**SNAVER: A Social Network Analysis Based Scholarly Venue Recommender System**

**ABSTRACT**

Academic venues act as the main platform of communities in academia and bridge of connecting researchers, which have rapidly developed in recent years. However, information overload in big scholarly data creates tremendous challenges for mining useful and effective information in order to recommend researchers to acknowledge high quality and fruitful academic venues, thereby enabling them to participate in relevant academic conferences as well as contributing to important/influential journals. The ever-growing number of venues publishing academic work makes it difficult for researchers to identify venues that publish data and research most in line with their scholarly interests. Rejection is the norm in academic publishing. One of the main reasons for rejections is that the topics of the submitted papers are not relevant to the scope of the journal, even when the papers themselves are excellent. A solution is needed, therefore, whereby researchers can identify information dissemination pathways in order to both access and contribute to an existing body of knowledge. While every researcher knows the few top-level venues for his specific fields of interest, a venue recommendation system may be a significant aid when starting to explore a new research field.We propose a venue recommendation system which requires only title and abstract, keywords differently from previous works which require full-text and reference list: hence, our system can be used even in the early stages of the authoring process and greatly simplifies the building and maintenance of the knowledge base necessary for generating meaningful recommendations. In this study, we present a system to recommend scholarly venues rated in terms of relevance to a given researcher’s current scholarly pursuits and interests. We assessed our proposal using a standard metric on Microsoft Academic Graph (MAG) dataset to show that our method provides recommendations whose quality is aligned with previous works, while requiring much less information from both the paper and the knowledge base. Evaluation results show that the proposed SNAVER approach outperforms traditional recommendation techniques that can be applied to journal recommendations in terms of quality and performance. The results demonstrate that, in comparison to relevant baseline approaches, SNAVER performs better in terms of precision, recall, F1 and NDCG. This research is the first attempt to provide an integrated framework with the inclusion of social network analysis for effective recommendation in the context of scientific venue recommendation.

**Categories and Subject Descriptors**

H.3.3 [**Information Search and Retrieval**]: Clustering; I.2.7

[**Natural Language Processing**]: Text analysis; I.5.3

[**Clustering**]: Similarity measures

**General Terms**

Key-Route, Algorithms, Measurement, Experimentation, Analysis, Approach

**Keywords:** Academic venue recommendation, Big scholarly data, Social network analysis, Cluster analysis, Latent dirichlet allocation, n-gram, Citation analysis, Factorization model

**1. INTRODUCTION**

Nowadays, the number of the researchers, articles and academic venues has risen beyond the imagination of various research communities due to rapid development of Information Technology (IT). However, the task of mining useful and effective information in big scholarly data is more complex and challenging due to information overload. Academic recommender systems have substantiates their necessity and importance because they objectively provide users with personalized information services. Most academic recommender systems focus on these four problems: collaborator recommendation, paper recommendation, citation recommendation and academic venue recommendation [1].

The immense growth of academic venues makes it troublesome for researchers to choose the most suitable venue, which is witnessed by DBLP. A service that provides open bibliographic information on major computer science journals and proceedings. It has recorded 3711 conferences and 11391 journals. Researchers usually desires to contact suitable academic venue, i.e. acknowledging high quality and fruitful academic venue, participating in academic conferences or workshops which are closely related to their research and publishing their papers and research achievements in important and relevant journals. Let’s verify these two scenarios. 1) An industrious researcher has made a breakthrough in his research area. Consequently, in order to share his work with other relevant researchers, such an industrious researcher has to find a suitable academic venue (conference). The question is, how he can find the relevant one with significant effects. 2) A junior researcher, i.e. a researcher who is at the initial stage of his research and has few publications, intends to extend his research. But the lack of appropriate academic venues information is a challenge for him to find relevant venue to consider and to publish his manuscript. 3) Additionally, although a veteran researcher knows his research area well, he may need a solution relating to cross field venue recommendation.

Considering the inherent requirements, a variety of approaches relating to academic venue recommendation have been proposed [2, 3, 4, 5, 6]. There are also some smart conferences systems or solutions that help improve participation experience and solve the conference recommendation problem [7]. However, most of the researcher did not take the aforementioned problem into consideration. In this work we propose a novel social network based academic venue recommendation system (SNAVER). We first integrate the academic entities (i.e. author, publication and venue) into a citation network, which contain paper to paper network and two kinds of association (in degree and out degree). Furthermore, we propose three notable hypotheses, 1) The closeness of a paper reflect the similarity of given paper with other papers.

In addition to the variety of challenges researchers face from the rising number of scholarly events and venues, the important task of identifying relevant publication opportunities is further complicated due to the expansion and overlap of what were previously discrete specializations. More and more collaboration is taking place between disciplines in the research landscape, which is leading to decreased compartmentalization overall. Increasingly complex academic sub-disciplines and emerging interdisciplinary research areas, though certainly a net gain for the community as a whole, compound this problem. In such a sophisticated research environment, researchers are finding it challenging to remain up to date on new findings, even within their own disciplines (Kuruppu& Gruber, 2006; Murphy, 2003). Furthermore, “context-drift” in scholarly com-munities is becoming more prevalent as researchers expand, evolve, or adapt their interests in rapidly changing subject areas.

Generally, researchers become aware of scholarly venues related to their research interests by word of mouth from lab members, departmental colleagues, and members of other scholarly communities; through online searches for scholarly material; and from rankings of venues and publishers’ reputations (Buchanan, Cunningham, Blandford, Rimmer, & Warwick,2005; Chu & Law, 2007). In the past, these approaches have yielded satisfactory results, as there were relatively few venues related to any given field. However, in today’s multifaceted, diverse, and interdisciplinary scholarly environment, researcher scan become acquainted with newly available and relevant specialized venues only by spending considerable time and effort explicitly searching for venues that align with their research interests.

It is also essential for funding agencies to become aware of new avenues of research across fields in order to determine future allocations. Further, new interdisciplinary research areas lead to greater challenges for research institutes as they strive to understand dynamic information needs and information-seeking behaviors. Information specialists need prompt and seamless measurements of researchers’ readings in order to make decisions on venue subscriptions, instead of relying blindly on the venue’s impact factor or on users’ explicit requests. For example, Springer provides its users with a form for recommending journals to librarians (Springer, 2015), but this feedback represents only the interests of the individuals who submit recommendations, rather than providing a picture of the entire constituency’s needs.

Many rankings of scholarly venues have been created and used to help researchers become more aware of specific scholarly communities. However, knowing that very prestigious journals, such as Science and Nature, are considered top venues for multidisciplinary fields does not help researchers seeking more specialized venues and communities. Moreover, traditional citation analysis cannot provide quick, adaptive results, especially for new scholarly venues that do not yet have an impact factor.

A number of online services provide collections of venues in an attempt to alleviate some of these problems. For example, the HCI Bibliography (Perlman, 1991) is a specialized bibliographic database on Human-Computer Interaction. AllCon-ferences and Lanyrd are global conference and event directories. Conference Alerts, EventSeer, and WikiCFP provide notifications of upcoming academic events based on keywords.5ConfSearch (Kuhn &Wattenhofer, 2008) enables researchers to search for computer science conferences using keywords, related conferences, and authors. ConfAssist (Singh, Chakraborty,Mukherjee, &Goyal, 2016) classifies conferences as top-tier or not.

However, in this era of big data, retrieving relevant results by manually searching and browsing online is no longer theonly approach to discover new information, not is it generally the most efficient approach. Studies have been conducted in an effort to offer techniques capable of accelerating scholarly discovery, such as summarization, visualization (Gove, Dunne,Shneiderman, Klavans, & Dorr, 2011), and collaborative information synthesis (Blake & Pratt, 2006). Recommender systems have been introduced to filter the overwhelming amount of data by using various data analysis techniques to alleviate information overload (Shenk, 1997; Speier, Valacich, &Vessey, 1999). Recommender systems are already entrenched in the digital landscape, as they provide millions of online users with continually updated suggestions for news, books, restaurants, tourism, movies, and television programs.

With the proliferation of publications, researchers are utilizing academic social networks and reference management systems in order to find, store, and manage references (Farooq, Song, Carroll, & Giles, 2007). Social and online reference management systems enable users to bookmark references to research content, as well as tag, review, and rate research content within their profiles. Scholarly tools such as these play an essential role in the organization of personal article collections and the generation of bibliographies across the research landscape today. Scholarly communities are sharing these digital reference libraries, and this open sharing encourages the formation of new research groups. Such online personal collections or repositories also accurately reflect researchers’ current and past reading, and indicate changes in their interests over time, making these datasets prime targets for recommendation analytics.

In previous work (Alhoori, 2016; Alhoori&Furuta, 2011), we found that several of the participating researchers expressed a notable desire to be aware of new and well-established scholarly venues and events related to their shifting research interests. In this paper, we build a personalized recommendation system which can take citation network into consideration while evaluating the relevance of the venue.

Our task to address is particularly challenging due to the following three practical considerations.

