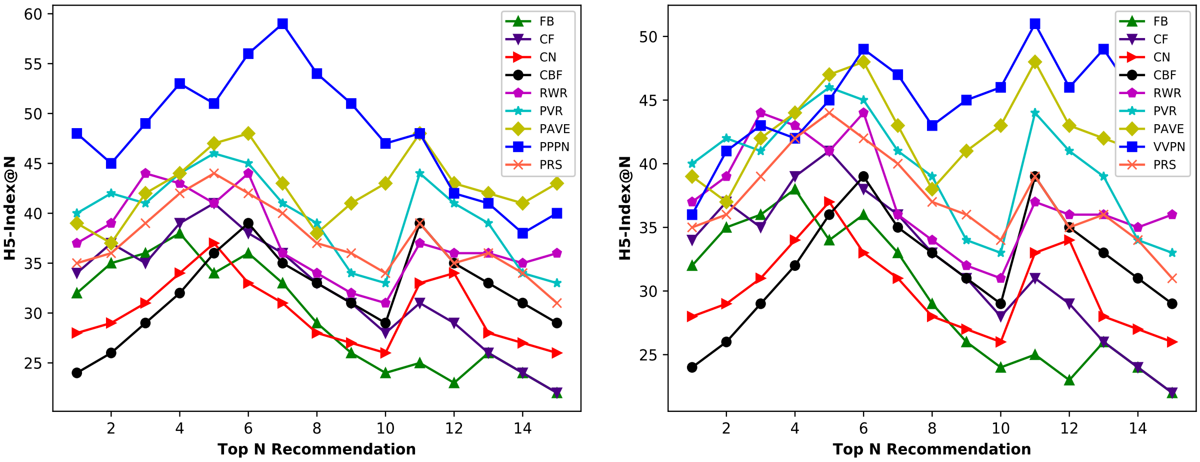
Methods	nDCG@3	nDCG@6	nDCG@9	nDCG@12	nDCG@15
FB	0.5913	0.5734	0.5743	0.5748	0.5786
CF	0.6109	0.6198	0.6213	0.6231	0.6217
CN	0.6255	0.6339	0.6296	0.6319	0.6325
CBF	0.6671	0.6511	0.6517	0.6587	0.6639
RWR	0.6589	0.6584	0.6527	0.6592	0.6657
PRS	0.6549	0.6585	0.6691	0.6604	0.6695
PVR	0.6612	0.6672	0.6632	0.6637	0.6543
PAVE	0.6759^{+}	0.6691^{+}	0.6701^{+}	0.6749^{+}	0.6771^{+}
PPPN	**0.6939**	**0.7013**	**0.6939**	0.6689	0.6706
VVPN	0.6584	0.6699	0.6892	**0.7209**	**0.7425**

Best results are highlighted in bold, and 2nd best are marked by ('+')

pattern and shows nDCG of 0.6685 at position 13. Afterward, it gradually increases and accomplishes a decent nDCG 0.6706 at position 15, as depicted in Fig. 4.4b.

Average Venue Quality (H5-Index) Analysis

We have additionally assessed the quality of venues recommended by PPPN model as compared to other existing methodologies. The PPPN model outperforms other methods in terms of average H5-Index of recommended venues, as illustrated in Fig. 4.5a. While assessing average venue quality, the PPPN model performs an upward trend from the beginning and shows an overall average H5-Index about 49. The top-quality venues recommended by PPPN is in position 7 with the highest H5-Index of 59. Then it indicates a descending pattern furthermore achieves an H5-Index of value 40 at position 15, as shown in Fig. 4.5a. The lowest quality of venues recommended by the FB method with an average H5-Index of 30, whereas the second-highest quality venues recommended by PAVE model with an average H5-Index about 43.



