Journal of AI and Data Mining
Published online:

A Recommendation System for Finding Experts in Online Scientific Communities

S. Javadi, R. Safa, M. Azizi, and S. A. Mirroshandel*

Department of Computer Engineering, University of Guilan, Rasht, Iran.

Received 02 November 2019; Revised 27 March 2020; Accepted 24 June 2020
Corresponding author: mirroshandel@guilan.ac.ir (S. A. Mirroshandel).

Abstract
Online scientific communities are the bases that publish books, journals, and scientific papers, and help promote the knowledge. The researchers use the search engines in order to find the given information including scientific papers, an expert to collaborate with, and the publication venue, but in many cases, due to the search by keywords and lack of attention to the content, they do not achieve the desired results at the early stages. Online scientific communities can increase the system efficiency to respond to their users utilizing a customized search. In this paper, using a dataset including bibliographic information of the user's publication, the publication venues, and other published papers provide a way to find an expert in a particular context, where the experts are recommended to a user according to his/her records and preferences. In this way, a user request to find an expert is presented with the keywords that represent a certain expertise, and the system output will be a certain number of ranked suggestions for a specific user. Each suggestion is the name of an expert who has been identified appropriate to collaborate with the user. In evaluation using the IEEE database, the proposed method reaches an accuracy of 71.50% that seems to be an acceptable result.

Keywords: *Big Scholarly Data, Online Scientific Communities, Recommender Systems, Expert Finding Systems, IEEE.*

1. Introduction

Today, we witness the increasing growth of information resources in the web, and access to information is an important challenge for the users. Search engines have also been developed to meet these needs but the volume of data retrieved by the system is high, and finding the relevant information to the user's query is very difficult [1]. The most common problem of most web search systems is the lack of attention to the difference between the users' interests and retrieving the same results for the same queries [1]. Customized search due to providing the results according to the users' interests plays a significant role in providing their required information [1, 2].

In online scientific communities, the researchers use the search engines to find scientific papers, an expert to collaborate with, and the publication venue, but in many cases, due to the keyword-based search and lack of attention to the content, they do not achieve the desired results at the early stages. This paper focuses on the Expert Finding System (EFS). The general process of EFS begins with collecting data and other elements that can be used to determine skill areas. Upon identifying the areas of expertise, these systems apply different techniques to calibrate the experts on a particular topic. The process of seeking expertise in an auto EFS is quite similar to what the humans do. The only difference is that EFS can be fast and more accurate, in addition to satisfying the users' requirements. EFS has the ability to extract different experts and fields of expertise from big and complex datasets in comparison with individual analysis [3]. In addition, getting answers very fast and easy, and effective communications for the users are very important for the users' collaboration, and EFS is used to find those who are good for collaboration [4, 28].

Combining the expertise of a team of researchers can often lead to better results than an individual work. If we know that each researcher is an expert to what extent and in what areas of expertise, we can potentially use this knowledge to find the researchers with appropriate expertise and cooperation suggestion [4, 5]. Academic EFS are developed for different tasks, such as finding paper reviewers, supervisors, similar experts, university–industry collaborators, and research collaborators. These systems can also help the universities in managing their knowledge assets and finding the gap in specific areas [6].

In this work, we show that how metadata is used to identify specialized areas, and then using the Information Retrieval methods and studying the available techniques in this field, a method will be presented to recommend an appropriate expert. The proposed method is applied to the IEEE dataset, which is considered as a standard scientific database. We also used the real-world data to evaluate the system performance.

In Section 2 of this paper, we discuss the research works related to the customized search and finding experts. In Section 3, the proposed model is presented in details. The configuration, dataset, pre-processing, applied technologies, and evaluation are described respectively in Sections 4. Finally, the presentations as well as the future works are summarized in Section 5.

2. Related works

The term "Big Scholarly Data" that is assigned for the rapidly growing scholarly source of information includes the authors, papers, citations, digital libraries, academic social networks, etc., which brings new challenges with respect to the data management and analysis tasks. The main motivation of working on this problem is to mine the knowledge in order to provide better academic services for the researchers. An article titled "Big Scholarly Data: A Survey" discussed this topic by investigating the characteristics of this type of big data, as well as their applications [7]. Based on this study, as we can see in figure 1, a variety of big scholarly data is drawn from its various types of entities and numerous types of relations among these entities, which makes it a complex system. The networks with such characteristics are hybrid ones that allow us to illustrate some general properties of the scholarly environment.