* Nowadays, the number of the researchers, articles and academic venues has risen beyond the imagination of various research communities due to rapid development of Information Technology (IT). However, the task of mining useful and effective information in big scholarly data is more complex and challenging due to information overload. Academic recommender systems have substantiates their necessity and importance because they objectively provide users with personalized information services. In this situation to recommend a suitable venue to a researcher is challenging.
* In addition to the variety of challenges researchers face from the rising number of scholarly events and venues, the important task of identifying relevant publication opportunities is further complicated due to the expansion and overlap of what were previously discrete specializations. More and more collaboration is taking place between disciplines in the research landscape, which is leading to decreased compartmentalization overall. Increasingly complex academic sub-disciplines and emerging interdisciplinary research areas, though certainly a net gain for the community as a whole, compound this problem. In such a sophisticated research environment, researchers are finding it challenging to remain up to date on new findings, even within their own disciplines. Furthermore, “context-drift” in scholarly com-munities is becoming more prevalent as researchers expand, evolve, or adapt their interests in rapidly changing subject areas.
* To recommend a suitable venue to researchers at the early stage of the paper being written is such a challenging task. Most of the traditional systems use the concept of co-authors past publications along with the concept of random walk model to do the above task. But to recommend relevant venues for a junior researcher who does not have even a single publication and without considering the full content of the written paper is a tedious task.

In this work, we propose a novel solution named a social network analysis based scholarly venue recommender system (SNAVER) for the new task of venues recommendation. It is developed based on the recent advance of social network analysis, which is further extended to model the context similarities by combining with the topic modeling and factorization techniques. We entail a few key technical components of our SNAVER as follows.

* For the modelling of information domain, we build a citation network to examine the importance of each candidate paper through various centrality measures like betweenness, closeness, degree, eigenvector, HITS score etc.
* We are also measuring the bibliographic coupling (BC) and co-citation score (CC) to check how strongly the seed paper (Sp) is related with other papers after introducing a new distance measure addition to the above scores . Bibliographic coupling is a measure of the similarity between two papers that refer to the same paper, whereas co-citation is the similarity measure for two papers cited together by other papers. We are dividing the sum of above two scores with the distance of the corresponding paper from the seed paper.
* For the topic modelling we have, we build a context aware recommender based on title, keywords and abstract similarities. To fully exploit the similarities of seed paper with all other papers two types of methods namely LDA similarity measure and Non Negative Matrix factorization has been used. Later on we are integrating the results from both the techniques to enhance the topic modeling.
* Main path analysis is capable of tracing the most significant paths in a citation network and is commonly used to trace the development trajectory of a research field. So finally to analyze a satellite view of a given citation network main path analysis has been experimented to identifies the key route via pajek tool. Key-route search is designed to avoid the problem of missing significant links in both the local and global search.

To sum up, the key contributions of this work are four fold:

* *Venues recommendation without past publication (VRWP):* To our knowledge, we are the first to introduce to address the issue of cold start problem in venue recommender system, which recommend relevant venues to target researchers irrespective of research experience. Our proposed system SNAVER could be beneficial even if a junior researcher do not have past publications.
* *Enhanced topic modelling and most influential papers identification (ETMIFI):* We propose a model which measures the importance of papers through various centrality measures and to detect the most influential papers in a citation network. For the above task we have used measures like bibliographic coupling, co-citation, social network analysis (SNA) and key route identification for main path analysis to recommend suitable venues to researchers. Generally the most cited papers are belongs to a relevant and highly reputed venues. And based on this assumption the system is able to recommend venues in all disciplines irrespective of the past records of a researcher.
* *Mixed types and multi publisher’s venues recommendation (MTMPVR):* The proposed system is able to recommend a combined nature of suggestion including both journals and top tier conferences as a recommendation. Generally the system suggests venues related to reputed publishers like Elsevier, Springer, IEEE, ACM and others. Able to provide a personalized recommender system after taking consideration of user input for filtering purposes. The ranking algorithm will perform better even if there are less number of papers of individual papers as it has to consider all papers as individual candidate papers.
* *Performance measurement of proposed venue recommender system (PMPVR):* We summarize the design principle of our model by analyzing various scenarios a researcher may face before identifying a suitable venue for their research work and successfully addressed those issue while building the system. We further compare the performance of different recommendation systems via a variety of recommendation experiments, including experimentation of a total of hundred topics related to twenty sub-fields of computer science and engineering. It has been observed that our model outperforms several other state of-the-art venue recommendation models with significant improvements of nDCG, precision, recall and F1 scores by 12.96%, 8.24% , 8.24% and 6.88%, respectively. The evaluated results show that our model achieves the best recommendation performance by accurately capturing more relevant venues and correctly predicting the original venue of the seed paper within top retrieved documents. Our findings provide insights for identifying suitable venues for a manuscript before publishing the research work. Social network analysis was beneficial for improving recommendation.

This paper is structured as follows: In Section 2, we discuss related work. In Section 3, we describe an approach for measuring an implicit rating for scholarly venues by monitoring researchers’ behavior. In Section 4, we explain the data collection and the experiments. In section 5, we present and discuss the results.

**2. RELATED WORK**

The comprehensive review paper on recommender systems by Adomavicius and Tuzhilin [1] notes four classes of recommender systems based on how they make recommendations:

* Content-based recommendations. Recommend items similar to items the user preferred in the past.
* Collaborative recommendations. Recommend items that people with similar preferences liked in the past.
* Hybrid approaches. Combine content and collaborative- based methods.
* Preference-based recommendations. Recommend items according to relative preference for the user instead of individual ratings.

Recommender systems streamline and augment a person’s decision-making process, especially when inadequate information is available with which to make an informed decision (Resnick& Varian, 1997). One well-known recommender technique is collaborative filtering (CF) (Resnick, Iacovou, Suchak, Bergstrom, &Riedl, 1994; Schafer, Frankowski, Herlocker, &Sen, 2007). User-based collaborative filtering stems from the idea that users whose respective ratings show a high level of agreement and/or who have a similar history of behaviors are likely to continue to show agreement in these regards. This algorithm searches for users who share similar patterns to those of a current user and uses their ratings to predict unidentified preferences for the current user. Item-based collaborative filtering uses similarities between item ratings to predict users’ preferences instead of using similarities between users’ ratings (Sarwar, Karypis, Konstan, &Reidl, 2001).

Other recommender systems use a matrix factorization approach based on the stochastic gradient descent (SGD) (Bottou&Bousquet, 2008), singular value decomposition (SVD) (Sarwar, Karypis, Konstan, &Riedl, 2000), or SVD++ (Koren, Bell, &Volinsky, 2009). SGD is an iterative learning algorithm for minimizing the error between actual and predicted ratings. SVD reduces the dataset by eliminating insignificant users or items. SVD++ constitutes an improvement on SGD in which it not only considers ratings but also considers who has rated what (e.g., rating an item is an indication of preference).

Recommender systems have been used to recommend movies (Wu &Niu, 2015), research papers (Beel, Gipp, Langer, &Breitinger, 2015), collaborators (Yan & Guns, 2014), experts (Protasiewicz et al., 2016), reviewers (Basu, Cohen, Hirsh, &Nevill-Manning, 2001), citations (Caragea, Silvescu, Mitra, & Giles, 2013), and tags (Song, Zhang, & Giles, 2011).

The need to connect authors and readers goes back to at least 1974 when Kochen and Tagliacozzo (1974) proposed a service to suggest journals for authors’ manuscripts using a mathematical model that took into consideration relevance, acceptance rate, circulation, prestige and publication lag. However, until just a few years ago very little progress had been made in this area. Since then, due in large part to the increasing information overload researchers face when searching for nnew venues, there has been resurgence in research and development surrounding the recommendation of scholarly events (Huynh & Hoang, 2012).

Klamma et al. (2009) recommended academic events based on researchers’ event participation history, whereas (Luong,Huynh, Gauch, Do, & Hoang, 2012; Luong, Huynh, Gauch, & Hoang, 2012) used co-authors’ publication history to recom-mend venues. Boukhris and Ayachi (2014) proposed a hybrid recommender for upcoming conferences related to computer science based on venues from co-authors, co-citers, and co-affiliated researchers. Pham et al. (2011) clustered users on social networks and used the number of papers a researcher had published in a venue to derive the researcher’s rating for that venue. eTBLAST (Errami, Wren, Hicks, &Garner, 2007) and the Journal Article Name Estimator (Jane) (Schuemie&Kors, 2008)recommend biomedical journals based on an assessment of abstract similarity. Silva et al. (2015) considered the quality and relevance of manuscripts in order to recommend journals. They also analyzed the authors’ social networks and identified journals in which similar researchers had published. Other venue recommendation approaches have based ratings on the topic and writing style of a paper (Yang & Davison, 2012), the title and abstract of a paper (Medvet, Bartoli, &Piccinin, 2014), an analysis of PubMed log data (Lu, Xie, & Wilbur, 2009), and personal bibliographies and citations (Küc¸ üktunc¸ ,Saule, Kaya,& C¸ atalyürek, 2012; Kucuktunc, Saule, Kaya, & C¸ atalyürek, 2013).

Recently, some online services have started to provide support for locating relevant journals using title, keyword, and abstract matching. These services include Elsevier Journal Finder (Kang, Doornenbal, &Schijvenaars, 2015), Springer JournalSelector, EndNote Manuscript Matcher,Jane,andEdanz Journal Selector.

