IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2019, accepted January 16, 2019, date of publication January 24, 2019, date of current version February 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2894680

A Pattern-Based Academic Reviewer Recommendation Combining Author-Paper and Diversity Metrics

MUSA IBRAHIM MUSA ISHAG[1], (Student Member, IEEE), KWANG HO PARK[1], JONG YUN LEE[1], AND KEUN HO RYU[2], (Life Member, IEEE)
[1]College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, South Korea
[2]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam

Corresponding author: Keun Ho Ryu (khryu@tdtu.edu.vn)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning under Grant 2017R1A2B4010826 and Grant 2017R1D1A1A02018718, and in part by the Research Year of Chungbuk National University in 2016.

ABSTRACT With the rapid increase of publishable research articles and manuscripts, the pressure to find reviewers often overwhelms the journal editors. This paper incorporates the major entity level metrics found in the heterogeneous publication networks into a pattern mining process in order to recommend academic reviewers and potential research collaborators. In essence, the paper integrates authors' h-index and papers' citation count and proposes a quantification to account for the author diversity into one formula duped impact to measure the real influence of a scientific paper. Thereafter, this paper formulates two kinds of target patterns and mines them harnessing the high-utility itemset mining (HUIM) framework. The first pattern, researcher-general topic patterns (RGP), is a pattern that includes only researchers; whereas, the researcher-specific topic patterns (RSP) is comprised of combinations of researchers and keywords that summarize their niche of expertise. The HUI algorithms of Two Phase, IHUP, UP-Growth, FHM, FHN, HUINIV-Mine, D2HUP, and EFIM were compared on two real-world citation datasets related to Deep Learning and HUIM, in addition to the open source mushroom dataset. The EFIM algorithm showed good performance in terms of run time and memory usage. Consequently, it was then used to mine the patterns within the proposed framework. The discovered patterns of RGP and RSP showed high coverage, proving the efficiency of the proposed framework.

INDEX TERMS High utility itemset mining, recommender system, expert finding, scholarly big data, reviewer assignment.

I. INTRODUCTION

Peer review is the corner stone of the process that leads to a high quality academic publication. During the review, journal editors and conference chairs seek help from subject matter experts (SMEs) in order to decide on accepting or rejecting an article in an objective manner [46], [47]. For example, the list of experts that help the conference chair is known as technical program committee (TPC). Likewise, journal editors consult a list of well-known SMEs. The role of these experts is to provide objective critics and recommendations to accept or reject a paper. Therefore, the quality of conference proceedings or journal articles depends on the rational reviews given by the members of the TPC and the reviewers respectively [48]. Journal editors and conference chairs are therefore, always looking for qualified academic reviewers to help in the peer-reviewed publications. After identifying the list of reviewers, journal editors and conference chairs assign papers to reviewers based on their knowledge. This assignment is the back bone of the process. Incorrect assignment of papers to reviewers will have significant negative effect on the quality of reviews [49], [50].

One typical criterion for selecting a reviewer is his expertise in a specific field of research. This subject matter mastery can be measured through the individual's own contributions in that filed which is typically the collection of his publications. The reviewer recommendation

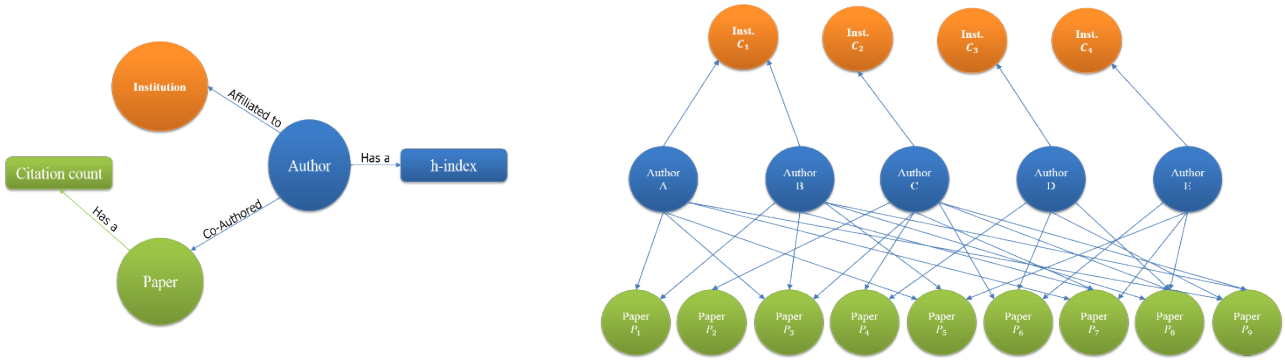


FIGURE 1. **(a) a conceptual representation of a heterogeneous academic social network, (b) Example of a heterogeneous academic network extracted from a citation dataset consisting of 9 research papers.**

problem [14], [31], [38] also known as referee assignment [32], or reviewer assignment (RA) [31], can be viewed as a sub-filed of the expert finding problem [26]. It can be performed either manually [47], or through a Computerized system. However, because of the sheer amount of the academic publications, the task of recommending reviewers manually is becoming overwhelmingly tedious [51].

Existing approaches of reviewer recommendation fall into three different categories. The first category, known as profile centric, generates profiles of experts and match them to keywords that represent the papers to be reviewed [33], [34], [35]. The second category on the other hand, clusters the papers to be reviewed into groups based on their topic similarity and then assigns the reviewers [28], [39], [40]. The third and most recent category, formulates the problem as a pattern mining problem [2], [23], [47]. The pattern mining problem is a subtask of the association rule mining [1] where combinations of items frequently occurring together are found. However, the frequencies of items alone would make it hard to find combinations of none frequent items which have other importance. This has motivated the high utility itemset mining (HUIM) [4]. In this context individual items have importance in terms of their counts and unit prices. Although the existing studies of reviewer recommendation encode the publication and the reviewer as keywords, they have overlooked the main factors that influence the research impact. Namely, they consider neither the author level metrics, nor the diversity of the authors.

In this paper, motivated by the characteristics of HUIM that takes into account the importance of individual items, we propose a pattern-based reviewer recommendation framework based on author, paper, and diversity metrics. The approach extracts information related to scientific papers from both citation databases gathered from digital libraries [59] and academic social networks [54], [55]. The extracted information is used to model a heterogeneous academic social network with three different node types. The nodes represent authors, intuitions and papers. Figure 1 shows the conceptual representation of the heterogeneous academic network where the three different node types are:

1) Author, i.e. the person who writes a paper. Authors typically have profiles in academic social networks that shows their *h-index*.
2) Paper, which is written by the author and has a *citation count*, and
3) Institution, which represents the *affiliation* of the author.

Figure 2 shows an example of a heterogeneous academic social network built from a citation dataset that contains 9 papers. The nodes represent the three different entities with colours to distinguish them. For instance, the link between the blue node labelled *Author A* and the orange node labelled $Inst.C_1$ indicates that the author A is affiliated to the institution C_1. On the hand, the links from the two blue nodes *Author A* and *Author B* that point to the green node $paperP_1$ indicate that the paper P_1 is co-authored by *Author A* and *Author B*.

In order to recommend reviewers, this paper defines two kinds of patterns: researcher general-topic pattern (RGP) and researcher specific-topic pattern (RSP). To find such patterns, the paper extends the high utility itemset mining (HUIM) [17], [52]. Analogously to the HUIM, the paper incorporates the importance of individual items.

However, there are major differences between the HUIM and the patterns found in the proposed framework. Firstly, the items that make the transaction data are of different types. Secondly, unlike HUIM, the importance of an item depends on its type. Hence, the proposed framework uses item-level importance. For instance, if an item is of type author, its importance is different from another item that represent papers. Consequently, new definition are given on how to construct the impactful transaction dataset.

Table 1 shows an example of impactful transaction dataset generated from the network illustrated in Figure 2. While the importance of the papers in terms of prestige, citation count and diversity are explicitly shown in Table 1, the

TABLE 1. Impactful transactional dataset.

PID	Authors	Keywd	Div	CC	Prestige	Impact
P_1	A, B	X	1	20	20	27
P_2	C	-	1	15	15	18
P_3	A, B, C	X	2	10	20	30

Keywd : Keywords, **Div**: Diversity, **CC**: Citation Count

TABLE 2. Authors' h-index table.