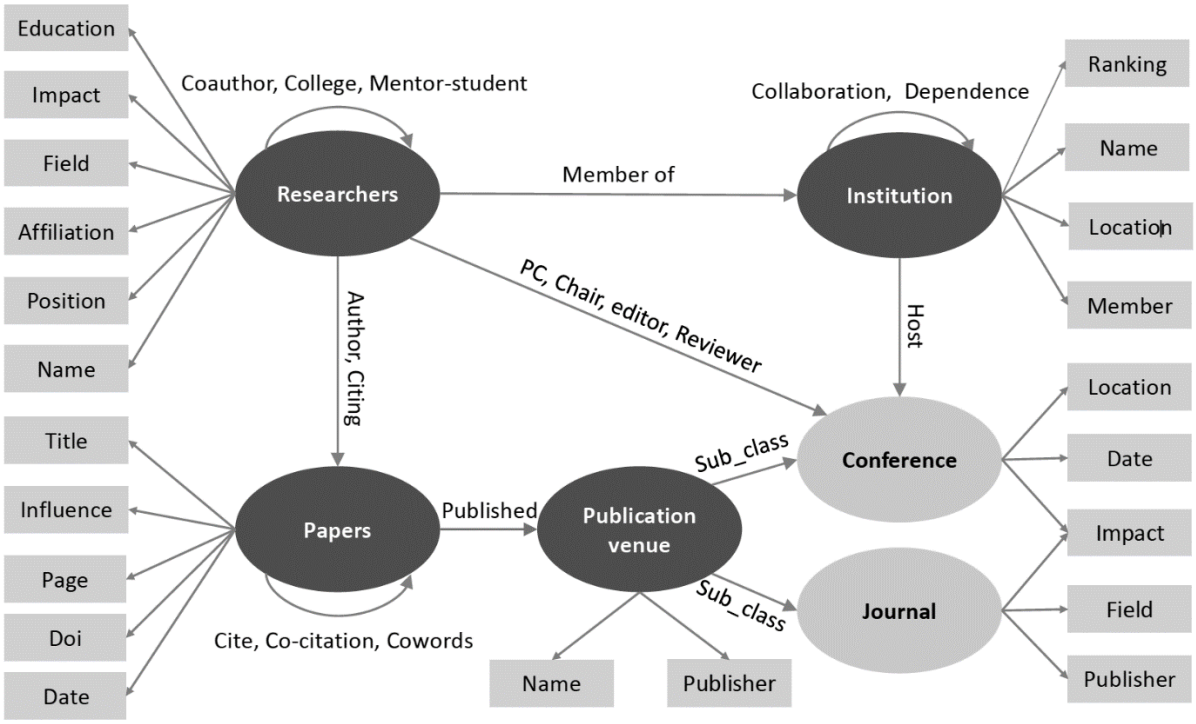

Figure 1. Major entities and their relationships in Big Scholarly Data [7].

Due to the rapid growth of information in online scientific communities, the researchers use search engines in order to find their needs. Search in these systems has problems, and often the users do not

reach their desired results at the right time. Hence, the need to customize search is felt in online scientific communities, and customized search tries to provide the search results according to the user's profile [8, 9]. In the following part, we will briefly describe an example of applying the recommender systems in the scholarly domain.

Given the enormous growth of the scientific conferences and journals, one of the important issues is to select the most appropriate venue to publish scientific papers, and this issue has been studied in some research works. One of the recent systems, in this case, is able to recommend the most appropriate venues to publish a paper written to the user by applying the concepts of social network analysis and content-based filtering methods. The proposed system in that study receives the author's identity and the written paper title in the input, and then using the defined database of library information identifies partners. Then after measuring the similarity of the paper title and publications of each of the authors' previous partners and identifying similar papers, conferences or journals to which the individual(s) have sent their similar paper are considered as the related venues. The operation can then be returned to the authors' partners and done more. The assessment using the real-world data also showed its good performance in providing the final effective recommendations [10].

Today, the demand for knowledge management has been increased. One of the important factors of knowledge management is to find a person with a high level of expertise in a specific field. The conventional way to do this is based on the relationship between the individuals. Hence, a systematic method is required to customize information filtering [11]. In a study in this field, search performance was enhanced by matching the users' historical preference on a set of items with similar patterns using collaborative filtering [12]. On the other side, some authors believe that the majority of studies in the field of scientific recommendations are limited to peer homogeneous networks. The outcome of the work indicates that more levels of social network proximity such as location can affect the intent of the researchers to collaborate [13, 14]. Therefore, only the use of network-based features cannot lead to a good recommendation. The key question of some research works is that how we can effectively and functionally extract and use the multiple features in bibliographic heterogeneous networks (which may affect scientific collaboration in the future implicitly). For example, the expertise of researchers, their ability as well as the frequency of collaboration are the important features that may affect the implementation of the collaboration recommendation. The importance of the bibliographic heterogeneous networks analysis and its applications have become a major trend in the recent years [14, 15]. QuickStep is a recommendation system of the scientific papers that uses the ontology of scientific papers' topics, computer science, and the classification created by Dimoz. The semantic interpretation of papers including finding the category in the ontology of scientific papers' topics is performed by the k-Nearest Neighbor (kNN) method [16]. Some researchers have proven that a combination of neighborhood and routing can have promising results [17]. Some additional factors such as the frequency of collaboration between two researchers may also affect the relationship between them [18].

A research work in this field by [19] has provided a combined method of five features of three heterogeneous networks that include the research topic network, the researcher collaboration network, and the institutions' network. In the mentioned research work, a language model-based method has been introduced which shows the expertise similarity in various aspects. In order to increase the accuracy of predictions, a new feature is made that combines the number of authors of a particular paper as well as the frequency of collaboration in the shortest route between two nodes. Finally, the ranking method Support Vector Machine (SVM) has been used to combine five features in the heterogeneous network (3-layer). The proposed method was used in the recommendation system in the ScholarMate platform, and finally, satisfactory experimental results were obtained. It was also noted that there were limitations, and to better illustrate the function of the proposed method, the experimental results should be tested on larger datasets to achieve more convincing evidence. This proposed method was used with a scientific social network platform in China to help the individuals search for a counterpart.