(a)

(b)

Figure 4.5: (a) PPPN average venue quality (b) VVPN average venue quality

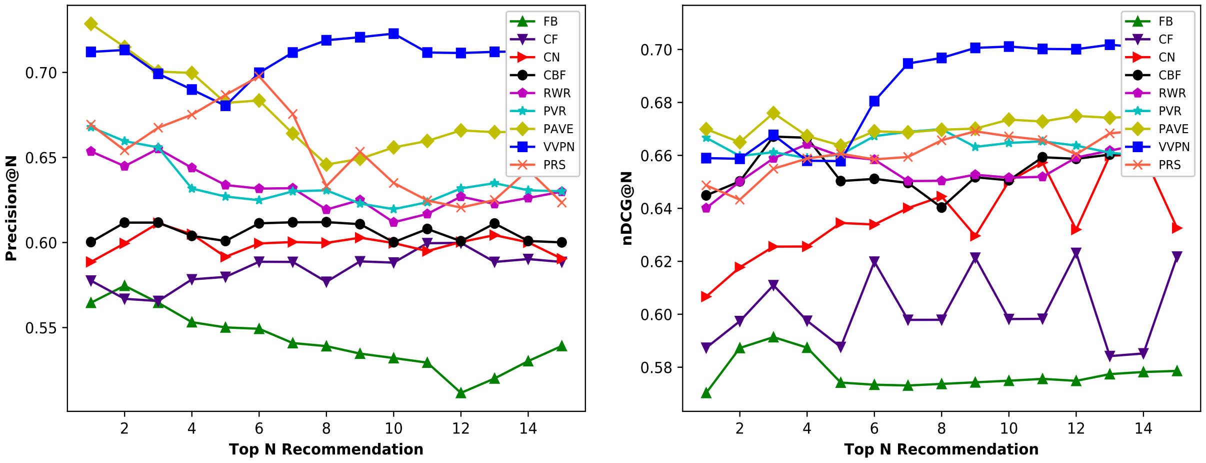
4.9.3 Offline Evaluation of VVPN Model

VVPN model shows a consistent accuracy over all other standard approaches (Table 4.10). More than 30% time it can predict the original venue of the seed paper within the top 3 recommendations. Initially, the VVPN model shows an accuracy of 0.3457 at position 6. Then slowly it shows an upward trend and exhibits an excellent performance with an accuracy of 0.7361 at position 15.

VVPN also shows excellent performance over other standard approaches in terms of MRR. VVPN model exhibits the overall MRR of 0.1169. The second-best performance is shown by the PAVE model with MRR 0.0906. The proposed approach could predict the original venue at early recommendations as compared to all other methods. In the case of MRR also, the least performance is exhibited by the FB method.

4.9.4 Online Evaluation of VVPN Model

We evaluate at a finer level, the effect of VVPN recommendations and compare it against other state-of-the-art methods. We use various metrics such as precision, nDCG, and average venue quality (H5-Index), respectively.



(a)

(b)

Figure 4.6: (a) precision@k of VVPN (b) nDCG@k of VVPN

Precision@k

The compared results are shown in Fig. 4.6. It can be easily observed that the proposed approach VVPN model has made a significant improvement of precision@k over the standard approaches as depicted in Fig. 4.6a. Initially, the VVPN model shows a precision of 0.7132 at a position 2. Then it slowly indicates a downward trend and reaches a precision of 0.6998 at position 6 as depicted in Table 4.11.

But afterward, it shows an upward trend and shows the highest precision of 0.7219 at position 15 and least precision value of 0.6804 after recommending 5 recommendations. The PRS model shows the second-best performance at position 5, 6, and 7 respectively. PAVE exhibits excellent performance at position 1, 2, 3 and 4 respectively. The worst performance among all methods is shown by the FB method.

nDCG@k

The overall nDCG of all methods is shown in Table 4.12. Initially, the VVPN model shows a lower nDCG 0.6587 at position 2. Then slowly it shows a downward trend and reaches the nDCG 0.6578 at position 5. Afterward, it shows an upward trend and is able to show consistency at other positions of the recommendations. It is clearly shown in Fig. 4.6b that the graph of CNAVER is consistent and shows the highest nDCG 0.7097 at position 15. But method PAVE shows higher nDCG than VVPN model at position 1, 2, 3, 4 and 5 respectively. It consistently shows the second-best performance throughout. The FB model exhibits the worst performance.