Author	*h*-index
A	*2*
B	*5*
C	*3*
D	*1*
E	*4*

TABLE 3. Keyword frequency table.

Keyword	*frequency*
X	*1*
Y	*1*

importance of the authors and keywords are shown in Table 2 and Table 3 respectively.

The main contributions of the paper can be summarized as follows:

- The citation dataset of published papers is modelled as a heterogeneous academic network comprising of three different nodes: authors, papers and institutions.
- Introducing impact – a measure of a publication importance that takes into account the authors' h-index, collaboration diversity, and the citation count of the underlying paper. A high utility-pattern based framework for mining impactful researcher general-topic patterns (RGP) and researcher specific-topic patterns (RSP).
- An experimental prove of concepts using real-world citation data to prove the effectiveness of the proposed approach.

The structure of the paper is organized as follows: Section 2 presents the relevant literature; section 3 explains the main building blocks of the proposed framework along with a detailed explanation of the mining approach. Section 4 reports the details of a proof of concept case study and explains the results obtained. Section 5 concludes the work by summarizing the major parts and highlighting the future directions..

II. RELATED WORK

The scientific literature in this paper is related to two main research topics:

- High utility pattern mining
- Reviewer recommendation.

The following subsections briefly introduce the concepts related to these topics.

A. HIGH UTILITY PATTERN MINING

The pattern mining problem was proposed by [1] in order to recommend products from transactional data. However, due to the binary representation of items in the market-based data which ignores the importance of items, a new pattern known as high utility itemset mining (HUIM) was introduced [4]. The HUIM is formally defined as [58], [52]:

Let $I = \{i_1, i_2, \ldots, i_M\}$ be a finite set of items. Then a set $X \subset I$ is referred to as an itemset. Let $D = \{T_1, T_2, \ldots, T_N\}$ be a transaction database. Each transaction $T_i \in D$, with unique identifier TID is a subset of I. The internal utility $q\left(i_p, T_d\right)$ represents the quantity of item i_p in transaction T_d. The external utility $p\left(i_p\right)$ is the unit profit value of item i_p. The utility of an item i_p in transaction T_d is defined as $u\left(i_p, T_d\right) = p\left(i_p\right) \times q\left(i_p, T_d\right)$. The utility of itemset X in transaction T_d is defined as

$$u\left(X, T_d\right) = \sum_{i_p \in X \cap X \subset T_d} u\left(i_p, T_d\right) \tag{1}$$

The utility of itemset X in D is defined as

$$u\left(X\right) = \sum_{X \subset T_d \cap T_d \in D} u\left(X, T_d\right) \tag{2}$$

The transaction utility (TU) of a transaction T_d is defined as

$$TU\left(T_d\right) = u\left(T_d, T_d\right) \tag{3}$$

To perform HUIM, the user provided minimum utility threshold δ is defined as a percentage of the total TU values of the database. The minimum utility value however, is calculated as

$$min\_util = \delta \times \sum_{T_d \in D} TU\left(T_d\right) \tag{4}$$

An itemset X is called HUI if u (X) $\geq$ min_util.

Given a transaction database D, the task of HUIM is to determine all items that have no less than *min_util*. The transaction-weighted utilization (TWU) of an itemset X is the sum of the transaction utilities of all transactions that contain X and is defined as:

$$TWU\left(X\right) = \sum_{X \subset T_d \cap T_d \in D} TU\left(T_d\right) \tag{5}$$

X is high transaction-weighted utilization itemset (HTWUI) if $TWU\left(X\right) \geq$ min_*util*; otherwise, X is a low transaction-weighted utilization itemset. A HTWUI with k items is called k-HTWUI.

Consider the transaction database in Table 4 and the unit profit table in Table 5. In the example database, the utility of item B in transaction T_1 is $u\left(B, T_1\right) = 2 \times 2 = 4$. The utility of itemset BC in transaction T_1 is $u\left(BC, T_1\right) = u\left(B, T_1\right) + u\left(C, T_1\right) = 4 + 2 = 6$. The utility of the same itemset in the entire database is $u\left(BC\right) = u\left(BC, T_1\right) + u\left(BC, T_2\right) + u\left(BC, T_5\right) = 6 + 11 + 5 = 22$. Assuming the min_util = 50, the itemset BC is not HUI because $u\left(BC\right) <$ min_*util*.

The TU of T_4 is $TU\left(T_4\right) = u\left(ABCDE, T_4\right) = 25$. Likewise, the utilities of the other transactions are shown

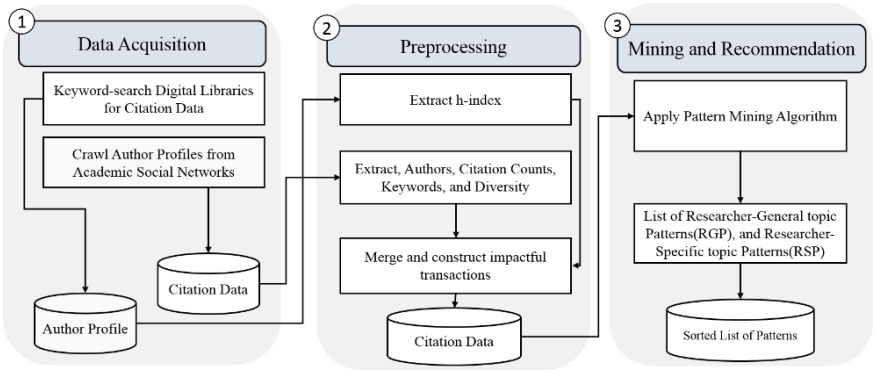


FIGURE 2. The architecture of the utility-pattern based academic reviewer recommendation.

TABLE 4. Example database.

TID	Transactions	TU
T_1	(B, 2), (C,2), (E,1)	9
T_2	(B,4), (C,3), (D,3), (E,1)	20
T_3	(A,1), (C,1), (D,1)	8
T_4	(A,2), (C,6), (E,2)	22
T_5	(A,1), (B,2), (C,1), (D,6), (E,1)	25

TABLE 5. Profit table.

Item	A	B	C	D	E
Profit	5	2	1	2	3

in the last column of Table 4. The TWU of the itemset BC is $TWU\ (BC) = TU\ (T_1) + TU\ (T_2) + TU\ (T_5) = 9 + 20 + 25 = 54$. Thus, BC is HTWUI.

The introduction of HUIM problem has motivated research in two directions [52]: 1) The development of efficient algorithms, and 2) The application of this technique to various domains.

1) HIGH UTILITY ITEMSET MINING ALGORITHMS

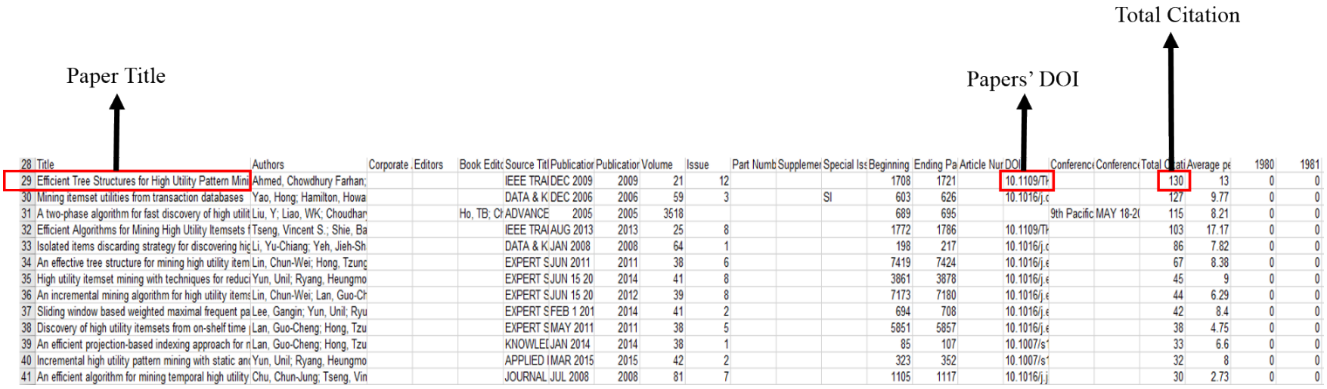
HUIM algorithms are mainly developed focussing on their efficiency in terms of execution time and memory space. The first such algorithms is Two-Phase [4]. It introduced the Transaction-Weighted Downward Closure (TWDC) property which is analogous to the apriority property [1], and used it to prune the search space. The algorithm follows two phases. In phase one, itemsets with TWU that satisfies the utility threshold are found. In phase two, the database is scanned to determine the real HUIs. Other algorithms that use tree-based data structures were also developed. Examples of such algorithms include Utility Pattern Growth (UP-Growth) [6], Incremental High Utility Pattern mining (IHUP) [5] among others [52]. Recently, more efficient algorithms such as Efficient high-utility Itemset Mining (EFIM) [11] that use different data structures have been developed [52]. The proposed work in this paper applies some of these algorithms to the problem of reviewer recommendation.