In addition, more research has been carried out in recent years on recommending events in general. For example, Xia et al.(2013) presented a socially aware recommendation system for conference sessions, and Quercia et al. (2010) used mobile phone location data to recommend social events. Minkov et al. (2010) proposed an approach to recommending future events, whereas Khrouf and Troncy (2013) used hybrid event recommendations with linked data. Most research on scholarly venue recommendation to date has used citation analysis and the publication or participation history of researchers to build recommendations. Unfortunately, this model cannot be widely generalized, as it would not be useful for new researchers or graduate students who lack an established record of scholarly activity. Furthermore, using only the venues in which a researcher has previously published work undermines the recommendation process, as a researcher might be interested in new research areas in which she or he has not yet published any articles. This research study explores pathways with the purpose of drawing on a researcher’s current personal article collections and readings to build tailored venue recommendations.

The main contributions of our research can be summarized as follows. First, we propose a novel recommendation system for scholarly venues that combines citation analysis and social network analysis. The proposed method is based on multilevel citation networks that compare all indirectly linked papers to the paper of interest to inspect the structural and semantic relationships among them. Our main research objective in this study is to consider the mutual relationships among the papers in a broad network beyond a single level, and to evaluate the significance of each paper through certain centrality measures. Second, the lack of a citation count notwithstanding, the proposed method can find influential papers using centrality measures that are derived from a citation network. Finally, we found that the proposed method outperformed the existing methods, Elsevier journal finder and Springer journal suggester, based on user satisfaction data. We asked users to receive recommendations from these algorithms and rate the recommended item lists based on their satisfaction with the results.

The remainder of this paper is organized as follows. In the related work section, we review the scholarly venues recommendation approaches proposed in existing literature. In the proposed method section, we detail our approach to scholarly venues recommendation by using multilevel simultaneous citation networks. The experimental evaluation section presents and discusses our experimental results and evaluation. The last section concludes this paper.

**3. PRELIMINARY**

**3.1 Problem Identification**

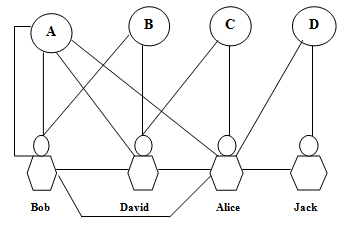
Our experiments are conceived based on a scenario in which the user has a particular paper on mind and has started the research and have the title, keywords and abstract in hand and is searching for a suitable venue (journal/ conference). The driving motivation behind our work is to investigate the possibility of replacing existing academic recommender system to recommend more relevant venues. More specifically we address the following research questions:

* RQ1: To what extent is it possible, using traditional methods like Random walk restart (RWR), modified version of RWR (AVER) with page rank, Friend based and topic based methods to recommend relevant venues in all scenarios?
* RQ2: To what extent are researchers satisfied with the existing Elsevier Journal Finder recommender system?
* RQ3: Does the existing Springer Journal Suggester able to fulfil the requirement of researchers.
* RQ4: What factors are involved in the selection of venues?
* RQ5:To what extent social network analysis can be exploited to influence relevant venues recommendation?
* RQ6: How does the information available influence the recommended venues suggested?

*3.1.1 Is RWR model can maintain a stability to recommend relevant venues in all scenarios*

AVER is designed to mine specific academic venues and make personalized recommendation for researchers. The model is inspired by the fact that , researchers usually desire to keep contact with suitable academic venues, i.e., acknowledging high quality and fruitful academic venues, participating in academic conference which are closely related to their research and contributing to some venues where it is possible for them to publish their research papers and achievements. Additionally, AVER is the evolution from a basic RWR model which has been proved to be suitable for calculating the similarity of nodes in networks.

Referring to figure 1, there are eight academic entities. With respect to recommending venues to Bob, he has never contacted venues C and D. According to the characteristics of the RWR model, the walker can walk from Bob to C and D via David and Alice respectively. After several times of iterative walking, venues C and D are recommended to Bob based on the sorted rank score. However, there are several academic factors that can be introduced to meet the real scene. Generally, researchers contacting the academic entities (researchers and venues) which have high frequency of interaction with them, i.e. high publishing frequency with the researchers.As shown in figure 1, Bob prefers contacting Alice rather than David and Jack because Bob collaborated with Alice twice, David once and Jack nothing.



**Figure 1: An example of co-publication network**

Alice seems to be more important than David and Jack. Furthermore, Bob prefers contacting venue A rather than B and C, since Bob published two papers in venue a. based on this assumption, we define co-publication frequency as equation (1) which is a part of the links weight.

F i, j = {cpi, j i € Author, j € venues (1)

{cti, ji, j € Author

In order to measure the similarity of academic entities, we define a simple metric as shown in equation 2.

LevSim i, j  =1-||ARi-ARj||/max€L(i)(|| ARi-ARx||) (2)

Equation 2 aims at discovering the neighbour with smallest rank score based on a normalization method. Where cpi, j is the count of author i’s publications in venue j. cti, j is author i’s collaborating times with author j. In addition, there are two kinds of associations in co-publication networks. i. e co-author relation and author- venue relation. In this case of basic random walk model, the difference between these two relations is ignored. Author venue relation seems to be more important than co-author relations, because the event of publishing a paper in the venue is more preferable when profiling the researcher’s interest.

*Example*. Let’s say Bob and Alice worked in machine learning domain and successfully published a paper in venue A. Bob himself published a paper by his own contribution in the same venue A. Now bob has two publications in a machine learning domain i.e. A. Bob and David worked together and published a paper in venue B i. e Information Retrieval. David and Alice published a paper in venue C, i.e. image processing. Alice and Jack published a paper in venue D which is wireless sensor network. After several times of iterative walking, venues D would be the recommended venue to Bob based on the sorted rank score. But let’s say Bob’s current interests are still machine learning and information retrieval. But as per the random walk model the suggestions are based on domain like Image processing and wireless sensor network. This is not relevant to researcher Bob. But still the preferred choices of Bob are venues related to machine learning and information retrieval.

There is a need to analyze the above issues and to recommend relevant venues to the researchers not only in terms of co-publication venues but also the content of the paper could be considered for better recommendation. There is still room for future study in this direction. There are many other features such as citation relations, influential nodes calculation that needs to be explored to increase the quality of the recommender system.There is a need to consider the researchers dynamic paper interest (POI) to make a content similarity as well as the social network analysis to measure the significant contributed papers before recommending venues to researchers.

*3.1.2 RQ2: To what extent are researchers satisfied with the existing Elsevier Journal Finder recommender system.*

TheElsevier journal finder, a freely available online service, is one ofthe most comprehensive journal recommender systems, coveringall scientific domains and more than 2,900 per-reviewed Elsevierjournals. The system uses natural language processing for featuregeneration, and Okapi BM25 matching for the recommendationalgorithm. The procedure is to paste text, such as an abstract, andget a list of recommend journals and relevant metadata.This information can help theauthors to decide to which journal to submit their papers, and mayreduce the probability of rejection.

* This system only recommends Elsevier journals.
* The system suggests a maximum of ten journals. The ranking algorithm used in this system only works well if there are enough sample papers (at least more than 100) in each journal. However, for some new journals, there are not enough published papers and the existing system may fail to recommend the relevant journals. Sometime its less than ten journals as results due to these constraints of available similar papers.
* Specific to only one publisher (Either Elsevier or Springer or any other Publisher).
* No choices for conference (In IR conferences are more relevant to the current topics).
* No option to filter the result as per the user specific constraints.
* Mainly working in the principle ofkeyword based search techniques.
* Most of the time suggested journals are out of scope.
* Sometimes show journal with no impact factor.

*3.1.3 Does the existing Springer Journal Suggester able to fulfil the requirement of researchers.*

The Springer journal suggestera a freely available online service is one ofthe most comprehensive journal recommender systems, coveringall scientific domains and more than 2,900 per-reviewed Springer journals. The system uses natural language processing for featuregeneration, and Okapi BM25 matching for the recommendationalgorithm. The procedure is to paste text, such as an abstract, andget a list of recommend journals and relevant metadata.This information can help theauthors to decide to which journal to submit their papers, and mayreduce the probability of rejection. Although this system has a option to refine the results in terms of personalization of journal selection but the existing shortcomings of the systems unable to attract user interests. A few of the shortcomings are as follows.

* This system only recommends Springer journals.
* The system suggests a maximum of twenty journals.
* No choices for conference (In IR conferences are more relevant to the current topics).
* No option to filter the result as per the user specific constraints.
* Mainly working in the principle ofkeyword based search techniques.
* Most of the time suggested journals are out of scope.
* Sometimes show journal with no impact factor.
* Most of the techniques require either full content of the paper or reference list.

**4. MODEL SPECIFICATION AND ANALYSIS**

SNAVER is designed to mine specific academic venues and make personalized recommendation for researchers. The model is inspired by the fact that , researchers usually desire to keep contact with suitable academic venues, i. e. acknowledging high quality and fruitful academic venues, participating in academic conferences which are closely related to their research , and contributing to some venues where it is possible for them to publish their research papers and achievements.