Average Venue Quality (H5-Index) Analysis

We investigate the performance of venue quality recommended by VVPN as compared to other existing approaches. VVPN model outperforms other methods in terms of average H5-Index of recommended venues. Overall, the average H5-Index of venues recommended by the VVPN model is 45. The top-quality venues recommended by VVPN are at position 11 with the highest H5-Index of 51 as displayed in Fig. 4.5b.

Table 4.13: Accuracy and MRR results of CNAVER and other compared approaches

Approach	Acc@3	Acc@6	Acc@9	Acc@12	Acc@15	MRR
FB	0.0555	0.0972	0.1250	0.1666	0.1944	0.0338
CF	0.0972	0.1111	0.1527	0.1805	0.2361	0.0451
CN	0.1111	0.1388	0.1805	0.2222	0.2500	0.0516
CBF	0.1527	0.1805	0.2083	0.2361	0.2916	0.0648
RWR	0.1944	0.2222	0.2500	0.2916	0.3194	0.0775
PVR	0.2083	0.2361	0.2368	0.3194	0.3472	0.0863
PRS	0.2063	0.2291	0.2486	0.2793	0.3419	0.0875
PAVE	0.2500$^+$	0.2916$^+$	0.3055$^+$	0.3611$^+$	0.4305$^+$	0.0906$^+$
CNAVER	**0.3572**	**0.3888**	**0.4583**	**0.5833**	**0.7916**	**0.1402**

Best results are highlighted in bold, and 2nd best are marked by ('+')

4.9.5 Offline Evaluation of Fusion Model: CNAVER

The complete results of accuracy and MRR after fusion are depicted in Table 4.13. It is evident from the overall results of accuracy and MRR that the proposed approach CNAVER shows a consistent performance over all other standard approaches. More than

Table 4.14: Macro-average analysis in terms of F-measure (F1)

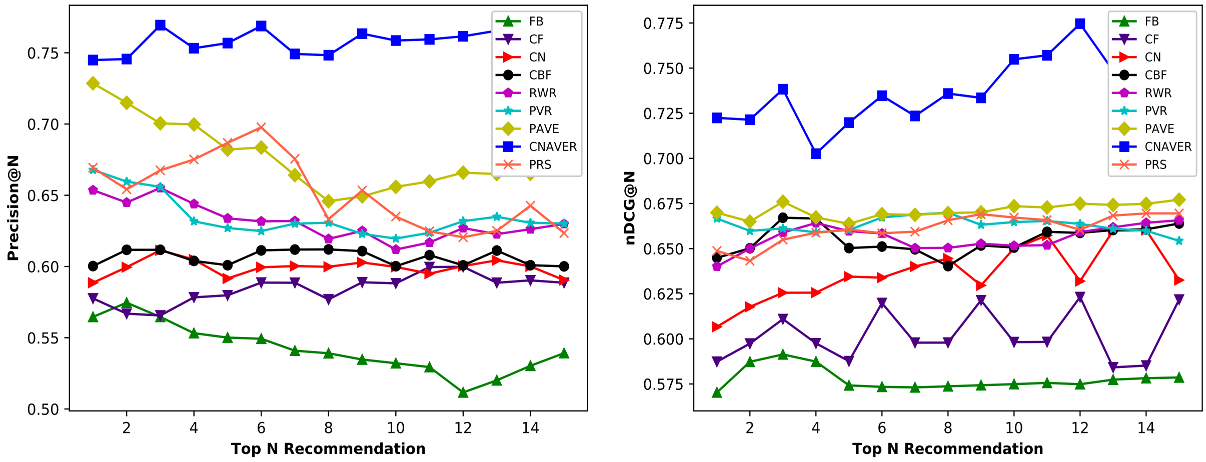
Approach	F1@3	F1@6	F1@9	F1@12	F1@15
FB	0.0128	0.0231	0.0458	0.0412	0.0408
CF	0.0351	0.0437	0.0759	0.0694	0.0621
CN	0.0561	0.0672	0.1045	0.1004	0.0938
CBF	0.0894	0.1025	0.1289	0.1167	0.1125
RWR	0.1148	0.1413	0.2141	0.1894	0.1663
PVR	0.1297	0.1568	0.1854	0.1945	0.1867
PRS	0.1231	0.1456	0.1959	0.1889	0.1826
PAVE	0.1631^{+}	0.2012^{+}	0.2674^{+}	0.2248^{+}	0.2179^{+}
CNAVER	**0.2769**	**0.3179**	**0.3627**	**0.3561**	**0.3524**

Best results are highlighted in bold, and 2nd best are marked by ('+')

35% of the time it could predict the original venue of the seed paper within the top 3 recommendations.