It has been reported that more features of other aspects like semantic similarity can be used to increase the prediction accuracy. For example, some researchers have considered local neighborhood and combined them with semantic similarities, and the results of experiments show its appropriate effect on the colleague recommendations [18]. In another study, a computational method has been presented to uncover the useful concepts of social networks and how to use these concepts in the design of an expert

recommendation system. The method of logical analysis was implemented on Ohloh large open-source community. It was observed that the nationality similarity, venue, and preferences of the programming language, as well as social recognition were essential in shaping the relationship between the members; although there was no guarantee for their collaboration [20]. Further, it was found that working with others could be more effective if they were close to each other, and in addition, the users considered consultation with familiar and trusted experts more [4]. We can say that finding the right experts to work is not only related to their authority of the topic but also the communication efficiency was very important.

Thus the recommender systems and the knowledge networks are very important to find an expert in scientific organizations and communities, and the right expert recommendation is not easy due to the need for the argument of complex heterogeneous networks as well as the need to consider the individuals' desire. Although a lot of efforts have been made over the past decade on the development of techniques for increasing the accuracy of recommendations customizing recommendations according to the individuals' motivation is still an open issue. While previous works focused on identifying the experts, customizing the selection of an expert was another method that was considered through a program from the social science aspect to model the users' motivation [21]. In this study, a recommender system has been proposed to customize the result through the users' motivation profile and their relationship.

An algorithm has been introduced by Wang et al. in order to find an expert called ExpertRank that assesses the expertise based on the relevance of documents and the expert validity in the online community [22]. The suggested EFS uses three indices for expert recommendation:

1. Find an expert based on custom disclosure of information: in this way, the experts openly proclaim their expertise in their profiles. It can be time-consuming, and the profile may remain constant with the development of the users' expertise.
2. Find an expert on the basis of documents: documents written or reviewed by an expert if available can be a good index and using text mining and information retrieval techniques, good results can be obtained.
3. Find an expert based on the analysis of social networks: Sociological studies show that the effect of social status plays an important role in the selections.

Moreover, the expert social activities can be examined by assistance from the Facebook or Twitter datasets. Yahoo! Answers, Stack Overflow and Quora have also attracted attention due to their applications to find experts in the question-answering systems. Recently, mining multimedia social networks have become more important. Multimedia sources like YouTube videos may have valuable information about the people's expertise [23, 24]. However, these social networks are not our main concern, as we focus on the academic communities and specifically bibliographic resources. To the best of our knowledge, there are a few papers that shape this area. One of the basic problem of the presented approaches is the lack of suitable datasets for evaluation.

The proposed methods generally use a set of pre-defined datasets that cause biases, while real-world datasets allow us to have a more realistic assessment. Another issue is the definition of new measures according to scientific EFS, which is different from other social network environments.

IEEE Xplore provides a free access to the required data and is an appropriate option for this work. In order to meet some of these demands, we intend to offer an effective approach with a light-weight method for expert recommendation. The system could generate a list of experts by analyzing the metadata from the IEEE scientific database. In the following, we will also meet the second issue by defining proper measures in the research scope.

3. Proposed model

According to the study, the main idea is that using the user publications' bibliographic information, the publication venue, and published papers, experts can be found in a specific field and presented to the user for collaboration. An overview of the presented model is shown in figure 2.

The system user is a person who seeks an expert for his current work in a specific field. The system input includes a *user name* and *field keywords*. A *user name* is an identity with which a person is known in the scientific community, and the *field keywords* are used to determine the scientific domain, which is specified by the user explicitly. The system output will be a certain number of rated recommendations. Each recommendation is the name of an expert that the system indicates appropriate for collaboration. In the proposed method, the system requires access to a dataset that

includes bibliographic information of scientific papers published in the recent years.

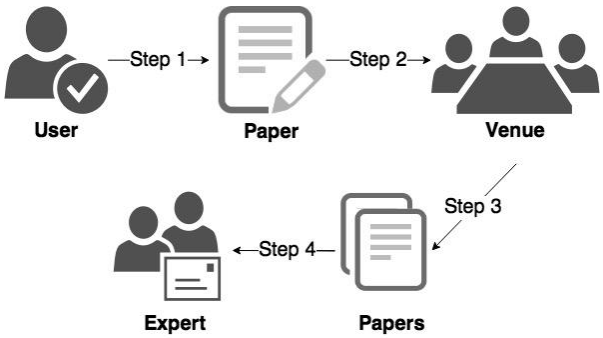

Figure 2. The proposed model overview.

The publication year data, venue of publication, authors' names, and keywords for each paper in the dataset are essential. Table 1 shows the main features and sub-features in the bibliography database. The system process with more details is shown in figure 3.