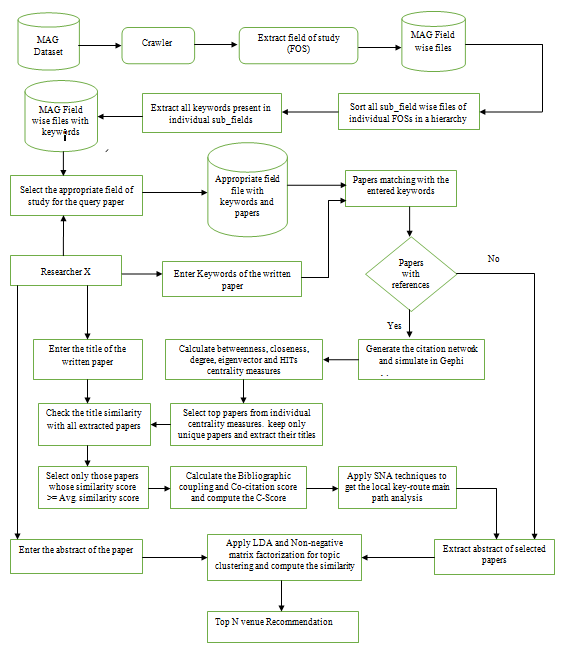
**4.1 Design of SNAVER**

Additionally, SNAVER ids the evolution from a basic social network analysis model which has been proved to be suitable for calculating the similarity of nodes in networks and also to measure the importance of individual nodes. Most of all, academic factors like betweeness, closeness, degree, eigenvector and HITs scores aim at biasing the citation network, so that the most influential nodes could be the final recommended nodes as a result. The detail process of SNAVER is described below. Additionally, the structure of our SNAVER model is illustrated in figure 1.

**4.2 Model Specification**

We propose a recommender system SNAVER which is a social network analysis with a multilevel citation network for venues recommendation. With reference to Figure 2, the detailed process of SNAVER is described below.

* *Step 1: Data Pre-processing.*Initially we need toconstruct the dataset bycrawling data from Microsoft Cognitive Services Academic Knowledge API. We furtherprocessed the dataset by storing papers as per the various field of study. Later the sub-field of study wise papers needs to be stored in a hierarchical manner.
* *Step 2: Formation of clusters among sub-fields.*We have used single link hierarchical clustering to merge sub-field and field of the above datasets. In the dataset a confidence score has been used to merge sub-fields and later on to form one cluster as a root field. Each time we need to merge two sub-fields and the process continued until either one cluster formed or it reached a maximum number of iteration. The storing of data should be in retaining only papers related to field of study of computer science.
* *Step 3: Field of study wise extraction of papers.* The initial input data is a set of publications with title, abstract and keywords respectively. Initially the researcher has to select the appropriate field of study from the lists given in the system.After selecting the appropriate field of study the system will extract only those papers which are present in the same domain. Next, the system will ask to enter a few related keywords of the paper, for which the researcher wants to know the venue.
* *Step 4: Check for the availability of references.* After matching keywords with all related sub field of study the system will extract all papers from the corresponding matching sub fields. Now, the system will check whether those extracted papers have references or not. It will divide those papers into two sets. It will keep those papers whose references are available in the dataset in one set. While the other set retain the other papers whose references are not available or the papers who does not have any citation.
* *Step 5: Citation graph formation and computation of various centrality measures.* The system will proceed with the result set with the available citation data. It will make a citation graph for all papers from that set. And the centrality measure of all individual papers to be calculated. The system has to check the centrality measures like betweenness, closeness, degree, eigenvector and hits score. Individual average score can be used as a threshold to select the candidate papers from the previous set of papers whose references were available. From individual groups separate papers has to be taken. Later the system will merge the results after removing duplicate papers from the resultant set.
* *Step 6: Calculation of title similarity and computing C-Score .*Now, SNAVER will ask to enter the title of the seed paper. Researchers give the title as a input. After getting the title of the seed paper the proposed system SNAVER checks the title similarities of filtered set of papers which are satisfies the average centrality measures. It retains only those papers, who satisfies a minimum average similarity score. And only those filtered papers can be taken into consideration. The bibliographic coupling (BC) and co-citation (CC) score can be computed to get the more similar papers and to get a C-score to filter those papers which are not that much related to the paper of interest (POI). We need to retain only those papers which are satisfies a minimum of average C-score and can be taken into further computation.
* *Step 7: Computation of main path analysis through identification of key route.* The citation network considers the relationship among the papers in a broad area and evaluates the significance of each paper through certain centrality measures. Centrality identifies the most important papers within the network. The process for the proposed SNAVER approach comprises three steps: (1) generation of multilevel citation networks, (2) selection of candidate papers, (3) selection of most significant papers with the help of key route selection, and (4) determination of ranking of each selected significant paper with the help of abstract similarity checks for final recommendation. The system has to perform the key route selection to get only those papers which are most influential in the result set. The major contributed papers can be considered for further computation.
* *Step 8: Selection of candidate papers to* check abstract similarity. After getting the key route papers in the previous step, the recommender system will add a few more papers whose references were not available in the dataset. We are doing this due to the high title similarity of the seed paper with them. We have two assumptions regarding the inclusion of these papers for abstract similarity check. 1) There may be a few papers present whose references are not there but may be involved with many reputed venues. 2) The seed paper’s title is matching with them so there is a probability that the seed paper may get accepted with the same level of venues.
* Step 9: C*omputing abstract similarities by LDA and non-negative matrix factorization (NMF) methods.* After selecting all candidate papers, the system has to perform the abstract similarities based on two key methods. Firstly LDA technique has been used to calculate the similarities of individual papers. Secondly, the same set of papers has to undergo another similarity named non-negative matrix factorization.

**

**Figure 2: The structure of SNAVER**

* *Step 10: Final venues recommendation by merging the results.*The system has to store the order of the recommended venues using both LDA and non-negative matrix factorization methods. Later on we need to merge the ranking of both of these methods to make a efficient recommendation. It has to take the sum of the rankings of these two methods to produce a final ranking of venues. The same ranking would be recommended by the proposed system.
* *Stpe11:* The target researcher can give a few inputs such as venues with specific publishers, only journal, only top tier conferences, minimum impact factor of the journals, number of suggested venues etc to refine the results to make it personalised venue recommendation.

**4.3 Model Analysis**

*4.3.1 Hierarchical clustering of sub-fields and to store the field of study wise papers in hierarchies.*

Add your graph, and explanation. If possible add the algorithm also.

*4.3.2 Algorithms to check title similarity*

Add the algorithm for title similarity.

*4.3.3 Computation of centrality measures to identify significant papers*

*4.3.3.1 Determination of recommends papers.* In the previous step, we selected all papers who satisfies a minimum average title similarity. A centrality analysis of the network is performed to examine the importance of each candidate paper (node) selected by the preceding step [31]. The centrality measure suggests the significance of individual papers due to their relationships with other papers [32]. We calculate four centrality measures (degree centrality, closeness centrality, betweenness centrality, eigenvector centrality and HITS score) of selected papers to determine the most significant papers for recommendation. This approach combines the concepts of citation analysis and network analysis, because network analysis is performed only on the papers selected by citation analysis.

* *Degree centrality (CD)*. The degree centrality is the most intuitive notion of centrality [33]. The more neighbors a given node has, the greater its influence is. We consider only the number of papers “cited,” called the in-degree centrality:



where d(P) is the number of papers referring to paper P, and n is the total number of papers in the network. A high value of the in-degree centrality implies popularity. The calculation of the degree centrality is limited by the number of nodes that are directly connected to the paper, and indirectly connected nodes are not included for the measurement.

* *Closeness centrality (Cc ).*We use closeness centrality to analyze the global hub and authority. Closeness centrality is based on the distance from a paper to all other nodes in the network, and is defined as the inverse total distance. The idea is that a paper is more central if it interacts with more of the other nodes, and it is considered relatively important [34].



where n is the total number of papers in the network. Therefore, n−1 is the minimum sum of distances for a paper that is adjacent to all other papers. P is a target paper of Cc, and J is all papers except paper P. d(P,J) denotes the distance between paper P and other papers J. A node with high closeness centrality is located to the center.

* *Betweenness centrality (CB).* It is based on the number of shortest paths passing through a vertex [35]. Vertices with high betweenness are potential deal makers. They are in a special position because most other nodes have to channel their communications through them. In other words, this measure is the extent to which a paper is positioned on the shortest path between other pairs of papers:



Where gJV is the number of links in the shortest route between paper J and V, and gJV(P) is the number of links in shortest route between J and V that pass through paper P. In the citation network, the papers with closely related citations constitute one community. A paper linking communities can control communication flow among communities, and thus is important. Typically, research papers that influence papers in various fields or that converge existing concepts tend to have a high score. In other cases, nodes play the role of bridging the flow and change in research trends.

* *Eigenvector centrality (CE).*It depends on the number of neighbor nodes that are directly connected to a paper and the quality of the neighbor nodes [36]. Eigenvector centrality measures the influence of set BT containing all papers linking to paper P.

****

where AP,J is the adjacency matrix in which its element is one if J is linked to P, and zero otherwise. xJ is the score of the eigenvector centrality of J, and λ is the eigenvalue of P. Eigenvector centrality measures not only how many papers are connected to a paper, but also how many important papers are connected to a paper.