The proposed approach shows an accuracy of 0.7916 after recommending the top 15 recommendations. Similarly, during the evaluation of MRR, we can see that CNAVER outperforms all other state-of-the-art methods and shows excellent behavior with a MRR 0.1402. The proposed approach could predict the original venue at early recommendations better than all other methods. The second-best performance is exhibited by the PAVE, whereas the FB performs the worst among all different standard approaches.

We have also investigated the efficacy of the proposed model CNAVER in terms of $F - measure_{macro}$ against other state-of-the-art methods. The complete results of $F - measure_{macro}$ are shown in Table 4.14. F_1 scores are generally seen to increase with rank up to a certain point (around 9-12) and drop thereafter. This is possibly due to the fact that precision and recall both increase till that point until the original venues are retrieved, causing an increase in F_1 score. However, with further increase in ranks, precision drops sharply without much increase in recall leading to an overall drop in F_1 scores. CNAVER demonstrates the efficacy in comparison to other state-of-the-art methods. The second-best performance is exhibited by PAVE, whereas FB performs the worst.



(a)

(b)

Figure 4.7: (a) Precision of CNAVER (b) nDCG of CNAVER

4.9.6 Online Evaluation of Fusion Model: CNAVER

In this section, the performance of CNAVER against other state-of-the-art methods is discussed. We demonstrate the performance of the proposed system considering various evaluation metrics such as precision, nDCG, and average venue quality (H5-Index), respectively.

Precision@k

The overall results precision and nDCG evaluations are shown in Fig. 4.7. In Fig. 4.7a, we can see the significance of CNAVER in terms of precision over all other standard approaches. Initially, the proposed CNAVER exhibits a precision of 0.7456 at position 2, and after that, it slightly shows an upward trend and shows a precision of 0.7634 at position 9 (Table 4.15).

The proposed model CNAVER exhibits the highest precision of 0.7704 after recommending 15 recommendations. It shows a lower precision of 0.7449 at position 1. Similarly, the PAVE method performs the second-best at positions 1, 2, 3 and 4 respectively. PRS method exhibits slightly higher precision than the PAVE method at position 5, 6, 7 and 9 respectively. The worst performance among all methods is shown by the FB method.

Table 4.15: Precision of CNAVER and other compared approaches

Methods	P@3	P@6	P@9	P@12	P@15
FB	0.5646	0.5493	0.5347	0.5116	0.5392
CF	0.5656	0.5887	0.5889	0.5998	0.5885
CN	0.6114	0.5994	0.6028	0.6003	0.5904
CBF	0.6117	0.6113	0.6108	0.6008	0.6001
RWR	0.6551	0.6317	0.6254	0.6273	0.6299
PRS	0.6675	0.6976^{+}	0.6533	0.6205	0.6234
PVR	0.6559	0.6248	0.6229^{+}	0.6318	0.6301
PAVE	0.7005^{+}	0.6835	0.6492	0.6659^{+}	0.6678^{+}
CNAVER	**0.7694**	**0.7687**	**0.7634**	**0.7629**	**0.7704**

Best results are highlighted in bold, and 2nd best are marked by ('+')

nDCG@k

The nDCG@k evaluations of all methods are shown in Table 4.16. Proposed CNAVER also exhibits better nDCG scores over all other state-of-the-art methods. The CNAVER model performs an upward trend and reaches a nDCG 0.7359 at position 8, and afterward, it shows an upward trend and reaches nDCG 0.7511 at position 15 (Fig. 4.7b). The performance of CNAVER is consistent and shows a nDCG 0.7467 at position 12. The PAVE model demonstrates the second-best performance. The FB model shows the worst performance among all other standard approaches.