2) HIGH UTILITY ITEMSET MINING APPLICATIONS

Although most of the HUIM algorithms can be adopted for real-world applications in different domains, for effectiveness, researchers have devised application-specific algorithms as well. Such algorithms include IHUP [5] which was developed to handle incremental datasets, and High Utility Itemsets with Negative Item Values (HUINV-Mine) [9]. Although other algorithms have also been developed, the effectiveness of HUIM is still a challenging task [52]. The work in this paper contributes to the application aspect of HUIM. In essence, it uses the domain knowledge of academic social network and tests the applicability of HUIM techniques in the reviewer recommendation problem.

B. REVIEWER RECOMMENDATION

The emergence of big data generated due to the scientific writings has inspired researchers to utilize such data for facilitating the tasks involved in the research process. This massive amount of data in academia is referred to as - Big Scholarly Data (BSD) [13]. Within this context, recommending reviewers and collaborators is the central problem faced by both journal editors and research institutions. Although the literature reports attempts to tackle the reviewer recommendation and collaboration, few studies [2], [25] have reported applying pattern-mining approaches to the problem. The reviewer recommendation problem [14], [31], [38] also known as referee assignment [32], or Reviewer Assignment (RA) [31], can be viewed as a sub field of the expert search and finding problem [26]. Which in turn is defined as sorting a list of candidates with a proven knowledge in a certain domain for a given query [27], [28]. For its importance, the expert finding problem has a dedicated track in Text Retrieval Conference (TREC) known as Expert Finding Track [29]. According to the literature on BSD analysis [12], the previous methodologies of reviewer recommendation can be grouped into two main methods [30]. The first category builds a profile of the experts' knowledge [14], [15], and known as profile based,



	Title	Authors	Corporate	Editors	Book Edit	Source Titl	Publicatior	Publicatior	Volume	Issue	Part Numb	Supplemer	Special Iss	Beginning	Ending Pa	Article Nur	DOI	Conferenc	Conferenc	Total Citati	Average pe	1980	1981
29	Efficient Tree Structures for High Utility Pattern Mini	Ahmed, Chowdhury Farhan;				IEEE TRAI	DEC 2009	2009	21	12				1708	1721		10.1109/TK			130	13	0	0
30	Mining itemset utilities from transaction databases	Yao, Hong; Hamilton, Howa				DATA & K	DEC 2006	2006	59	3			SI	603	626		10.1016/j.c			127	9.77	0	0
31	A two-phase algorithm for fast discovery of high utilit	Liu, Y; Liao, WK; Choudhar			Ho, TB; Cl	ADVANCE	2005	2005	3518					689	695			9th Pacific	MAY 18-2(	115	8.21	0	0
32	Efficient Algorithms for Mining High Utility Itemsets f	Tseng, Vincent S.; Shie, Ba				IEEE TRAI	AUG 2013	2013	25	8				1772	1786		10.1109/TK			103	17.17	0	0
33	Isolated items discarding strategy for discovering hig	Li, Yu-Chiang; Yeh, Jieh-Sh				DATA & K	JAN 2008	2008	64	1				198	217		10.1016/j.c			86	7.82	0	0
34	An effective tree structure for mining high utility item	Lin, Chun-Wei; Hong, Tzung				EXPERT S	JUN 2011	2011	38	6				7419	7424		10.1016/j.e			67	8.38	0	0
35	High utility itemset mining with techniques for reduci	Yun, Unil; Ryang, Heungmo				EXPERT S	JUN 15 20	2014	41	8				3861	3878		10.1016/j.e			45	9	0	0
36	An incremental mining algorithm for high utility items	Lin, Chun-Wei; Lan, Guo-Ch				EXPERT S	JUN 15 20	2012	39	8				7173	7180		10.1016/j.e			44	6.29	0	0
37	Sliding window based weighted maximal frequent pa	Lee, Gangin; Yun, Unil; Ryu				EXPERT S	FEB 1 201	2014	41	2				694	708		10.1016/j.e			42	8.4	0	0
38	Discovery of high utility itemsets from on-shelf time	Lan, Guo-Cheng; Hong, Tzu				EXPERT S	MAY 2011	2011	38	5				5851	5857		10.1016/j.e			38	4.75	0	0
39	An efficient projection-based indexing approach for n	Lan, Guo-Cheng; Hong, Tzu				KNOWLEI	JAN 2014	2014	38	1				85	107		10.1007/s1			33	6.6	0	0
40	Incremental high utility pattern mining with static an	Yun, Unil; Ryang, Heungmo				APPLIED I	MAR 2015	2015	42	2				323	352		10.1007/s1			32	8	0	0
41	An efficient algorithm for mining temporal high utility	Chu, Chun-Jung; Tseng, Vin				JOURNAL	JUL 2008	2008	81	7				1105	1117		10.1016/j.j			30	2.73	0	0

FIGURE 3. Report Dataset.

and the second class follows document clustering and then querying [16], and thus known as document centric. Very recently, however, a new category that follow pattern based approach has been introduced [2], [25]. These approaches are briefly introduced in the following sub sub-sections.

1) PROFILE-CENTERIC APPROACHES

In this category, researchers build profiles of potential reviewers based on their expertise [33], [34] and then match keywords related to the paper on hand, to a list of top matching reviewers. For example, Kou *et al.* [35] considered the RA problem as Weighted-coverage Group-based Reviewer Assignment Problem (WGRAP) where reviewer expertise are weighted and compared against the terms of the paper to be reviewed. Afterwards they optimize the number of papers per reviewer [37]. However, methods that follow this approach [36] do not consider information about the authors in terms of their importance. Furthermore, in this paper we follow a pattern-based approach that does not require optimization.

2) DOCUMENT CENTRIC APPROACHES

The literature that follow this approach [28], [39] cluster the papers based on their topics and then rank the authors based on a relevance score. An example of this category is the recent work in [40] where the authors considered additional information like time, which can affect research interest. More recently, Gui *et al.* [30] followed this approach with the addition of authority ranking. Although, document centric models are better than the profile models because they consider the text of whole publications, these models also do not consider extra information related the author. In contrast to the work proposed in this paper, the proposed methods is pattern-based which incorporate the importance of both papers and authors in addition to the keywords.

3) PATTERN BASED APPROACHES

In the pattern based reviewer recommendation approach, bibliographic datasets [56], [57] are collected and the RA is treated as a pattern-mining problem [1]. Associations between Reviewers and or Keywords are found. Although the work in [2], and its extension in [25] are considered pilots in this direction, they treat the problem as a weighted association rule mining [17]. Where the impacts of the publications are considered as weights. The work proposed in this paper, however, although it follows pattern-based approach, it differs from [2], and [25] in several aspects. 1) this paper considers the impacts of authors and keywords in addition to the papers' citation count, 2) this paper finds all patterns of high utility compared to the closed patterns found in [25], and 3) the patterns found in this paper are represented in terms of sets rather than explicit rules as in [25].