* *Computation of Average Score (AR).*The values of each of four centrality measures are converted to ranks to combine them. Note that the ranges of the four centrality measures described above are different. After the centrality values are converted to ranks, all four centrality measures have the same scale [1,25]. Now we use the following average ranks (AR) [37] that combines multiple rankings yielded by Eqs. (4)–(7):



where M is the number of centrality measures, and rank (P) is ranking result with kth centrality measure on paper P. Finally, candidate papers are sorted by the AR and the top-n papers in the list are recommended to users. The number of n can be determined by numerous factors, such as the characteristic of the domain field or the system environment.

*4.3.4 Calculation of C-score via computing bibliographic coupling (BC) and co-citation (CC) score*

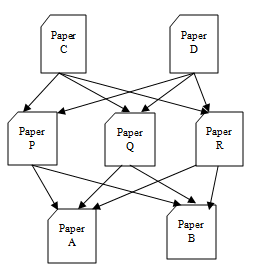
Citation analysis, used in large applications such as patent analysis and document analysis, refers references in one item to another item. The citation network is based on relational information. Therefore, it is useful for understanding the relationship between subjects, the flow of history, and publication trends [20]. The citation analysis of academic papers in particular is important because it can directly reveal papers closely related to the query paper. Several previous studies recommended venues for a manuscript containing a partial list of citations. Co-citation analysis, introduced by Small [21], is one of the first applications of co-occurrence. Small suggested that the more two papers are related to each other, the more often they are co-cited. Liang [22] presented graph networks that show how the papers are connected through citations. Connections are based on bibliographic coupling and co-citation strength [23,24]. Once a graph was built, graph metrics were used to find recommendation candidates. One or several input papers are given as the paper of interest and random walks were conducted to find the most popular items in the network graph [25,26] Much of the literature on citation analysis considers just one level, directly linked to nodes [27]. However, in single-level analysis, the relationship between indirectly connected papers cannot be comprehended.

**4.3.1 Generating multilevel citation networks**

References are a list of cited papers appearing at the end of academic papers. The relations among papers are “cites” and conversely, “cited.” Graphs describing these relationships are citation networks. Fig. 2 shows an example of a multilevel citation network with six levels. The nodes represent papers and the links represent citations with direction. Beginning with paper of interest, I, we use its reference list to start a citation network. “Backward” is the name used to identify citation relationships for papers cited by I, and “forward” is the name used to identify citation relationships for papers citing I. Therefore, the level of a multilevel citation network is the sum of all levels in the backward and forward directions. The paper of interest cites three papers and is cited by four papers. Traditional citation analysis would only utilize a single-level network to consider the directly linked papers. In our proposed method, we extend the network to multiple levels. In the present study, we initially generate the multilevel citation network up to ten levels. We thought that ten is generally acceptable because using more than 10 levels would include the majority of papers, which are unrelated to the paper of interest.

**4.3.2 Selection of candidate papers**

Once we generated the multilevel citation network, we computed a candidate score for each paper (i.e., each node of a multilevel citation network) to select candidate papers. Candidate scores are calculated for all papers appearing in the multilevel citation network to quantify their relevance with the paper of interest. Two different citation relations between papers have been used to measure their similarity; namely, bibliographic coupling and co-citation. Bibliographic coupling is a measure of the similarity between two papers that refer to the same paper, whereas co-citation is the similarity measure for two papers cited together by other papers [29].



**Fig 3. Example of bibliographic coupling and co-citation analysis**

Fig. 3 illustrates bibliographic coupling and co-citation, showing that papers C and D both cite papers P, Q, and R. Their bibliographic coupling strength (B.C strength) can be calculated as follows:



B. C (Ck, Dk) is one if papers C and D cite paper k. In Fig. 3, the B.C strength for papers C and D is three. For papers A and B, which are both cited by papers P, Q, and R, and have a co-citation, the co-citation strength (C.C strength) can be calculated as follows:



C.C (Ak, Bk) is one if paper k cites papers A and B. In Fig. 3, the co-citation strength for papers A and B is three. The purpose of B.C strength and C.C strength is not to analyze the indirect relationship between the papers in a multilevel network, but to find the papers related to the paper of interest in a single-level network. Therefore, B.C strength and C.C strength of papers C and A are calculated independently

We propose merging these two measures to reflect both characteristics by defining the following candidate score (C-score):



In the C-score, the numerator represents the similarity of two papers based on citation information, and the denominator is the distance between the paper and the paper of interest on the network. Therefore, C-score can be considered to be a combination of citation analysis and network analysis. The numerator is the sum of the bibliographic coupling strength and co-citation strength of paper P. J represents all papers except paper P, which is a target of the C-score. The C-score measures how strongly P is related with other papers, J, in both aspects. Thus, a high value of this numerator is an indication that P has a related subject matter with its neighbors. On the other hand, a low value indicates that P is not relevant to the contents of other papers. The C-scores consider the relevance of P with not only J, but also I. The denominator of the C-score determines the boundary of the research area, which has papers that are more relevant and closer to I. d (I,P) is the distance, considered to be the number of links between I and P. The more distance there is between them, the more indicative that the topic or domain field of the two papers is different.

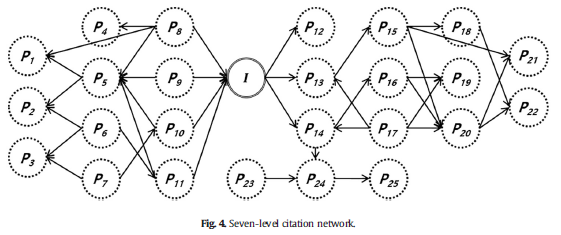
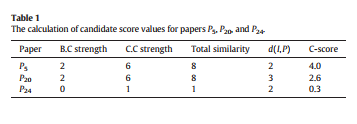


Table 1 shows an example of how the C-scores for papers P5, P20, and P24 are calculated in the seven-level citation network displayed in Fig. 4. For example, the B.C strength of P5 is two ([P5, P8 → P1], [P5, P6 → P2]) and the C.C strength of P5 is six ([P5, P1 ← P8], [P5, P4 ← P8], [P5, I ← P8], [P5, I ← P9], [P5, I ← P10], [P5, I ← P11]). The sum of the B.C strength and the C.C strength is eight, which is the numerator value of the C-score of P5. The denominator value of the C-score of P5 (d(I,P)) is two because the number of links between paper of interest I and P5 is two. For example, P5 and P20 have the same value for the total similarity. However, the Cscore of P20 is lower than that of P5 because P20 is farther than P5 from paper of interest, I. This means that the degree of similarity of P5 and P20 with their neighbor papers is the same, but P5 is more dissimilar to the user topic than P20. On the other hand, although P5 and P24 are at the same distance, P24 has a lower C-



score because the total similarity of P24 is lower than that of P5. The papers with low C-scores tend to be isolated from the network community. In this case, P24 is likely to be an irrelevant paper produced by self-citations and ceremonial citations. Ceremonial citations are citations that were used even though the authors did not read the cited publication [30]. In this way, the C-score is calculated for each of the papers to quantify the degree of the relationship between the paper of interest and all others. Having found the C-scores of all the papers, candidate papers are selected. The number of candidate papers is determined empirically. In this study, we selected top papers with large C-scores. Various experiments indicate that a proper network size for a given problem is between 500 and 800. In addition, our experiments have shown that there is no significant performance difference between networks of sizes 500 and 800. Networks with less than 500 papers may not accurately represent the field. In contrast, networks with more than 800 papers are too complicated to use in practice. The citation level for a network containing 500–800 papers (nodes) is usually six to eight. Note that C-scores will not be used in further steps because they are only used to determine whether the paper is relevant to paper of interest or not.

*4.3.5 Application of SNA to identify key route papers*

Main path analysis is a bibliometric method capable of tracing the most significant paths in a citation network and is commonly used to trace the development trajectory of a research field.The efficient algorithms were implemented in Pajek, a public-domain social network analysis software (Batagelj&Mrvar, 1998) .There are several benefits to applying main path analysis to a citation network. First, it simplifies a complicated citation network to a small number of nodes and links. The analysis provides a satellite view to a given citation network. Under such a view, the paths are like roads on the ground, and only the most significant paths remain whereas paths of lesser significance disappear. Second, it highlights a sequence of major historical-development events, which is very useful to scientists who are considering entering into a particular science and technology domain. Third, it identifies works standing at an important juncture of a field’s historical development. The method begins by measuring the significance of all the links in a citation network through the concept of ‘traversal count’ and then sequentially chains the most significant links into a "main path", which is deemed the most significant historical path in the target citation network. Main path analysis operates in two steps. The first step obtains the traversal counts of each link in a citation network. Several types of traversal counts are mentioned in the literature. The second step searches for the main paths by linking the significant links according to the size of traversal counts. One needs to prepare a citation network before proceeding for main path analysis.

*4.3.5.1 Traversal count.* Traversal counts measure the significance of a link. The literature discusses several types of traversal counts, including search path count (SPC), search path link count (SPLC), search path node pair (SPNP), and other variations.

*4.3.5.2 Search path node pair (SPNP).* A link’s SPNP is the number of times the link is traversed if one runs through all possible paths from all the ancestors of the tail node (including itself) to all the descendants of the head node (including itself). SPNP is first proposed by Hummon and Doreian.