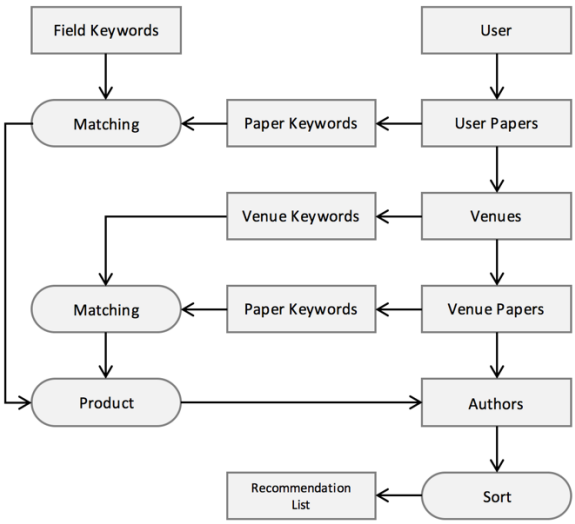

Figure 3. The system process.

First, using the authors' name, the user published papers in a limited period of time (a given date to the current date) are retrieved from the dataset; The papers are called *the user papers*. Also the period specified duration is called *the user papers' duration*. The reason for this time limitation is that the user research interests may change over time. Next, each of a user papers are given a weight that is obtained through measuring the similarity between the keywords of a paper and keywords of a field. We will discuss this in more details in the next section. In fact, the weight specifies that each of the user papers - that are his previous works - to what extent is related to his current work. Then the user papers are classified based on the venue of publication, and the total weight given to each category papers is assigned to the corresponding venue of publication. As a result, the venue of publication with more weight will be known as more relevant to the user's current work. The obtained venue of publication at this stage is called *the venue of publication of the user*. Then, for each venue of publication of the user, *the venue keywords* are obtained, which is a set of the keywords of the papers published in a limited period of time (a given date to the current date). The duration of this period of time is called *the venue keywords' duration*. The reason for this time limitation is to retrieve update keywords, that are likely to be related to the user requirement.

Next, using the venue of publications of the user, the papers published in the venue are retrieved in a specific time period for the same reason. The papers are called *the venue papers*. Also the period specified duration is called *the venue papers duration*. Then a weight is given to each venue paper that is obtained from the product of the paper venue weight - that was obtained before - and the weight obtained from measuring the similarity between the paper keywords and the venue keywords.

$$W_{venue\ paper} = W_{paper venue} \cdot S(paper\ keywords, venue\ keywords) \quad (1)$$

In this equation, W is an abbreviation for weight and S denotes the similarity calculation (or matching process) that will be explained in the next section.

Table 1. Main features and sub-features.

Main Feature	Sub-feature(s)
User	Paper(s)
Paper	Author(s), Keywords, Publication Venue
Venue	Papers, Keywords

It should be noted that the publication venue of each of these papers is available in the user publication venue so the weight has already been obtained. The weight obtained from measuring the similarity between a paper keywords and the venue keywords specifies that the paper to what extent is related to the publication venue. As mentioned earlier, the weight of each publication venue signifies its relationship with the user's current work. As a result, the product of the two weights provides a measure of determining the relationship between a paper and a user's current work. Then, the publication venue papers are classified according to the author, and the total weights given to each category papers are dedicated to the

corresponding author. As explained earlier, we can say that the weight signifies the relationship between an author expertise and the user's current work, and therefore, it is used as the expert recommendation measure.
In the last step, the authors are sorted based on the given weight, and a certain number with the most weight is provided as the experts list; The certain number is called *Recommendation List*.

3.1. Measuring similarity

The simplest measure of the similarity between the two keywords is equality that accordingly the similarity between the two same keywords is 1, and otherwise 0. However, this measure is not efficient because, for example, the keyword "iterative decoding" is more related to "channel coding" rather than "cloud computing" but the measure for both of them brings the value 0.
Since in the proposed method, the weight obtained from measuring the similarity between keywords plays a decisive role in the final recommendations, and the efficiency of the similarity measure is very important for the system, in the trade-off of cost and efficiency, the efficiency is taken into consideration. In the proposed method, to measure the similarity between the two keywords, we offer two different measures: *co-occurrence* and *shared neighbor*. The *co-occurrence* measure for two keywords signifies the number of papers that both keywords have appeared in them. *Shared neighbors*' measure for two keywords signifies the number of keywords that are neighbors, i.e., with each of them at least have appeared in a paper. The order of appearance of a keyword in a paper is clearly its presence in the keywords of the paper. In order to obtain each of these measures, we check the papers published in a limited period of time (a given date to the current date). The duration of this period of time is called *the similarity measure duration*. Table 2 shows the values of the similarity measure with a duration of two years for a few keyword pairs.

Table 2. Values of the similarity measures with a duration of two years for a few keyword pairs.

First keyword	Second keyword	Co-occurrence	Shared neighbors
iterative decoding	channel coding	105	365
	computational complexity	37	388
	equalisers	15	289
	multi-access systems	14	265
	galois fields	7	141
	underwater acoustic communication	5	236
	delays	3	362
	radiocommunication	3	297
	statistical analysis	2	373
	multiprocessing systems	2	262
	polynomial approximation	1	173
	amplitude modulation	1	173
	newton method	1	170
	multi-threading	1	142
	cloud computing	0	271
	hardware-software codesign	0	135
	mathematical morphology	0	63
	bipolar logic circuits	0	5
	passive solar buildings	0	0
	strontium alloys	0	0

The values of the defined measures are not normal. Eq. (2) is used to normalize the values, where S is the similarity measuring normalizing function, M is a measure of the similarity, and k_1 and k_2 are the keywords.

$$S(k_1,k_2)=\begin{cases}1 & ,k_1=k_2\\ 1-\dfrac{1}{M(k_1,k_2)+1} & ,k_1\neq k_2\end{cases} \quad (2)$$

In order to measure the similarity of two sets of keywords, two measures of central tendency have also been taken into account: one the mean pairwise similarity of sets' keywords, and another,

similarly, the median of pairwise similarity sets' keywords.