Table 4.16: nDCG of CNAVER and other approaches

Methods	nDCG@3	nDCG@6	nDCG@9	nDCG@12	nDCG@15
FB	0.5913	0.5734	0.5743	0.5748	0.5786
CF	0.6109	0.6198	0.6213	0.6231	0.6217
CN	0.6554	0.6339	0.6296	0.6319	0.6325
CBF	0.6671	0.6511	0.6517	0.6587	0.6639
RWR	0.6589	0.6584	0.6527	0.6592	0.6657
PRS	0.6549	0.6585	0.6691	0.6604	0.6695
PVR	0.6612^{+}	0.6672	0.6632	0.6637	0.6643
PAVE	0.6759	0.6691^{+}	0.6701^{+}	0.6749^{+}	0.6771^{+}
CNAVER	**0.7283**	**0.7247**	**0.7235**	**0.7467**	**0.7511**

Best results are highlighted in bold, and 2nd best are marked by ('+')

Average Venue Quality (H5-Index) Analysis

We also investigate the performance of venue quality recommended by CNAVER as compared to other existing approaches. CNAVER outperforms other methods in terms of average H5-Index of recommended venues as depicted in Table 4.17. Overall, the average H5-Index of venues recommended by CNAVER is 54. The top-quality venues recommended by CNAVER are at position 7 with the highest H5-Index of 63 as displayed in Fig. 4.8.

Table 4.17: H5-Index of CNAVER and other compared approaches

Approach	HI@3	HI@6	HI@9	HI@12	HI@15
FB	36	36	26	23	22
CF	35	38	31	29	22
CN	31	33	27	34	26
CBF	29	39	31	35	29
PRS	39	42	36	35	31
RWR	44^{+}	44	32	36	36
PVR	41	45	34	41	33
PAVE	42	48^{+}	41^{+}	43^{+}	43^{+}
CNAVER	**51**	**58**	**52**	**52**	**53**

Best results are highlighted in bold, and 2nd best are marked by ('+')

4.9.7 Evaluation of Diversity

Diversity is defined in terms of content dissimilarity. We group all papers published at a particular venue and extract their corresponding keywords. We apply the similarity score to define diversity in Eqn. 2.29, and the scores are in Table 4.18. CNAVER is seen to show the best diversity, whereas the second-best performer is the method PVR.

4.9.8 Evaluation of Stability

We have also provided a comprehensive investigation of the stability of CNAVER as defined in Eqn. 2.30. CNAVER shows the lower MAS than all other standard approaches (Table 4.18). It shows a MAS of 4.359 on the DBLP dataset, meaning that on an average every predicted venue will shift by a position of 4.359 after adding new data into the training data of the system. We have considered the average MAS-score as a threshold to decide whether a particular method provides stability or not.

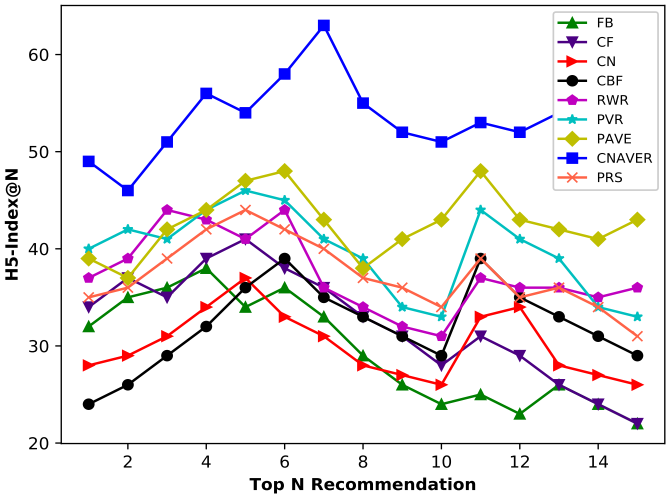


Figure 4.8: Average venue quality of CNAVER and other approaches

Table 4.18: Diversity (D) and Stability (MAS) of CNAVER and other approaches

Methods	Diversity (D)	Stability (MAS)
FB	0.219	9.821
CF	0.338	8.757
CN	0.273	9.452
CBF	0.204	5.769
RWR	0.309	7.884
PVR	0.394^{+}	8.236
PRS	0.273	5.351^{+}
PAVE	0.316	8.349
CNAVER	**0.497**	**4.359**

Best results and 2nd best are marked by bold, and (‘+’)

4.9.9 Study of the Proposed Approach

The main findings concerning our SQs as introduced in Sec. 4.8.4 are summarized below:

SQ1: How Effective is CNAVER in Comparison to State-Of-The-Art Methods?