*4.3.5.3 Key-route search***.** Key-route search is designed to avoid the problem of missing significant links in both the local and global search. An additional advantage of the key-route approach is that one is able to control the detail of the main paths by varying the number of key-routes. The larger the number of key-route is specified, the more detail is revealed. A serious potential problem that the main path approach suffers is that the link with the highest traversal count may not always be included in the main path. To overcome this problem, the suggested solution is to view the main path as an extension of the most significant link and begin a search from both ends of the key-route rather than from the sources.

We call this the key-route search. It guarantees that this key route is included in the main path. The key-route search procedure is as follows.

* Select the key-route; it is the link that has the highest traversal count.
* Search forward from the end node of the key-route until a sink is hit.
* Search backward from the start node of the key-route until a source is hit.

The search in Steps 2 and 3 can be either local or global. One also can select multiple key-routes and execute the procedure multiple times, each time selecting the link with the next-highest traversal count, to obtain multiple key-route main paths.

**5. EXPERIMENTS**

**5.1 Data Description**

*5.1.1 Datasets Construction.*We constructed the dataset bycrawling data from Microsoft Cognitive Services Academic Knowledge API. We furtherprocessed the dataset by retaining only papers related to field of study of computer science.

The Microsoft Academic Graph can be accessed via the Microsoft Cognitive Services Academic Knowledge API. The graph is currently being updated on a weekly basis.The data that we used is the Microsoft Academic graph dataset which is a heterogeneous graph that had information relating to a collection of scientific publication records and also the relationships between the records in that collection. With the service of Microsoft Cognitive Services Academic Knowledge API, you will be able to interpret user queries for academic intent and retrieve rich information from the Microsoft Academic Graph (MAG). It is a large heterogeneous graph which models scholarly communication activities and which consists of six types of entitiespublications, authors, institutions (affiliations), venues (journals and conferences), fields of study and events (specific conference instances); and the relations between these entitiescitations, authorship, etc. The dataset contains publication metadata, such as year of publication, title and DOI.

It comprised of over 120 million publication and related authors, papers, institutions, venues and fields of study. There is an ever growing interest in datasets of scholarly publications. As of today, the MAG is the largest publicly available dataset of scholarly publications and the largest dataset of open citation data. The MAG also has very good coverage across different domains with a slight bias towards technical disciplines. On the other hand, there are certain limitations to completeness. Only 30 million papers out of 127 million have some citation data. Similarly as with the author and affiliation entities, the papers in MAG are linked to publication venuesjournals and conferences. MAG also contains a list of conference instances containing information on when and where the conference took place. There are 51,900,106 publications in MAG which are linked to a journal entity and 1,716,211 publications linked to a conference. Interestingly, 103,131 publications are linked to both a journal and a conference.

Information about which field or fields of study does a publication belong to is very valuable 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 investigate the fields of study provided by MAG for papers in the graph in order to understand what the coverage of the dataset is.The fields of study found in MAG are organised hierarchically into four levels (level 0 to level 3, where level 3 has the highest granularity). There are 47,989 fields of study at level 3 (e.g. "concerted evolution"), 1,966 at level 2 (e.g. "evolutionary developmental biology"), 293 at level 1 (e.g. "genetics") and 18 at level 0 (e.g. "biology"). Out of the 126,909,021 total papers, 41,739,531 (about ~ 33%) are linked to one or more field of study entities. In cases where the publication was linked to more than one level 0 field of study, we have counted it towards each linked field of study.

One part of the dataset which is very interesting to us is the citation network. In order to understand how reliable the citation data in the MAG are, we study the citation network from several perspectives. First, we study the network by itself by looking at the citation distribution, to see whether it is consistent with previous studies. We then compare the citations received by two types of entities (institutions and journals) in the graph with citations from external datasets.The MAG contains 528,682,289 internal citations (citations between the papers in the graph). This means each paper in the graph is cited on average 4.17 times. However, a significant portion of the papers are disconnected from the network (neither cite nor are cited by any other papers). There are over 80 million such nodes.

However, despite the limitations, the MAG is currently the most comprehensive publicly available dataset of its kind and represents an astonishing effort which will prove useful in many areas of research where full text access to publications is not required. For our study we used the latest available version of MAG (5 February 2016). Based on the initial investigation on the dataset, we were able to generate a subset of the papers in the dataset based on their field of study. Computer science has been taken as a main field of study to build the venue recommender system. Information Retrieval and Machine Learning were the two fields of study that we considered for building our citation network to recommend venue to researcher. The major attributes of the MAG datasets are mentioned in Table 2.

***Table 2. Microsoft Academic Graph (MAG) Dataset***

|  |  |
| --- | --- |
| Papers | 126,909,021 |
| Authors | 114,698,044 |
| Institutions | 19,843 |
| Journals | 23,404 |
| Conferences | 1,283 |
| Conferences Instances | 50,202 |
| Fields of Study | 50,266 |

**5.2 Experimental Settings**

Our experiments are conceived based on a scenario in which the user has a particular research title on mind along with the features like keywords, field of study and abstract of the written paper and is searching for a suitable venue (journal/ conference).

The driving motivation behind our work is to investigate the possibility of replacing existing academic recommender system to recommend more relevant venues. To comprehensively evaluate our proposed method, more specifically we address the following research questions:

* RQ1: How does SNAVER perform in handling cold-start problem in case of a new researcher?
* RQ2: What are the effects of keywords (number and related keywords with the query topic) with the end results of the proposed personalized venue recommendation?
* RQ3: How does our SNAVER approach perform as compared with other state-of-the-art venue recommendation methods?
* RQ4: Does SANAVER consistently outperform other existing algorithms irrespective of domain with respect to available information?
* RQ5: What are the effects of different hyper-parameter settings (e.g., Key route selection, LDA similarity measure, NMF similarity measure, combined similarity measure information for personalized venue recommendation?
* RQ 6: What is the performance of our final integrated recommender system for the task of personalized venue recommendation?

*5.2.1 Baseline.*In order to demonstrate the effectiveness of our proposed approach, we compare results across several baseline algorithms:

* ***Elsevier Journal Finder:***

To evaluate the accuracy of our proposed system, we applied a strategy of comparing the nDCG values of our approach with Elsevier Journal Finder results. We randomly selected 100 topics from various domains of computer science field which needs to be extracted from the testing dataset. The target researcher has to give the following parameters as input.

* *Input:* Paper title, Paper abstract, Field of research
* *Output:* A maximum of ten journals

After giving the title and abstract as input to the Elsevier journal finder we have retrieved the results. As you know the Elsevier journal finder can suggest up to a maximum of ten journals as relevant venues.

* ***Springer Journal Suggester:***

To evaluate the accuracy of our proposed system, we applied a strategy of comparing the nDCG values of our approach with Springer Journal Suggester. We randomly selected 100 topics from various domains of computer science field which needs to be extracted from the testing dataset. The same topics which we tested in Elsevier journal finder have to be tested in Springer journal suggester. After giving the title and abstract as input to the Springer journal suggester we have retrieved the results. As you know the Springer journal suggester can suggest up to a maximum of twenty journals as relevant venues.

* *Input:* Paper title, Paper abstract, Field of research
* *Output:* A maximum of twenty journals

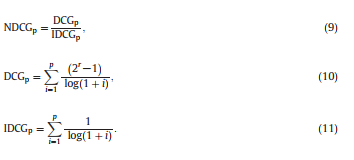
*5.2.2 Comparison with other approaches*

* ***Simple Counting:***For each target paper pi, we simply count the occurring frequency of venues of three kinds of neighboring papers of paper pi, i.e., the reference neighboring papers (papers cited by pi, referred as Simple Count-Ref), sibling neighboring papers (papers that share at least on citation with pi, referred as Simple Count-Sibling) and author neighboring papers (other papers written by authors of pi, referred as Simple Count-Author).We also count the frequency of venues of the combination of all three kinds of neighboring papers (referred as Simple Count-All).We would then rank and return the venues in terms of their frequency.
* ***Content-based LDA:***We construct a profile for each venue by concatenating all the papers published in it. We use LDA topic model implemented by Mallet [15] to retrieve the topic distribution for each paper and venue over 100 topics. We then compute and rank venues by their similarities with the target paper.
* ***Traditional memory-based CF:***We use the original traditional memory-based CF approach, in which we do not incorporate stylometric features of papers to compute their similarity, nor do we categorize neighboring papers and differentiate their different contributions. Under this scheme, P(vj |pi) can be computed as: P(vj |pi) = Ppk\_S(pi) s(pi, pk)I(pk, vj), where papers’ similarity is determined by their topic distribution obtained from LDA.
* ***Graph-based Folk Rank algorithm:***We used the Folk Rank algorithm [20], which is an adaptation of PageRank, and has been shown empirically to generate high quality recommendations in tag recommendation systems. The basic idea of this approach is to run PageRank algorithm twice, giving uniform initial weights to all nodes in the first time, and giving higher weight to targeted nodes in the second time. The difference in terms of the weight of the nodes is then used to generate the final ranking.

**5.3 Evaluation Metrics**

We employed four metrics, precision, recall, F1 score and NDCG, to evaluate the performance of SNAVER. Detailed information about these metrics has been discussed. All experiments we performed on a 64-bit and 3.2-GHz intelcpu, 4-G bytes memory, and implemented with python.