Since two different measures have been introduced to measure the similarity of two keywords (co-occurrence and shared neighbors), and two different measures have been considered to measure the similarity of two sets of keywords (the mean and median), there are totally four methods to obtain the similarity of two sets of keywords. As it will be discussed, after the four methods' assessment, the best result is obtained using the measures of co-occurrence and the mean.

4. Experimental results

In this section, we first address providing the required dataset and its pre-processing as well as the characteristics of the dataset. Then some challenges will be reviewed.

4.1. Configuration

Table 3 shows the parameters and their various values; Given this data, in total, there are eight different states of assessment. To each case, a *configuration number* has been given, as seen in table 4.

The values of other parameters of the system that were kept constant in assessing all the test cases are given is table 5.

As we will discuss later, the time complexity of the recommendation algorithm is an indirect function of the parameters described in table 5. On the one hand, choosing a large value for these parameters will make the system slow. On the other hand, too small of a value will lead to poor results. The values shown in table 5 are chosen experimentally to reach a good balance between latency and the quality of recommendations.

Table 3. Assessment configuration parameters.

Parameter	Value
Similarity measure	Co-occurrence
	Shared neighbors
Central tendency measure	Mean
	Median
Keywords type	Controlled index terms
	Thesaurus terms

Table 4. Different configurations of the assessment.

No.	Keywords	Measure of measuring similarity	Central tendency measure
0	Controlled index terms	Co-occurrence	Mean
1	Controlled index terms	Co-occurrence	Median
2	Controlled index terms	Shared neighbors	Mean
3	Controlled index terms	Shared neighbors	Median
4	Thesaurus terms	Co-occurrence	Mean
5	Thesaurus terms	Co-occurrence	Median
6	Thesaurus terms	Shared neighbors	Mean
7	Thesaurus terms	Shared neighbors	Median

Table 5. Constant parameters in the assessment.

Parameter	Value
Duration of a user papers	2 years
Duration of keywords of the publication venue	1 years
Duration of the publication venue papers	2 years
Duration of the similarity measures	2 years

4.2. Dataset

For a reliable assessment of the proposed method, we needed a large set of real-world data that included the bibliographic information of the scientific papers published in the recent years. One available option for the dataset was Digital Bibliography & Library Project (DBLP). DBLP is a computer science bibliography database that the University of Trier in Germany hosts it. Since all DBLP data is stored in an XML file, access to it is simple. The dataset contains the data from the year of publication, the venue of publication, and the name of the authors, but no keywords data [25]. For this reason, we did not use it. We chose IEEE Xplore Digital Library for our work. This digital library includes the papers of computer science, electrical and electronic engineering that have been published by IEEE (institute of electrical and electronics engineers) and other partner publishers. IEEE Xplore provides web access to more than 3 million scientific and technical documents, and about twenty thousand new documents are added to it monthly. The content of IEEE Xplore contains the following [26]:

- More than 170 magazines
- More than 1,400 conference proceedings
- More than 5,100 technical standards
- About 2,000 books
- More than 400 training courses

The most important data available for IEEE Xplore documents for free are listed below:

- Topic
- Author(s)
- Affiliation of the author(s)
- Keywords
- Publication venue
- Type of document (conferences, magazines, books, early access, standards or training courses)
- Publishers (IEEE, AIP, IET, AVS or IBM)
- Publication year
- Abstract
- ISBN
- ISSN
- Digital Object Identifier (DOI)

Since IEEE Xplore provides a free access to the required data for many papers available in the Digital Library - the year of publication, the venue of publication, the name of the author(s), and keywords - it is an appropriate option for our work. Although IEEE Xplore provides the required data, unlike DBLP, downloading the dataset as a file is not possible. Thus it was necessary to somehow fetch the data of the digital library and stored in a local database. The search gateway of IEEE Xplore provides an application programming interface (API) to search the database of the digital library [27]. Although to assess this study, only a few years of publications are enough, we decided to fetch and store IEEE Xplore data if required to be used in the future works.
IEEE Xplore search gateway responses are in the form of XML, which is readable for both the humans and machines. It is an open standard. It should be noted that in the fetched data from the IEEE Xplore search gateway, there are two types of keywords: "thesaurus terms" and "controlled index terms". We used both types of keywords for the assessment. As would be discussed in the future, after assessing both types of keywords individually, we found that by using the controlled index terms, the best result was achieved.