The overall results of CNAVER and other state-of-the-art methods are displayed in Table 4.13, 4.14, 4.15, and 4.16 respectively. It demonstrates the best performance in terms of precision@k, nDCG@k, accuracy, MRR, and $F-measure_{macro}$ respectively. Also, the difference with the second-best is statistically significant even at 1% level of significance.

Table 4.19: Cold-start and other issues available in CNAVER and other approaches

Methods	Cold-start	Sparsity	Diversity	Stability
FB	yes (new researcher)	no	yes	yes
CF	yes (researcher and venue)	yes	no	yes
CN	yes (new venue)	no	yes	yes
CBF	yes(new venue)	no	yes	no
RWR	yes (new researcher)	no	yes	yes
PRS	yes(new venue)	no	yes	no
PVR	yes (researcher and venue)	yes	no	yes
PAVE	yes(new researcher)	no	yes	yes
CNAVER	no	no	no	no

SQ2: How is the Quality of Venues Recommended by CNAVER as Compared to State-Of-The-Art Methods?

The complete results of venue quality in terms of H5-Index is depicted in Table 4.17. The venues recommended by CNAVER are of high quality when contrasted with other cutting edge techniques as portrayed in Fig. 4.8. The average H5-index of CNAVER is 54 after recommending 15 venues. PAVE recommends venues having the second-best H5-index. The least quality of recommendation performed by the FB model. The most elevated H5-index recommended by CNAVER is 63, and the least is 46, whereas the most noteworthy H5-index suggested by the second-best PAVE is 48 and the least is 37.

SQ3: How does CNAVER Handle Cold-start Issues and Other Issues Like Data Sparsity, Diversity, and Stability

(i) **Cold-start Issues**: To specifically address "cold-start" issues like a new researcher and a new venue, PPPN and VVPN are fused to give venue recommender framework customizable for personalization. Examination of Table 4.8 and Table 4.9 reveal that, regardless of whether the seed paper identified with the new researcher and new venue, CNAVER can predict the original venue at an early stage of recommendations. It does not require past publication records or co-authorship networks for the recommendations. Rather it only focuses on the work at hand. It considers only the current area of interest along with the title, and abstract as inputs to recommend the same.

(ii) **Data Sparsity**: To explicitly address the data sparsity issue, both importance and relevance parameters are considered at the beginning phase of the proposed method. Social network analysis through different centrality measures and content features like abstract and title were used to capture the quality of essentiality, relevance, and importance separately. To extract only related papers, the entire citation network is apportioned, and later on, intra-graph clustering is performed. It has been noticed that the number of papers found after centrality measures are around $32,069$ out of total 2,236,968 papers as input. The average number of papers involved after Intra-graph clustering for abstract similarity is in the range of $80 - 120$. After the initial step, we are left with important papers for further computation, which is close to the area of interest. Hence there is no data sparsity issue in our proposed approach, as indicated in Table 4.19.

(iii) **Diversity**: To resolve the issue of diversity, both connection and contextual similarity-based relevance parameters are taken into consideration. Mainly age-discounted Venue2Vec, meta-path features, and biased random walk are incorporated in VVPN to recommend venues from diverse publishers. 1-degree and 2-degree meta-paths capture different rich latent information in VVPN model. In the PPPN model, topic modeling alongside intra-graph clustering captures both contexts as well as links to suggest relevant papers from diverse publishers. CNAVER, therefore shows the highest value of D (diversity) as compared to all other approaches (Table 4.18).

(iv) **Stability**: A series of techniques like content-similarity, various centrality measures, meta-path, random walk are used to invite stability as well as robustness to the system. Each of these techniques participates in a co-operative manner where the contribution of any single technique is not immensely decisive. Rather, we have some amount of redundancy such that a paper is potentially shortlisted by several techniques. To counter the destabilizing nature of network-based approaches, content-based approaches are incorporated at several places in the pipeline. In all, these batteries of techniques together provide stability to the recommendations. CNAVER shows the minimum MAS than all other standard approaches (Table 4.18).