III. A UTILITY PATTERN BASED ACADEMIC REVEWER RECOMENDATION

This paper acknowledges the heterogeneous nature of the scholarly network which is comprised of various entities [3]. For instance Figure 1-(b) shows a typical representation of the academic network. The figure shows how researchers affiliated to different institutions can collaborate to generate a written document in a form of paper that summaries a major research task they have accomplished. A closer look at the figure motivates the conceptual representation of its major entities shown in Figure 1-(a). These entities are: The researcher who is affiliated to an institution and can co-author a paper. The second major entity is the paper that have been co-authored by several authors. The two major entities in this model can be evaluated in various ways. For example, authors or researchers in general are evaluated by their h-index [12], [41]. On the other hand, the quality of papers can be measured through the number of citations a particular paper receives. This paper combines the metrics of authors and papers along with the keywords mentioned in each paper and formulate the reviewer recommendation as a utility pattern mining problem. The paper generalizes the proposed analytical approach as a framework. The framework is comprised of three major steps as shown in Figure 2. The steps are Data Acquisition, Pre-processing, and Mining and Recommendations.

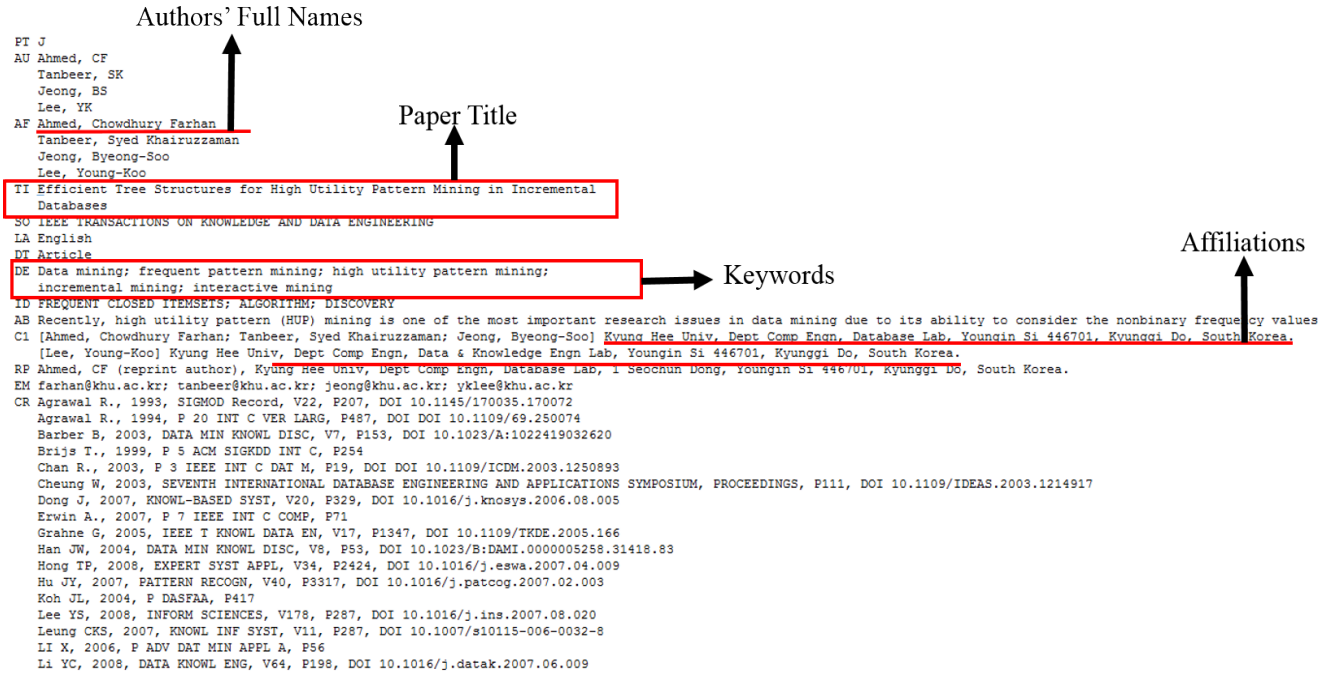


FIGURE 4. Citation records Dataset.

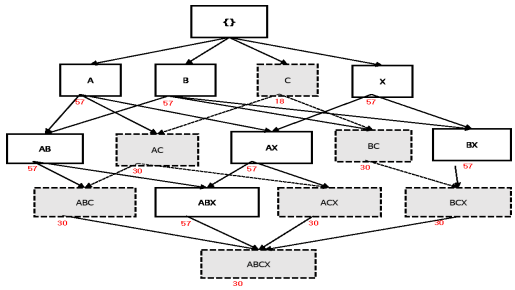


FIGURE 5. Itemset lattice of the impactful transactional databse of Table1.

A. DATA ACQUISITION

The datasets that can help in reviewer recommendation can be collected from two major sources. 1) Digital Libraries such as: Clarivate Analytics' Web of Science Core Collection [56], and Scopus [57]. 2) User profiles found in Online Academic Social Networking Sites, which include Google Scholar [54], ResearchGate [53], and Arnetminer [55].

While researchers can obtain citation data from digital libraries through their institutional subscriptions, the user profile data can only be obtained through web crawlers [20]. Figure 4 shows a typical example of a record in a citation dataset retrieved from Clarivate Analytics' Web of Science Collection. Part (a) of the figure shows the citation record from where major information about the paper are extracted. Part (b), however, shows the associated citation report data from where the paper's citation count and DOI are extracted.

B. DATA PREPROCESSING

Because the work in the paper is motivated by the semantic representation of the citation data, a pre-processing step has to be done in order to obtain a clean copy of the raw data that can be used in the subsequent utility-based pattern-mining step. In particular, this step includes extracting author level quality measures such as the h-index of an author from his academic social network profile, which was obtained in the previous data acquisition step. This step also includes extracting certain information from the citation datasets obtained from digital libraries. This information include the names of the authors, the diversity of their affiliation in terms of geographic locations of their institutions. This is represented by countries. Another piece of information is the citation count of a particular paper along with its Digital Object Identifier (DOI) [22] and a list of keywords used in this paper. Following the extraction of author and paper level information, an integration step is performed in order to generate a transaction-like dataset, which can be used for the pattern mining. Table 1 represents the result of the integrated dataset. Particularly, the first four columns represent the paper ID, list of authors, keywords, diversity, and citation count. On the other hand, Table 2 represents the *h*-index of each author and Table 3 shows the frequencies of the keywords. Algorithm 1 illustrates the data processing steps. The algorithm takes both the citation record file shown in Figure 3, and the report file shown in Figure 4. It refers to the files as $\mathcal{D}_{sr}$ and $\mathcal{D}_{rep}$ respectively. The output is an impactful transactional dataset referred to in the algorithm as $\mathcal{D}$. The algorithm then starts by initializing two lists to keep information about Authors and Keywords. The first list, $Author_{list}$ extracts information about authors. Namely, the author's name from $\mathcal{D}_{sr}$, and his h_{index} from GoogleScholar. It also encodes the authors as numbers represented as the index of the list. The second list, $Keywords_{list}$, encodes the keywords extracted from $\mathcal{D}_{sr}$

Algorithm 1 Generating Impactful Transaction Dataset

Input: $\mathcal{D}_{sr}$; //Citation records Dataset
$\mathcal{D}_{rep}$; //Report Dataset about the Citation Records
$\mathcal{D}$; *An impactful Transaction Dataset*

Output:

01 $Author_{list} \leftarrow \langle (i_1, A_{i_1}, h_{index}(A_{i_1})), (i_2, A_{i_2}, h_{index}(A_{i_2})), \ldots (i_r, A_{i_r}, h_{index}(A_{i_r}))\rangle \leftarrow \emptyset$;
$i = 1, 2, 3 \ldots r$, is the index of the author and represents his code
02 $Keywords_{list} \leftarrow \langle (j_1, K_{j_1}, c(K_{j_1})), (j_2, K_{j_2}, c(K_{j_2})), \ldots (j_r, K_{j_r}, c(K_{j_r}))\rangle \leftarrow \emptyset$;
Where, $c(K_{j_1})$, represents the counts of the keywords.
03 $i \leftarrow 1$;
04 $j \leftarrow 1$;
05 **While** $(\mathcal{D}_{sr} \neq \emptyset)$ **do**
06 **For** each record in $\mathcal{D}_{sr}$ **do**
07 **For** each Author in the record *[p_i] in* $\mathcal{D}_{sr}$ **do**
08 **If** the author has GoogleScholar profile **then**
09 Crawl his profile and extract his h_{index};
10 **else**
11 Set his $h_{index}$to a default value;
12 **If** the Author is not in the $Author_{list}$ **then**
13 Create an entry in $Author_{list}$ with index i and add him, along
14 with his h_{index};
15 $i \leftarrow i + 1$;
16 **For** each Keyword in the record **do**
17 **If** the Keyword is not in the $Keywords_{list}$ **then**
18 Create an entry in $Keywords_{list}$ with index j and add the Keyword;
19 $j \leftarrow i + 1$;
20 **else**
Increment its count;
21 **While** $((\mathcal{D}_{sr} \neq \emptyset) \wedge (\mathcal{D}_{rep} \neq \emptyset))$ **do**
22 **For** each record*[p_i]* in $\mathcal{D}_{sr}$ **do**
23 Extract its Document Object ID(DOI) [it_j]and Citation Count[$C(p_i, it_j)$] from $\mathcal{D}_{rep}$
24 Compute p_i's diversity
25 $D(p_i, it_j) = \sum_{k=1}^{l} (Aff_k, it_j)$ (1);
26 Computep_i'sprestige
27 $\Pr(p_i, it_j) \leftarrow D(p_i, it_j) \times C(p_i, it_j)$ (2);
28 Using $Author_{list}$ and $Keywords_{list}$
29 Compute p_i's Author impacts
30 $AI(A_i, it_j) \leftarrow h_{index}(A_i) + \Pr(p_k, it_j)$ (3);
31 Compute p_i's Keyword impacts
32 $KI(K_i, it_j) \leftarrow frequency(K_i) + \Pr(p_k, it_j)$ (4);
33 Compute it_j's Impact
34 $I(it_j) \leftarrow \sum_{i=1}^{m} AI(A_i, it_j) + \sum_{l=1}^{n} KI(K_l, it_j) + \Pr(p_q, it_j)$
35 (5);
$\mathcal{D} \leftarrow \mathcal{D} \cup it_j$;
36 **Return** $\mathcal{D}$

and stores their counts. The two lists are defined in lines 01, and 02 of the algorithm. The process of filling these lists requires a complete scan of the $\mathcal{D}_{sr}$ dataset. This process is represented in the algorithm by lines 03, through 20. The algorithm then scans both the $\mathcal{D}_{sr}$, and $\mathcal{D}_{rep}$ datasets in order to generate the impactful transactional dataset. This process is shown from line 21 to line 35. The algorithm takes each paper p_i, in $\mathcal{D}_{sr}$, extracts its title as shown in Figure 4, and uses it to extract the papers' DOI from $\mathcal{D}_{rep}$. It then extracts p_i's citation count as shown in Figure 3. Thereafter, p_i's Diversity is calculated using formula (1). Then its Prestige is calculated using formula (2), and the impacts of its authors and keywords are calculated by looking up the necessary information from $Author_{list}$ and $Keywords_{list}$ using formulas (3) and (4). Finally the impact of the transaction is calculated using Formula (5) and a transaction is formed and added to $\mathcal{D}$.

This process continues until the end of records in $\mathcal{D}_{sr}$ and $\mathcal{D}_{rep}$ and then the algorithm returns the complete transactional dataset $\mathcal{D}$.

C. MINING AND RECOMENDATION

The pattern-mining problem was proposed by [1] in order to recommend products from transactional data. It has been extended to account for weighted datasets [17]. Recently a version of the weighted algorithm was applied to the problem of reviewer recommendation [2]. Within their framework, the quality of the paper- its citation count was considered as a weight to extract compact patterns. Although the proposal of [2] is promising, it has overlooked the typical heterogeneity of the academic social network. Therefore, this paper proposes to account for the quality of publication, which stems from the semantics inherent in the network representation as shown in

Figure 1. The quality of the paper can be thought of as comprising from the authors who wrote it and the paper itself. Thus in this work we define a utility-like transaction database wherein the utilities of the papers duped impact can be calculated based on the impacts of the authors, and the prestige of the paper. Like [2], an item in the utility-like transaction is represented as attribute-value pair, where the attribute can be either an author, or a keyword.

Definition 1 (Impactful Transactional Dataset): Let A be the set of authors and S be the set of specific topics. Let P be the set of all scientific publications and let $D(p_i)(p_i \in P)$ be the number of different countries hosting the institutions of the authors who published the paper p_i. Let $C(p_i)(p_i \in P)$ be the number of citations received by paper p_i, and let $Pr(p_i)(p_i \in P)$ be the prestige of paper p_i. Let $I(p_i)(p_i \in P)$ be the impact of paper p_i. An item i_k is a pair feature : v_q, where $v_q \in A$ if feature is Author, and $v_q \in S$, if feature is special topic keyword. A transaction t_j is a set of items related to a paper $p_j(p_j \in P)$. An impactful transactional dataset $\mathcal{D}$ is a set of impactful transactions $it_j \in \mathcal{D}$ corresponds to different paper p_i and consists of the triplet $\langle t_j, Pr(p_j), I(p_j)\rangle$.

Example 1: Looking at the scenario presented in figure1, a total of 9 papers are depicted which are authored and co-authored by five authors labelled A, B, C, D, and E. each one of the researchers is affiliated to an institution that resides in a different country. The institutions are labelled C_1 through C_4. Table 1 shows an impactful transactional dataset extracted from the scenario of Figure 1-(b). Each transaction in the table is identified by a PID shown in column 1. For instance paper the first paper in the scenario on figure 1 which is labelled P_1 and couther by authors A and B both of which belong to the same institution C_1 is given a PID $= P_1$, Authors $=$ A, B, and a specific topic $=$ X. the other two papers are populated the same way.

Because the goal of this paper is to recommend reviewers who are qualified, we formulate a utility-pattern mining approach to achieve the task. To this end, we define the concepts of itemset, the Author-General topic pattern, the Author-Specific topic pattern and the mining problem

Definition 2 (Itemset): Let $\mathcal{D}$ be an impactful transactional dataset and let I be the set of distinct items in the form feature : v_q contained in any impactful transaction $it_j \in \mathcal{D}$. A k-itemset is a set of k distinct items in I.

Itemsets can represent combination of Authors and Keywords.

Definition 3 (Paper Diversity (D)): The Diversity of a paper p_i in an impactful transaction it_j denoted as $D\left(p_i, it_j\right)$ is the sum of the affiliations Aff_k of its authors. And is given by the following formula.

$$D\left(p_i, it_j\right) = \sum_{k=1}^{l} \left(Aff_k, it_j\right) \tag{6}$$

where, $k = 1, 2, \ldots l$ is the number of unique Affiliations in p_i.

For example, the diversity of the paper P_3 in table 1 is calculated as $D(P_3, P_3) = 2$, because from the scenario at Figure 1-(b), P_3 was co-authored by three researchers: A, B and C. Both authors A and B are affiliated to the same institution, labelled C_1. Whereas, the third author; C is affiliated to a different institution, C_2.

Definition 4 (Paper Prestige (Pr)): The prestige of a paper p_i in an impactful transaction it_j denoted as $\Pr\left(p_i, it_j\right)$ is the product of its diversity and citation count. And is given by the following formula.

$$\Pr\left(p_i, it_j\right) = D\left(p_i, it_j\right) * C\left(p_i, it_j\right) \tag{7}$$

For example, the prestige of the paper P_2 in table 1 is calculated as $\Pr(P_2, P_2) = D(P_2, P_2) * C(P_2, P_2) = 1 * 15 = 15$.

Definition 5 (Author Impact (AI)): The h_{index} of an author and the prestige of the publication in a particular impactful transaction represent the impact of an author

$$AI\left(A_i, it_j\right) = h_{index}\left(A_i\right) + \Pr\left(p_k, it_j\right) \tag{8}$$

For example, the author impact of C in P_2 in table 1 is calculated as $AI(A, P_2) = \text{h-indec}(A) + \Pr(P_2, P_2) = 3 + 15 = 18$. And it is shown in red in the itemset lattice depicted in figure 3.