4.3. Preprocessing

In the IEEE Xplore database, for a conference in different years, a distinctive venue is considered. The following example has been extracted from the real data of the database:

- "Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on" (Publication Number: 6204368)
- "Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on" (Publication Number: 6853543)

As it can be seen, for a single conference in two different years, two different publication numbers have been intended. However, since the scope of a conference, held in different years, is constant, we desire to consider all of them as a unique venue. Therefore, before using the dataset, we should solve this problem. Due to the large number of available documents, manual review and correction is not feasible. Thus by examining a considerable number of documents we identified a specific pattern. In all the examined cases, the same as the previous example, the year of holding a conference appears in a part of the conference topic, which is separated by a comma. Knowing this pattern, the correction can be performed automatically.
If a comma has not appeared in the publication topic, there is no need to review; otherwise, we separate the topic into parts and looking for a year (specifically, an integer between 1800 and 2099). If there is a year in a part, the part is excluded from the topic. Also, if several different publications' topics become the same after correction, the publication number of all of them becomes the same as well. As an example, both the previous titles are changed to Biomedical and Health Informatics (BHI). After applying pre-processing, the number of unique venues was reduced from 27,102 to 17,252, which made the assessment more accurate.

4.4. Challenges

A major challenge in the implementation was the high complexity of computation of the keywords' similarity. Due to the time-consuming similarity computation between two keywords and the large number of keywords that should be measured in the system process for each run, online calculating is not practical and affordable. Thus the pairwise similarity of all keywords of the papers published in the specified duration was calculated and stored. In order to be able to do this in a reasonable time, we used a multi-processing method to write the program.
In this study, for the similarity measure duration of two years (2012 and 2013), two measures of co-occurrence and shared neighbours were calculated and saved for the existing 38,531,031 keyword pairs.

4.5. Evaluation

We considered 2014 as *the year of assessment*, which means the system access to the data related to the papers published by the end of 2013 is limited.

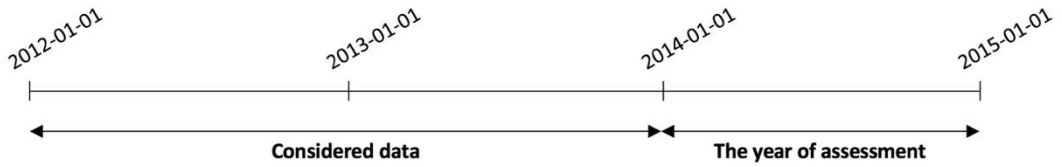

Figure 4. Dataset Segmentation.

In order to select a test case, first, among all the authors who have published a paper in the assessment year, one is randomly selected if in the ten years leading up to the year of assessment (excluding the year of assessment), at least four papers have been published. Then among all the author's published papers during the assessment year, one is randomly selected if both types of thesaurus and controlled index terms are available for the paper. For the assessment, the author's name selected and the keywords selected are used as the system input - a *user name* and the *field keywords* - and the output of the system - the name of recommended experts - are saved for each test case.

A major challenge in assessing the proposed method is the lack of a clear measure to judge the quality of a recommendation. Unfortunately, the existing approaches use different parameters for the expert recommendation task, and their methods are not easily reproducible. Additionally, their utilized dataset is not publicly available. As a result, we were unable to apply our method on their datasets to have a fair comparison. Thus, the only way available to assess was manual review recommendations by the human. In order to do this, we asked three experts to label the data manually. To assess each test case, with the search for the *user name* and the name of any expert recommended on the web, the interests and context of the recent works were found, and then, using common sense, the quality of recommendations was judged, and each item was labeled *good* or *bad*. In order to evaluate the performance of the system in various configurations, we implemented the system for twenty test cases with each of the possible eight configurations. One of the test cases, for example, is listed below:

- User name: Marques, E
- Controlled index terms:
    - high level synthesis
    - c language
    - field programmable gate arrays
    - hardware description languages
- Thesaurus terms:
    - clocks
    - radiation detectors
    - hardware
    - benchmark testing
    - pipeline processing
    - field programmable gate arrays

The system recommendations for the above test case, with configuration 0, along with the assessment results, are presented below:

- Luk, W. – Good
- Kumar, A. – Good
- Amano, H. – Good
- Maruyama, T. – Not Good
- Cheung, P.Y.K. – Good
- Stroobandt, D. – Good
- Bruneel, K. – Good
- Betz, V. – Good
- Benkrid, K. – Good
- Chow, P. – Good

The results of the assessment are shown in table 6.

Table 6. Assessment results.

Configuration number	Accuracy (%)
0	71.50
1	67.00
2	65.00
3	65.00
4	64.00
5	61.00
6	62.00
7	62.00
Average	**64.60**

The accuracy of the system in any configuration, the percentage of recommendations with the label good in that configuration, is for twenty test cases.

As it can be seen, the configuration number zero has provided the best result with an accuracy of 71.50%.

4.6. Analysis

The first phase of the algorithm extracts the authors, i.e. potential recommendations with their

associated weights. The time complexity of this phase is $O(n)$, where n is the total number of retrieved papers. In other words, n equals the number of *User Papers* (published in the *User Paper Duration*) plus the number of *Venue Papers* (published in the *Venue Paper Duration*). This phase can be done with space complexity of $O(1)$ because we only need to work with one paper at a time.