From the recommender system literature, we learn that performance evaluation is mostly conducted by using fewer than 25 items because recommending too many items can confuse users [39,40]. Hundreds researchers with expertise in the subjects of the papers provided their recommendations. The titles, authors, year of publication, and venue names of the recommended papers are provided to the experts. The experts determine whether they are satisfied or not with recommended venues. For evaluation, normalized discounted cumulative gain (NDCG) and mean reciprocal rank (MRR) are used [41,42]. NDCG measures the performance of a recommender system based on the graded relevance of the recommended items.



NDCGp represents the total normalized gain accumulated at a particular rank p. DCGp is the total gain accumulated at a particular rank p. The relevance value, r, of recommended item is binary; r∈{0, 1, 2}. It is set to two if the user is satisfied with the recommended paper, set to one if there is partial matching or it is set to zero otherwise. IDCGp generates the maximum possible DCG until rank p for normalization. All NDCG calculations are then relative values on the interval 0.0 to 1.0. In a perfect recommendation, the NDCG value is one because DCGp will be the same as IDCGp. Mean reciprocal rank (MRR) is widely used in the study of information retrieval and measures the ability of a recommender system toreturn a relevant item at the top of the ranking.

To measure the recommendations’ performance, we measured precision, recall, and normalized discounted cumulative gain (NDCG) (Järvelin&Kekäläinen, 2002; McNee, Riedl, &Konstan, 2006). Precision is derived by dividing the number of relevant venues recommended according to the researcher’s interests by the number of recommended venues, as shown in Eq. Recall is derived by dividing the number of relevant venues recommended by the number of relevant venues, asshown in Eq. (8).

Precision = |relevant *v*enues ∩ top *v*enues|/ |top *v*enues| (8)

Recall = |relevant *v*enues ∩ top *v*enues| |relevant *v*enues| (9)

Discounted cumulative gain (DCG) measures the extent to which a venue ranking is relevant to a user’s ideal ranking, as shown in Eq. (10). Relv is the relevance assigned by a researcher to the venue at position p. We measured the normalized discounted cumulative gain (NDCG), which ranges from 0.0 to 1.0, with 1.0 as the ideal ranking, as shown in Eq. (10). As recommendation lists vary in length, we used NDCG. IDC Gpis the maximum possible ideal DCG at position p.

NDCGp = DCGp/IDCGp (10)

We also incorporated user coverage (Good et al., 1999; Herlocker, Konstan, Terveen, &Riedl, 2004; Sarwar et al., 1998), which is the percentage of users for whom the system was able to recommend venues. Additionally, we tested for the normalized mean absolute error (NMAE) and the normalized root mean square error (NRMSE), which are independent rating scales.

We used 70% of the data as a training set and 30% as a test set. We selected recommendations by choosing a threshold per user that was equal to the user’s average PVR.

**5.4 Performance Comparison**

We conducted experiments to evaluate the recommendation capabilities of the proposed SNAVER compared with Elsevier Journal Finder and Springer Journal Suggester. Elsevier Journal Finder and Springer Journal Suggester are chosen as comparison systems because they are the most recognized proprietary databases for journal content and provide recommendation services to customers [38].

We considered twenty sub domains of computer science field that were selected as papers of interest in our experiments. We chose multiple sub domain papers because we wanted to compare the recommendation results irrespective of the topic of the paper. Moreover, 100 papers are selected from various fields, including Information retrieval, image processing, security, wireless sensor network, machine learning, software engineering, computer vision, artificial intelligence, data mining, natural language processing, parallel and distributed systems, multimedia, world wide web, operating system, databases, programming languages, real time and embedded system, human-computer interaction, bioinformatics and computational biology. We performed a blind test to assess the validity of the proposed method. For each paper of interest, we compute the title similarity of all papers from the field of study computer science. top selected papers whose similarity score are more than the average similarity score considered as candidates for further computation.

**5.5 study of the proposed approach**

The main findings with respect to our RQs are summarised as follows:

5.3.1 How does SNAVER perform in handling cold-start problem in case of a new researcher (RQ1)

* The analyses in Section 7 show that, in case of a new researcher also the proposed system will perform with the same level of accuracy while suggesting relevant venues. Because in this approach the co-authors records are not required while constructing the citation network. The proposed approach considering the titles matching with the query paper and extracting only relevant papers to generate the network for further computation. While the traditional approach has been taken the co-authors publications to draw the networks and applies random walk restart algorithm with a variation to suggest venues to the target researchers. But here the system will perform with the same relevance results irrespective of the types of researchers.

**5.3. 2** What are the effects of keywords (number and related keywords with the query topic) with the end results of the proposed personalized venue recommendation (RQ2)

* We have also observed the suggested venues for a same input with changing their input keywords and conclude that with less number of keywords or unrelated keywords the system may mislead the suggestion and result with less number of relevant venues. While the sufficient number of keywords specific to the topic may results with a high number of relevant venues. So the system as that a sufficient number of key words required. Perception of related keywords and sufficient number of keywords may influence the system. So the system may be influenced by the input information given by the user.

**5.3.3** How does our SNAVER approach perform as compared with other state-of-the-art venue recommendation methods (RQ3)

* The proposed approach outperforms other state-of-the-art venue recommendation methods like Elsevier journal finder and Springer journal suggester etc. The analyses in Section 8 show that the proposed approach is efficient not only terms of precision, recall and F1 score but also outperforming Elsevier journal finder in terms of nDCG @5 and @10. Similarly it’s outperforming springer journal suggester in terms of nDCG @5, @10 and @20 respectively. So it is possible to replace the existing venue recommender systems with more relevant suggestion of venues for target researcher.

**5.3.4** Does SANAVER consistently outperform other existing algorithms irrespective of domain with respect to available information (RQ4)

* The proposed approach outperforms other state-of-the-art venue recommendation methods like Elsevier journal finder and Springer journal suggester etc. Here we are using the huge dataset Microsoft Academic Graph (MAG). It contains various domains with sufficient number of papers to carry out our research in any discipline. As we have already tested the same after checking around hundred topics from around twenty sub domains of computer science. And we have observed that the proposed system is performing stability in terms of performance measures like precision, recall, F1 score and nDCG respectively. Overall the performance has been recorded as consistent irrespective of domains and topics.

**5.3.5** What are the effects of different hyper-parameter settings (e.g., Key route selection, LDA similarity measure, NMF similarity measure, combined similarity measure information for personalized venue recommendation (RQ5)

* We have tested the proposed system with all variation of key route selection. We have identified the main path analysis through various options like forward path, backward path and key route respectively. But observed that the key route is the only one method which is returning significant papers with reputed venues. So tested all hundreds topics with the key route main path analysis techniques. We have also checked the abstract similarities with both LDA and Non-negative matrix factorization methods and observed that while merging it results in a descent ranking order of venues. So opted the integrated approach for ranking final venues before recommending relevant venues to target researchers.

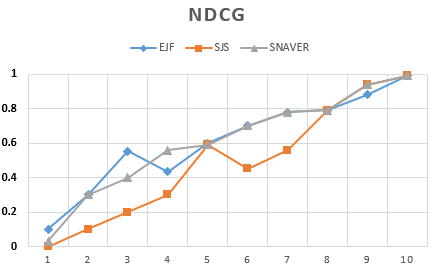
5.3.6 What is the performance of our final integrated recommender system for the task of personalized venue recommendation (RQ6)

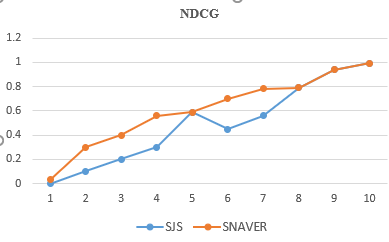
* The proposed approach outperforms other state-of-the-art venue recommendation methods like Elsevier journal finder and Springer journal suggester etc. The analyses in Section 8 show that the proposed approach is efficient not only in terms of precision, recall and F1 score but also outperforming Elsevier journal finder in terms of nDCG @5 and @10. Similarly it’s outperforming springer journal suggester in terms of nDCG @5, @10 and @20 respectively. So it is possible to replace the existing venue recommender systems with more relevant suggestion of venues for target researcher.

**6. Results and Discussion**

The rest of this paper is organized as follows. We introduce the data collection with basic statistics in Sec. 2. .e empirical study on the data and the motivation of our model are shown in Sec. 3. In Sec. 4, we describe and analyze the proposed model of MPF. .en, experimental results are illustrated in Sec. 5. Finally, we present the related work in Sec. 6 and conclude the paper in Sec. 7.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Top K=500 | | | | | | | |
|  | ACM | | | CiteSeer | | | MAG |
|  | Style | Content | S+C | Style | Content | S+C | SNAVER |
| Accuracy @5 |  |  |  |  |  |  |  |
| Accuracy @10 |  |  |  |  |  |  |  |
| Accuracy @20 |  |  |  |  |  |  |  |
| MRR |  |  |  |  |  |  |  |
| Top K=1000 | | | | | | | |
|  | ACM | | | CiteSeer | | | MAG |
|  | Style | Content | S+C | Style | Content | S+C | SNAVER |
| Accuracy @5 |  |  |  |  |  |  |  |
| Accuracy @10 |  |  |  |  |  |  |  |
| Accuracy @20 |  |  |  |  |  |  |  |
| MRR |  |  |  |  |  |  |  |
| Top K=2000 | | | | | | | |
|  | ACM | | | CiteSeer | | | MAG |
|  | Style | Content | S+C | Style | Content | S+C | SNAVER |
| Accuracy @5 |  |  |  |  |  |  |  |
| Accuracy @10 |  |  |  |  |  |  |  |
| Accuracy @20 |  |  |  |  |  |  |  |
| MRR |  |  |  |  |  |  |  |
| All neighbours | | | | | | | |
|  | ACM | | | CiteSeer | | | MAG |
|  | Style | Content | S+C | Style | Content | S+C | SNAVER |
| Accuracy @5 |  |  |  |  |  |  |  |
| Accuracy @10 |  |  |  |  |  |  |  |
| Accuracy @20 |  |  |  |  |  |  |  |
| MRR |  |  |  |  |  |  |  |





In this section, we initially performed several experiments for SNAVER, AVER, basic RWR, topic based and friend based recommendation model on data set discussed above. Secondly, we measured the performance of SNAVER when recommending academic venues for researchers at different levels. We randomly choose 100 researchers as target nodes.