Definition 6 (Keyword Impact (KI)): The frequency of a keyword and the prestige of the publication in a particular impactful transaction represent the impact of that keyword.

$$KI\left(K_i, it_j\right) = frequency\left(K_i\right) + \Pr\left(p_k, it_j\right) \tag{9}$$

For example, the keyword impact of X in P_1 in table 1 is calculated as $KI(X, P_1) = \text{frequency}(K) + \Pr(P_1, P_1) = 1 + 20 = 21$. And it is shown in red in the itemset lattice depicted in Figure 5. Where for the sake of simplicity, in this example the Keyword Impact is assumed to comprise of prestige only.

Definition 7 (Transaction Impact (I)): The impact of an impactful transaction is the combination of the prestige of the paper it represents along with the impacts of the authors and the keywords.

$$I\left(it_j\right) = \sum_{i=1}^{m} AI\left(A_i, it_j\right) + \sum_{l=1}^{n} KI\left(K_l, it_j\right) + \Pr(p_q, it_j) \tag{10}$$

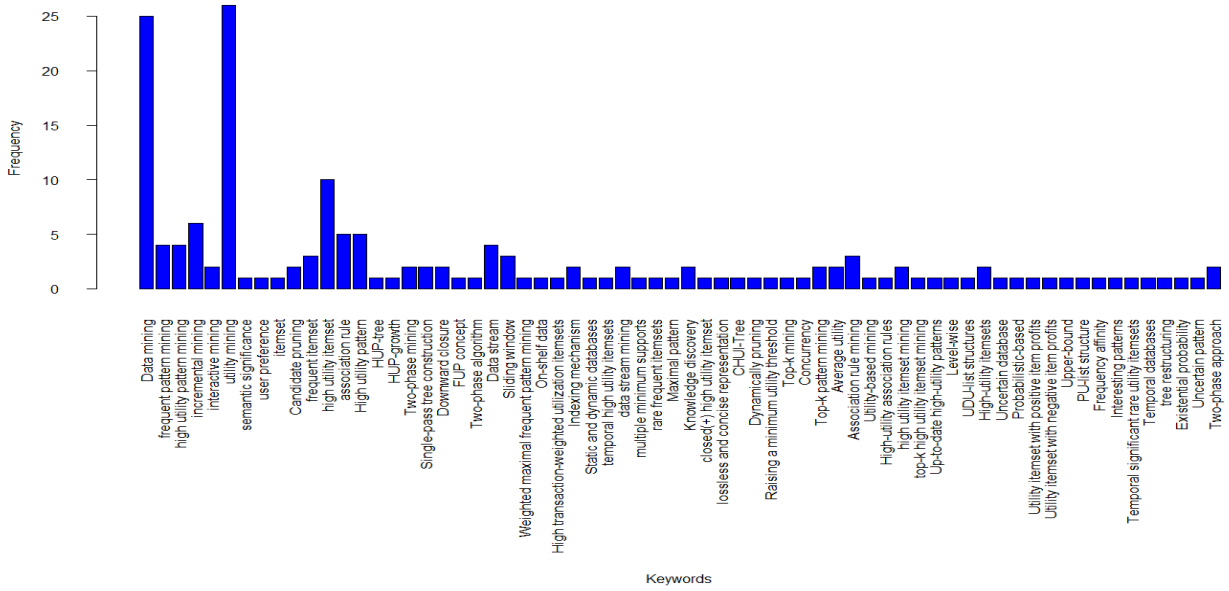


FIGURE 6. The top Keywords used in the citation dataset related to High Utility Itemset along with their frequencies.

where $i = 1, \ldots m$ *represents the number of authors, and* $l = 1, 2, \ldots n$ represents the number of specific topic keywords.

For example, the impact of the impactful transaction P_1 in table 1 is calculated as $\mathrm{I}(P_1) = \mathrm{AI}(A, P_1) + \mathrm{AI}(B, P_1) + \mathrm{KI}(\mathrm{X}, P_1) + \mathrm{Pr}(P_1, P_1) = 2 + 5 + 20 = 27$. And it is shown in the last column of table 1.

At this point, setting a minimum impact threshold, referred to hereafter as *minImpact*, and applying utility pattern mining, one would be able to extract patterns of various item combinations. However, in this paper, for the sake of academic reviewer recommendation, two special patterns are of interest. Namely, Author-General topic Patterns, and Author-Specific topic Patterns.

Definition 8 (Author General Topic Pattern (AGP)): Is a pattern that consists of combinations of authors only. The term General refers to the general keywords used to obtain the citation data at the first step of the proposed framework.

For example, in the itemset lattice of Figure 5, when the *minImpact* $= 20$, the none shaded items represent patterns of interest. Particularly patterns where only authors are shown represent AGP.

Definition 9 (Researcher-Specific topic Pattern (RSP)): Is a pattern that consists of combinations of authors and a specific topic. The term Specific refers to the specific keywords that are associated with the citation dataset obtained from an online database.

For example, in the itemset lattice of figure 3, when the *minImpact* $= 20$, the non-shaded items represent patterns of interest. Particularly patterns where authors and keywords are shown represent RSP patterns.

IV. CASE STUDY

The proposed framework have been applied to a real-world citation dataset. The datasets were prepared according to the framework and then a mining algorithm was selected to perform the task. The discovered RGP and RSP patterns were evaluated in terms of coverage.

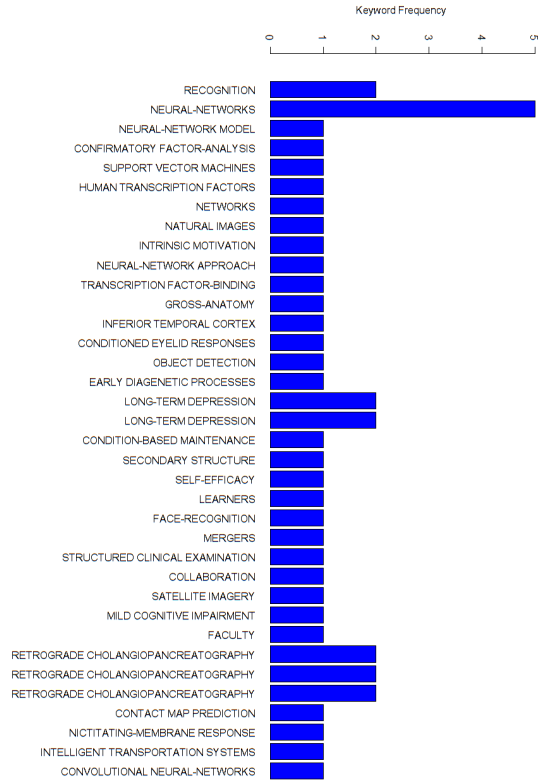


FIGURE 7. The Keywords used in Deep Learning citation data along with their frequencies.

A. EXPERIMENTAL DATASET

For the purposes of the experiments, two real-world datasets and one standard dataset were used. The two real-world

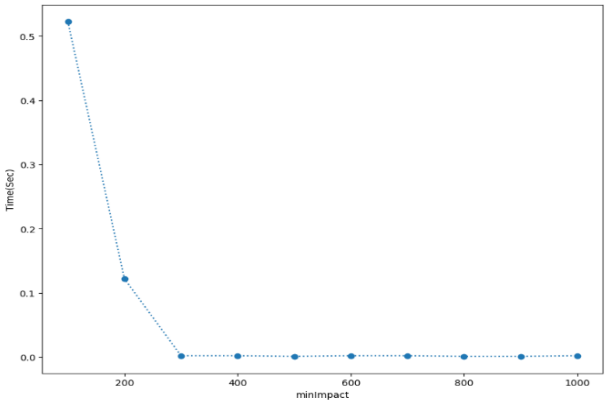


FIGURE 8. Run Time vs minImpact threshold using citation dataset related to Deep Learning.