4.9.10 Discussion on DISCOVER as Baseline

DISCOVER and CNAVER are designed for academic venue recommendation tasks; however, they are treated differently due to their motivations, objectives, and architectural designs. In the introduction of Chapter-4, we already mentioned the reasons for using two different datasets. Information about which field or fields of study a publication does belong to is precious for many tasks. At the same time, this information is often complicated to get, as it is dependent on either having access to the text of the publication or access to manually created metadata. We were more interested in investigating the fields of study provided by MAG for papers in the graph in order to understand their impact on the overall performance of DISCOVER. The fields of study found in MAG are organized hierarchically into four levels (level 0 to level 3, where level 3 has the highest granularity). The fields related to each other have a confidence score signifying relatedness among fields. In DISCOVER, a hybrid binary tree architecture based keyword-based search strategy is adopted to extract relevant papers from the huge volume of data quickly. There are no such fields of study and/or keywords available in DBLP-Citation-network V10 dataset. Therefore we did not include DISCOVER as a baseline for CNAVER and DeepRec

4.9.11 Some Insights

The overall performance results obtained and discussed in Sec. 4.9 showcase the efficacy of the proposed CNAVER. However, there are a few limitations of our work.

(i) The proposed system has multiple parameters involved in both PPPN and VVPN models. Most of the steps involved in the PPPN model are purely based on empirical assumptions but backed by observations from rigorous experimentation.

(ii) If 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. Hence, the proposed model CNAVER may fail to capture the relevant papers resulting in possibly irrelevant venue recommendations.

(iii) In the VVPN model, while choosing the venue of interest (Z) if the topmost paper is contextually similar with the seed paper but its corresponding venue associated with entirely different domains then few of the topmost recommendations by random walk algorithm may not be relevant.

(iv) If the original venue of a seed paper is comparatively new and the venue does not have sufficient number papers, the system may perform poorly. Although the venue of interest Z is contextually similar in content but due to meta-paths features, other venues may be recommended in the VVPN model, but the original venue may not appear at the top of the recommendation list. Hence, it may results in low accuracy and low MRR during on-line evaluation.

4.10 Conclusions

Academic venue recommendation is an emerging area of research in recommendation systems. The prevalent techniques are few in numbers and suffer from various limitations. One of the major issues is cold-start having two sub-parts: a new venue and a new researcher. Additionally, there exist problems of sparsity, diversity, and stability in venue recommender systems that are not adequately addressed in existing state-of-the-art methods.

We proposed a fusion-based scholarly venue recommender system CNAVER incorporating paper-paper peer network (PPPN) model and venue-venue peer network (VVPN) model that reasonably addresses the above-mentioned issues. Several techniques like topic modeling based contextual similarity, link analysis, and topic-oriented intra-graph clustering, abstract similarity using Okapi BM25+ algorithm are used to reinforce the PPPN model. To identify relevant venues, age-discounting-based Venue2Vec, different meta-paths features, and biased random walk with restart (RWR) algorithm are incorporated into the VVPN model. We conducted an extensive set of experiments on a real dataset DBLP and showed that CNAVER consistently outperforms state-of-the-art methods. It shows substantially higher scores of precision@k, nDCG@k, accuracy, MRR, and diversity than other best in class techniques. CNAVER proposes top-notch venues as per H5-index.

Nonetheless, there is scope for continuous update of the model. Considering the fast development of digital information technology, we would like to employ a web crawler to update the training dataset and the learning model continuously. This crawler will automatically extract and collect the relevant data to generate the training dataset. To continually enhance the quality of the recommendation of CNAVER, we plan to collect feedback from users through a web-based application. We plan to adopt some information retrieval techniques like relevance feedback or pseudo relevance feedback to improve the relevance of final recommendations. In future, we would like to incorporate advanced machine learning techniques such as gradient descent optimization in such a way that it will enforce the random walker not to go too far from the initial venue of interest (Z).

We intend to explore with different datasets and to broaden it for various controls with the objective of enhancing precision, accuracy, diversity, novelty, coverage, and serendipity.