**6. CONCLUSION**

Multidisciplinary research areas are growing at a tremendous rate, and the number of scholarly venues is increasing every year. Researchers need to discover venues that are of interest to them, and research institutions need to be aware of these venues. In this paper, using data from an academic social network, we described an approach to recommend scholarly venues for researchers to follow and/or to publish their work in based on their current interests.

We developed a new approach called SNAVER which can use the concept of social network analysis along with topic modelling to recommend venues to target researchers. And measures the performance agains t standardized recommender system like Elsevier journal finder and Springer journal suggester. We have noticed that for few domains springer achieved slightly better performance in terms of NDCG than Elsevier. But SNAVER achieved a consistent relevance over all sub fields of computer science. Our experiments with this strategy using a real dataset produced results that showed improvements in accuracy and ranking quality compared with a standard baseline. A number of factors will be investigated to improve the results and recommendation quality, including the total number of papers published in a venue, the number of online references to a venue in an academic social network, the average number of references added by researchers to an online reference management system, the dates on which references were added to the researchers’ repositories, and the readership statistics for an article.

In future research, we plan to enhance the quality of our generated recommendations by using a researcher’s trustworthiness and reputation (Alhoori, Alvarez, Furuta, Miguel Mu, & Urbina, 2009), cited references (Thor, Marx, Leydesdorff,&Bornmann, 2016), and various altmetrics (Thelwall, Haustein, Larivière, & Sugimoto, 2013) with the goal of improving accuracy, diversity, novelty, and serendipity (Ge, Delgado-Battenfeld, &Jannach, 2010). The system will begin similarly, using meta-data of articles, such as title, abstract, keywords, and tags, to recommend venues, but will diverge into an analysis of explicit user-provided ratings. These experiments will use a hybrid approach implementing both collaborative filtering and content-based filtering. In addition, other factors that affect researchers’ choices will be considered, such as budget availability and the ability to travel in cases such as conferences or workshops.

**REFERENCES**

1. Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-based systems*, *46*, 109-132.
2. Errami, M., Wren, J. D., Hicks, J. M., & Garner, H. R. (2007). eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic acids research*, *35*(suppl\_2), W12-W15.
3. Schuemie, M. J., &Kors, J. A. (2008). Jane: suggesting journals, finding experts. *Bioinformatics*, *24*(5), 727-728.
4. Ricci, F., Rokach, L., &Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). springer US.
5. Pham, M. C., Cao, Y., Klamma, R., &Jarke, M. (2011). A clustering approach for collaborative filtering recommendation using social network analysis. *J. UCS*, *17*(4), 583-604.
6. Yang, Z., Yin, D., & Davison, B. D. (2014, August). Recommendation in Academia: A joint multi-relational model. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 566-571). IEEE.
7. Yang, Z., & Davison, B. D. (2012, December). Venue recommendation: Submitting your paper with style. In *Machine learning and applications (ICMLA), 2012 11th international conference on* (Vol. 1, pp. 681-686). IEEE.
8. Luong, H., Huynh, T., Gauch, S., Do, L., & Hoang, K. (2012). Publication venue recommendation using author network’s publication history. *Intelligent Information and Database Systems*, 426-435.
9. Asabere, N. Y., Xia, F., Wang, W., Rodrigues, J. J., Basso, F., & Ma, J. (2014). Improving smart conference participation through socially aware recommendation. *IEEE Transactions on Human-Machine Systems*, *44*(5), 689-700.
10. Wongchokprasitti, C., Brusilovsky, P., & Parra-Santander, D. (2010). Conference Navigator 2.0: community-based recommendation for academic conferences.
11. Asabere, N. Y., Xia, F., Wang, W., Rodrigues, J. J., Basso, F., & Ma, J. (2014). Improving smart conference participation through socially aware recommendation. *IEEE Transactions on Human-Machine Systems*, *44*(5), 689-700.
12. Hornick, M., & Tamayo, P. (2012). Extending recommender systems for disjoint user/item sets: The conference recommendation problem. *IEEE Transactions on Knowledge and Data Engineering*, *24*(8), 1478-1490.
13. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.
14. Adomavicius, G., &Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, *17*(6), 734-749.
15. Lemarchand, G. A. (2012). The long-term dynamics of co-authorship scientific networks: Iberoamerican countries (1973–2010). *Research Policy*, *41*(2), 291-305.
16. Yang, Z., & Davison, B. D. (2012, June). Distinguishing venues by writing styles. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (pp. 371-372). ACM.
17. Xia, F., Chen, Z., Wang, W., Li, J., & Yang, L. T. (2014). Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 364-375.
18. Huynh, T., & Hoang, K. (2012). Modeling collaborative knowledge of publishing activities for research recommendation. *Computational collective intelligence. Technologies and applications*, 41-50.
19. Alhoori, H., &Furuta, R. (2011). Understanding the dynamic scholarly research needs and behavior as applied to social reference management. *Research and advanced technology for digital libraries*, 169-178.
20. Alhoori, H., &Furuta, R. (2013, September). Can social reference management systems predict a ranking of scholarly venues?. In *International Conference on Theory and Practice of Digital Libraries* (pp. 138-143). Springer, Berlin, Heidelberg.
21. Alhoori, H. (2016, June). How to identify specialized research communities related to a researcher's changing interests. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*(pp. 239-240). IEEE.
22. Boukhris, I., &Ayachi, R. (2014, November). A novel personalized academic venue hybrid recommender. In *Computational Intelligence and Informatics (CINTI), 2014 IEEE 15th International Symposium on* (pp. 465-470). IEEE.
23. Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J., &Riedl, J. (1999, July). Combining collaborative filtering with personal agents for better recommendations. In *AAAI/IAAI* (pp. 439-446).
24. Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, *8*(2), 295-309.
25. Silva, T., Ma, J., Yang, C., & Liang, H. (2015). A profile‐boosted research analytics framework to recommend journals for manuscripts. *Journal of the Association for Information Science and Technology*, *66*(1), 180-200.
26. Schein, A. I., Popescul, A., Ungar, L. H., &Pennock, D. M. (2002, August). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 253-260). ACM.
27. Protasiewicz, J., Pedrycz, W., Kozłowski, M., Dadas, S., Stanisławek, T., Kopacz, A., &Gałężewska, M. (2016). A recommender system of reviewers and experts in reviewing problems. *Knowledge-Based Systems*, *106*, 164-178.
28. Minkov, E., Charrow, B., Ledlie, J., Teller, S., &Jaakkola, T. (2010, October). Collaborative future event recommendation. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 819-828). ACM.
29. Medvet, E., Bartoli, A., &Piccinin, G. (2014, November). Publication venue recommendation based on paper abstract. In *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on* (pp. 1004-1010). IEEE.
30. Luong, H. P., Huynh, T., Gauch, S., & Hoang, K. (2012). Exploiting Social Networks for Publication Venue Recommendations. In *KDIR* (pp. 239-245).
31. Kuruppu, P. U., & Gruber, A. M. (2006). Understanding the information needs of academic scholars in agricultural and biological sciences. *The Journal of Academic Librarianship*, *32*(6), 609-623.
32. Koren, Y., Bell, R., &Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8).
33. Küçüktunç, O., Saule, E., Kaya, K., &Çatalyürek, Ü. V. (2013, July). TheAdvisor: a webservice for academic recommendation. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (pp. 433-434). ACM.
34. Klamma, R., Cuong, P. M., & Cao, Y. (2009). You never walk alone: Recommending academic events based on social network analysis. *Complex Sciences*, 657-670.
35. Kang, N., Doornenbal, M. A., &Schijvenaars, R. J. (2015, September). Elsevier journal finder: recommending journals for your paper. In *Proceedings of the 9th ACM Conference on Recommender Systems* (pp. 261-264). ACM.
36. Küçüktunç, O., Saule, E., Kaya, K., &Çatalyürek, Ü. V. (2012). Recommendation on academic networks using direction aware citation analysis. *arXiv preprint arXiv:1205.1143*.
37. Järvelin, K., &Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422-446.
38. Huynh, T., & Hoang, K. (2012). Modeling collaborative knowledge of publishing activities for research recommendation. *Computational collective intelligence. Technologies and applications*, 41-50.