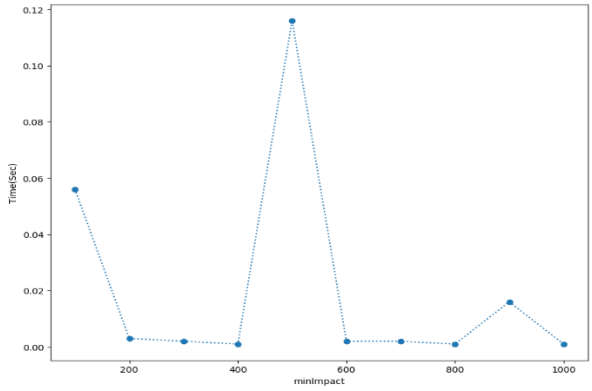


FIGURE 9. Run Time vs minImpact threshold using citation dataset related to HUIM.

datasets were retrieved from Clarivate Analytics' Web-of-Science Core Collection (https://clarivate.com). The first one is a citation dataset related to the publications on Deep Learning [42], and the other one is related to High Utility Itemset mining [6]. The search queries used to retrieve the two datasets were "Deep learning" and "High Utility Itemset Mining" in that order. Although the dataset retrieved were big in size, for the purposes of proving the applicability of the proposal, only the top cited articles were selected for this study. Statistics about these two datasets are shown in Table 6. In parallel to that, researchers' profiles were retrieved from Google Scholar. The necessary information about the authors' names, diversity, and citation counts were extracted from the citation records [19]. The h-index of authors' however, was obtained from their Google Scholar profiles. For the keywords, the most frequent keywords were identified and two were chosen for the experiment in each dataset. Figures 6, and 7 show a histogram of the most frequent keywords in the two datasets. From the Deep learning citation dataset, only "Recognition and "Neural Networks" were chosen for the analysis. Whereas, from the High Utility Itemset mining dataset, "Data Mining" and "Utility Mining" were chosen. Furthermore, for the purposes of selecting a suitable mining algorithm, the Mushroom

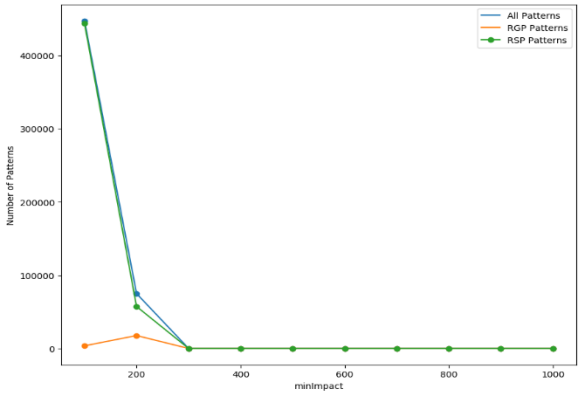


FIGURE 10. Number of patterns vs minImpact threshold using citation dataset related to Deep learning.

TABLE 6. Information about the citation datasets considered.

Property	Citation Dataset	
	Deep learning	HUI
Number of Authors	170	63
Number of papers	44	44
Number of papers with diverse authors	14	27
Keywords considered	2	2

dataset (http://www.philippe-fournier-viger.com/spmf/) was also used to compare eight mining algorithms. The distribution of the three datasets is given in table 5.

B. CHOOSING A MINING ALGORITHM

Although a number of HUI mining algorithm are being developed [43], [44], few tools are available where implementations of these algorithms are freely available [24]. Therefore, in order to choose an appropriate algorithm, the UP-Miner toolkit [24] was used where the major item utility algorithms were compared for suitability of our proof of concept experiment. The algorithms considered for this purpose were, Two Phase [4], Incremental High Utility Pattern (IHUP) mining [5], Utility Pattern Growth (UP-Growth) [6],

Fast high Utility Miner (FHM) [7], Faster High-Utility itemset miner with Negative unit profits (FHN) [8], (High Utility Itemsets with Negative Item Values (HUINIV)-Mine [9], Direct Discovery of High Utility Patterns (D2HUP) [10], and Efficient high-utility Itemset Mining (EFIM) algorithm [11]. These algorithms were compared across three dimensions; number of patterns obtained, and the space and time resource consumed. Table 8 shows the results of the comparison. It is clear from the table that only three out of the eight algorithms considered reported the same number of patterns. Those were namely, FHM, FHN, and EFIM. Although FHN might produce different number of patterns generally, in this case it produced exactly the same number of patterns as FHM and EFIM. This is because our dataset

TABLE 7. Information about the citation datasets considered.

Dataset	Transaction Count	Distinct item count	Average Transaction Length	Maximum length of Transaction
Deep learning	44	170	4	20
High Utility Itemset	44	63	4	7
Mushroom	88,162	16,470	23	23

TABLE 8. Comparison of different high utility itemset mining algorithms.

Dataset	Algorithm	Maximum Memory in Megabytes (MB)	Execution Time in Seconds (Sec)	Number of Patterns Returned
Deep Learning	Two Phase	560.54	446.93	203648
	IHUP	510.57	2.70	706409
	UP-Growth	830.78	108.28	447320
	FHM	**68.81**	0.63	**447328**
	FHN	85.31	0.65	**447328**
	HUINIV-Mine	411.01	422.53	447328
	D2HUP	138.09	1.95	697471
	EFIM	73.56	**0.60**	**447328**
High Utility Itemset	Two Phase	77.16	0.06	254
	IHUP	78.02	0.07	254
	UP-Growth	**15.49**	0.30	254
	FHM	38.58	0.04	254
	FHN	38.59	0.04	254
	HUINIV-Mine	38.15	0.06	254
	D2HUP	39.01	**0.03**	254
	EFIM	38.58	**0.03**	254
Mushroom	Two Phase	**NA**	**NA**	**N/A**
	IHUP	2018.72	7209.09	1045780
	UP-Growth	1376.84	4902.57	1045780
	FHM	1704.66	123.05	1045780
	FHN	1918.02	512.09	1045780
	HUINIV-Mine	**N/A**	**N/A**	**N/A**
	D2HUP	**528.28**	26.13	1045780
	EFIM	**567.47**	**1.53**	1045780

- For the Deep learning, and High Utility Itemset data, *minImpact* was set to *100*.
- For the Mushroom dataset, the *minUtil* was set to *100000*.

TABLE 9. Impactful RGP patterns and RSP patterns generated from the Deep learning Citation Datasets.

Patterns Discovered				
Type	ID	Pattern	impact	Coverage
RGP	**1**	{(**R:** "Hinton, Geoffrey E"), (**R:** "Bengio, Yoshua")}	486	100%
	2	{(**R:** "Bengio, Yoshua")}	424	100%
	3	{(**R:** "Hinton, Geoffrey E")}	411	100%
	4	{(**R:**" LeCun, Yann "), (**R:** "Hinton, Geoffrey E"), (**R:** "Bengio, Yoshua")}	341	100%
	5	{(**R:**"Osindero, Simon"), (**R:** "Teh, Yee-Whye"), (**R:** "Hinton, Geoffrey E"), (**R:** "Bengio, Yoshua")}	**306**	**70%**
RSP	**6**	{(**R:** "LeCun, Yann"), (**R:** "Hinton, Geoffrey E"), (**R:**"Bengio, Yoshua"), (**S:**"Neural networks") }	346	100%
	7	{(**R:**"Bengio, Yoshua"), (**S:**"Neural Networks")}	333	100%
	8	{(**R:**"Osindero, Simon"), (**R:** "Teh, Yee-Whye"), (**R:** "Hinton, Geoffrey E"), (**R:** "Bengio, Yoshua"), (**S:** "Neural")}	**308**	**60%**
minImpact threshold used to generate these patterns was **300**				

does not include negative utility values. Two algorithms, Two Phase, and HUINIV-Mine reported memory errors and we could net obtain their results when the mushroom dataset was

C. PATTERNS DISCOVERED

EFIM algorithm was used to mine the resulting citation datasets for patterns. Firstly, in order to analyse the patterns,

TABLE 10. Impactful RGP patterns and RSP patterns generated from the HUI Citation Datasets.