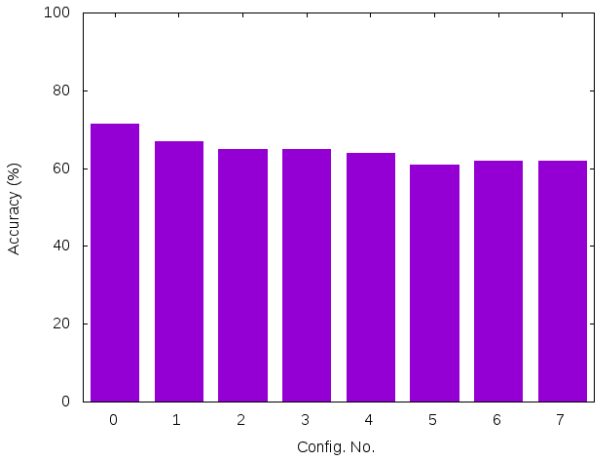

Figure 5. Accuracy for different configurations.

The main operation in the first phase is matching, i.e. calculating the similarity of two sets of keywords. The time complexity of matching itself is $O(m.n)$, where m and n are the sizes of the two keyword sets. For the first step of this phase, the two sets are *User Paper* keywords and *Field Keywords*. For the second step, these are *Venue Paper* keywords and *Venue Keywords*. *Field Keywords* is expected to be a small set, as the keywords are input by the user. The system can also enforce a reasonable limit. The size of *User*/*Venue Paper* keywords is the number of keywords appeared in the paper, which is usually a small number. *Venue Keywords* is a relatively large set whose size is controlled by the *Venue Keywords Duration* parameter. The main operation for matching is calculating the similarity of two keywords. Since the similarity of all keyword pairs is calculated and cached in advance, matching is very fast in practice. The space complexity of this operation is also $O(1)$.

Both similarity measures we employed, i.e. *co-occurrence* and *shared neighbors*, have the time complexity of $O(n)$, where n is the total number of papers published in the *Similarity Measure Duration*. Since n can be very large, similarity calculation is costly in practice. However, once the similarity of two keywords is calculated, it can be reused for a long time, because the relative similarity of keywords used in the literature does not change very often. As an example, "iterative decoding" is closer to "channel coding" than to "cloud computing" – this is true now and will probably be true the next year as well. Also, new keywords are coined only occasionally. We exploit these properties of keywords to create a cache of keyword similarities that contains all keywords appeared in the papers published in the *Similarity Measure Duration*.

Creating such a cache is costly but it only needs to be updated infrequently, e.g. yearly. This cache significantly speeds up our recommendation system. The space complexity of the c*o-occurrence* method, the one finally selected, is $O(1)$. As for the *shared Neighbors* method, the space complexity is $O(m)$, where m is the number of distinct keywords appearing in the papers published in the *Similarity Measure Duration*.

The second phase of the algorithm returns the final result: k recommendations or top k ordered authors. This is a partial sorting problem. Using a heap-based solution, it can be done in $O(n.\log(k))$, where n is the total number of retrieved authors, and k is the number of recommendations. Since k is a pre-set constant, this is equivalent to $O(n)$.

5. Conclusions and future works

There are various opportunities in the field of providing the scientific recommendations including the paper recommendation, the venue of publication, and collaboration that the majority of topics discussed in this paper have been on the scientific colleague recommendations. It is clear that the scientific colleague recommendations' system helps the researchers communicate with other experts by identifying them and respond to a part of their research needs. Also the other main motivation for working on the problem of colleague recommendations is capability management and its widely used applications in various fields.

In this study, using a dataset containing the data of the publication year, the venue of publication, authors' name, and keywords of the published papers in the recent years, we could provide a way to find an expert in a specific field, where the experts are recommended according to a user records. For access to the required data, by fetching the data from the IEEE Xplore digital library, the database containing the data on more than three million scientific and technical documents have been provided. In addition, by reviewing and processing of the database, the data quality has been improved. Also by the expensive calculation of keywords' similarity measures, a valuable

dataset of the similarity of more than 38 million scientific keyword pairs has been obtained. This dataset can also be useful for future research works. The proposed system is able to find the experts in a specific field and introduce them to a user for collaboration using the bibliographic information of a user publication, venue information, and other papers published in that venue. Of course, it should be noted that a problem unresolved in the proposed method is cold start. This means that the system is only used for the users with the paper publication history and for those who have not published a paper yet; it cannot provide a recommendation, and this can be a topic for future research works.

Since assessing the recommendations is possible only with human judgment, a major challenge for us is the assessment in this study. Totally, after assessing using the real-world data, the proposed method has shown an accuracy of 71.50 percent, which is an impressive result, and can be a proof-of-concept for similar implementations. It shows that based on the experts' validation, in 71.50 percent of cases, the model achieves 100 percent relevant results; this indicates that if we were less strict in the selection phase, accuracy would be improved. This is in a situation where we do not limit the scientific domain. In the future, by assessing more test cases, we can better estimate the accuracy of the proposed method. Also by changing the parameters that we kept constant in the assessment, we can see their effect on the accuracy of the system.

Thus some future opportunities available to continue the research works include strengthening the accuracy of the method to recommend the experts to use the other bibliographic features available in databases (such as Abstract), using advanced algorithms based on graphs, provide a solution for cold start problem and the ability to help the less experienced individuals, changing the related parameters discussed in table 5, implementation of the proposed method on the other datasets and turning it into a real system, using more data for a more comprehensive assessment of the proposed system, using more test cases for a more accurate estimation of the proposed system performance, changing the proposed system assessment mechanism to gain more knowledge of the problem, and finally, using the general system mechanism aimed to determine the other scientific recommendations (including the paper recommendation).

References

[1] Sathiyabama, M. T., & Vivekanandan, K. (2011). Personalized Web Search Techniques-A Review. Global Journal of Computer Science and Technology.

[2] Dadiyala, C., Patil, P., & Agrawal, G. (2013). Personalized web search. Int. J. Adv. Res. Comput. Sci. Softw. Eng, vol. 3, no. 6.

[3] Afzal, M. T., & Maurer, H. A. (2011). Expertise Recommender System for Scientific Community. J. UCS, vol. 17, no. 11, pp. 1529-1549.

[4] Zhan, Z., Yang, L., Bao, S., Han, D., Su, Z., & Yu, Y. (2011). Finding appropriate experts for collaboration. In International Conference on Web-Age Information Management. Springer, Berlin, Heidelberg.

[5] Davoodi, E., Kianmehr, K., & Afsharchi, M. (2013). A semantic social network-based expert recommender system. Applied Intelligence, vol. 39, no. 1, pp. 1-13.

[6] Husain, O., Salim, N., Alias, R. A., Abdelsalam, S., & Hassan, A. (2019). Expert Finding Systems: A Systematic Review. Applied Sciences, vol. 9, no. 20, pp. 4218-4250.

[7] Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. IEEE Transactions on Big Data, vol. 3, no. 1, pp. 18-35.

[8] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In Recommender systems handbook. Springer US.

[9] Zanker, M., Felfernig, A., & Friedrich, G. (2011). Recommender systems: an introduction.

[10] Safa, R., Mirroshandel, S., Javadi, S., & Azizi, M. (2018). Venue Recommendation Based on Paper's Title and Co-authors Network. Journal of Information Systems and Telecommunication, vol. 6, no. 1, pp. 33-40.

[11] Yukawa, T., Kasahara, K., Kato, T., & Kita, T. (2001). An expert recommendation system using concept-based relevance discernment. In Tools with Artificial Intelligence, Proceedings of the 13th International Conference on (pp. 257-264). IEEE.

[12] Rohini, U., & Ambati, V. (2005). A collaborative filtering based re-ranking strategy for search in digital libraries. In International Conference on Asian Digital Libraries (pp. 194-203). Springer, Berlin, Heidelberg.

[13] Yang, C., Ma, J., Silva, T., Liu, X., & Hua, Z. (2013). A multilevel information mining approach for expert recommendation in online scientific communities. The Computer Journal, vol. 58, no. 9, pp. 1921-1936.

[14] Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., & Han, J. (2011). Co-author relationship prediction in heterogeneous bibliographic networks. In Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on (pp. 121-128). IEEE.

[15] Lee, D. H., Brusilovsky, P., & Schleyer, T. (2011). Recommending collaborators using social features and mesh terms. Proceedings of the Association for Information Science and Technology, vol. 48, no. 1, pp. 1-10.

[16] Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. ACM Transactions on Information Systems (TOIS), vol. 22, no. 1, pp. 54-88.

[17] Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. Information processing & management, vol. 41, no. 6, pp. 1462-1480.

[18] Han, S., He, D., Brusilovsky, P., & Yue, Z. (2013). Coauthor prediction for junior researchers. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 274-283). Springer, Berlin, Heidelberg.

[19] Yang, C., Sun, J., Ma, J., Zhang, S., Wang, G., & Hua, Z. (2015). Scientific collaborator recommendation in heterogeneous bibliographic networks. In System Sciences (HICSS), 2015 48th Hawaii International Conference on (pp. 552-561). IEEE.

[20] Hu, D., & Zhao, J. L. (2008). Expert recommendation via semantic social networks. ICIS 2008 Proceedings.

[21] Fazel-Zarandi, M., Devlin, H. J., Huang, Y., & Contractor, N. (2011). Expert recommendation based on social drivers, social network analysis, and semantic data representation. In Proceedings of the 2nd international workshop on information heterogeneity and fusion in recommender systems (pp. 41-48). ACM.

[22] Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. Decision Support Systems, vol. 54, no. 3, pp. 1442-1451.

[23] Nikzad–Khasmakhi, N., Balafar, M. A., & Feizi–Derakhshi, M. R. (2019). The state-of-the-art in expert recommendation systems. Engineering Applications of Artificial Intelligence, vol. 82, no. 1, pp. 126-147.

[24] Amato, F., Cozzolino, G., & Sperlì, G. (2019). A hypergraph data model for expert-finding in multimedia social networks. Information, vol. 10, no. 6, pp. 183-193.

[25] Ley, M. (2009). DBLP: some lessons learned. Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1493-1500.

[26] IEEE Xplore (2020), Available: https://ieeexplore.ieee.org/Xplore/home.jsp

[27] IEEE Xplore API Portal (2020), Available: https://developer.ieee.org.

[28] Tahmasebi, M., Fotouhi, F., & Esmaeili, M. (2019). Hybrid Adaptive Educational Hypermedia Recommender Accommodating User's Learning Style and Web Page Features. Journal of AI and Data Mining, vol. 7, no. 2, pp. 225-238.