Patterns Discovered				
Type	ID	Pattern	impact	Coverage
RGP	1	{(**R:** "Yu, Philip S"), (**R:** "Tseng, Vincent S")}	756	100%
RSP	2	{(**R:** "Yu, Philip S"), (**R:** "Tseng, Vincent S"), (**S:** "utility mining")}	860	100%
	3	{ (**R:** "Tseng, Vincent S"), (**S:** "utility mining")}	828	100%
	4	{(**R:** "Tseng, Vincent S"), (**R:** "Wu, Cheng-Wei"), (**R:** "Yu, Philip S"), (**S:** "utility mining")}	702	100%
minImpact threshold used to generate these patterns was **700**				

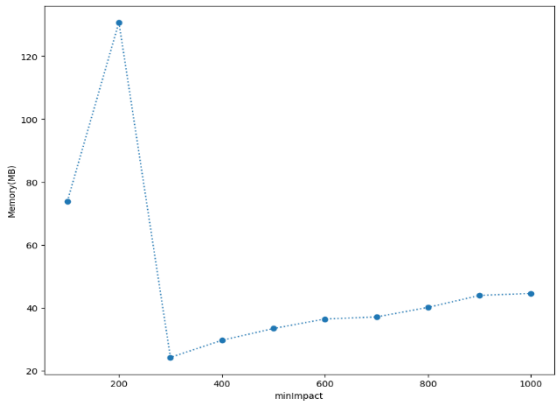


FIGURE 11. Memory vs minImpact threshold using citation dataset related to Deep Learning.

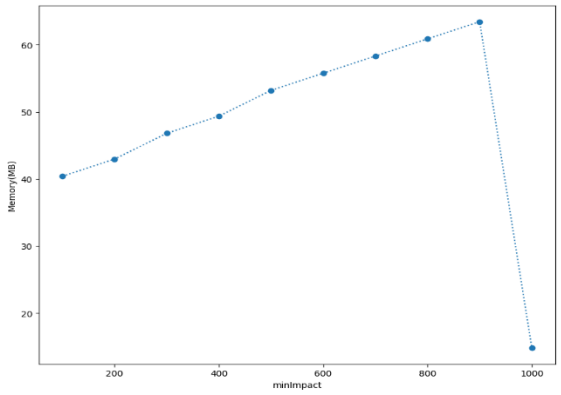


FIGURE 12. Memory vs minImpact threshold using citation dataset related to HUIM.

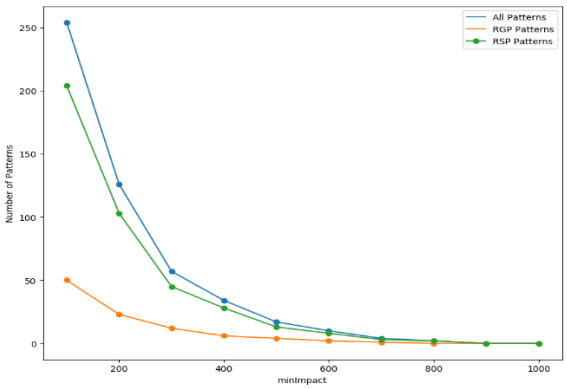


FIGURE 13. Number of patterns vs minImpact threshold using citation dataset related to HUIM.

the number of patterns generated while the minimum impact threshold varied were observed and represented in Figure 10, and Figure 13 for Deep Learning and High Utility Itemset Mining datasets respectively. From the figures, it was confirmed that patterns decrease when the impact threshold increase. In parallel, Figure 11 and Figure 12 depict the memory space consumed when performing the pattern search at different values of the *minImpact*. The memory decreased steadily. Likewise, Figure 8 and Figure 9 observe the run time of the algorithm across different values of *minImpact*. The runtime also decreased steadily. Thereafter, the patterns discovered from the Deep Learning citation dataset were observed and presented in table 7. Exactly 8 patterns were discovered when the minImpact threshold was set to 300. In the table, patterns are sorted according to their *minImpact*threshold. Those with higher values of minImpact top the list. The patterns are also classified into RGP and RSP. Five RGP patterns were found and only three RSP patterns were reported. At the top of the Researcher General topic pattern was, {(**R:** "Hinton, Geoffrey E"), (**R:** "Bengio, Yoshua")} with minImpact of 486. This pattern can be interpreted as: both researchers, Mr. "Hinton, Geoffrey E", and "Bengio, Yoshua" are top experts in the general topic of "Deep learning" and thus, they can be contacted to review articles related to the topic. On the other hand, Table 10 reports the patterns found when the dataset was about High Utility Itemset mining and the minImpact threshold was set to 700. The RGP pattern, {(**R:** "Yu, Philip S"), (**R:** "Tseng, Vincent S")} with an impact of 756 reports two experts in the field. Namely, "Yu, Philip S" and "Tseng, Vincent S". Not surprisingly, however, the both experts were reported at the top of the RSP patterns as, {(**R:** "Yu, Philip S"), (**R:** "Tseng, Vincent S"), (**S:** "utility mining")} indicating the subject matter expertise in the specific topic of "Utility Mining".

D. PATTERN VERIFICATION

In order to verify the patterns reported by the algorithm, Google Scholar's search results were assumed to represent

TABLE 11. Top 11 results retrieved from GoogleScholar when searching about RGP pattern 1 related to HUI.

Pattern	CC	Authors	Paper Title	DOI
{(**R:** "Yu, Philip S"), (**R:** "Tseng, Vincent S"), (**S:** "utility mining")}: (***impact*** :860)	311	**Vincent S. Tseng** Bai-En Shie Cheng-Wei Wu **Philip S. Yu**	Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases	10.1109/TKDE.2012.59
	270	**Vincent S. Tseng** Cheng-Wei Wu Bai-En Shie **Philip S. Yu**	UP-Growth: an efficient algorithm for high tility itemset mining	10.1145/1835804.1835839
	108	Cheng Wei Wu Bai-En Shie **Vincent S. Tseng** **Philip S. Yu**	Mining top-K high utility itemsets	10.1145/2339530.2339546
	69	Bai-En Shie **Vincent S. Tseng** **Philip S. Yu**	Online mining of temporal maximal utility itemsets from data streams	10.1145/1774088.1774436
	65	**Vincent S. Tseng** Cheng-Wei Wu Philippe Fournier-Viger **Philip S. Yu**	Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets	10.1109/TKDE.2014.2345377
	64	Bai-En Shie, Hui-FangHsiao **Vincent S.Tseng** **Philip S. Yu**	Mining High Utility Mobile Sequential Patterns in Mobile Commerce Environments	10.1007/978-3-642-20149-3_18
	61	Cheng-Wei Wu Yu-Feng Lin **Philip S. Yu** Vincent S. Tseng	Mining high utility episodes in complex event sequences	10.1145/2487575.2487654
	56	Bai-En Shiea **Philip .Yu** **Vincent S.Tseng**	Efficient algorithms for mining maximal high utility itemsets from data streams with different models	10.1016/j.eswa.2012.05.035
	70	**Vincent S. Tseng** Cheng-Wei Wu Philippe Fournier-Viger **Philip S. Yu**	Efficient Algorithms for Mining Top-K High Utility Itemsets	10.1109/TKDE.2015.2458860
	35	Cheng Wei Wu, Philippe Fournier-Viger **Philip S. Yu** **Vincent S. Tseng**	Efficient Mining of a Concise and Lossless Representation of High Utility Itemsets	10.1109/ICDM.2011.60

CC : Citation Count from GoogleScholar,
DOI: Document Object Identifier

the ground truth. Thereafter, the reported patterns by the algorithm on each datasets were used to search Google Scholar. The results of the search are then checked to confirm whether they matched the underlying search term (i.e. the pattern). However, in order to quantify this process, a measure of coverage is used. Although the coverage as a measure [45] is used in the literature to evaluate Association Rules [1], this paper uses the measure in a slightly modified way. In essence, we define Pattern Coverage, r, as : Pattern Coverage (r) is the fraction of ground truth records (GTR) that satisfy an RGP, or RSP pattern.

$$r = {}^{GTR}/_{N} \tag{11}$$

where, N is the total number of the GTR.

During the experiment for coverage, only the top ten results reported by Google Scholar in its first page are used to calculate the pattern coverage. Therefore, in our experiments, N is assumed to be 10. In addition, a ground truth pattern is said to be covered by the underlying search pattern, if and only if at least one of the authors in the pattern is among the coauthors of the ground truth.

The values of coverage of each pattern are reported in the last columns of Tables 11 and 12. The patterns have also been sorted according to their coverage values. From table 8. The top RSP pattern according to its impact was, {(**R:** "Yu, Philip S"), (**R:** "Tseng, Vincent S"), (**S:** "utility mining")}: (**impact** :860). When this pattern was used as a search query in Google Scholar to calculate its coverage, the top ten results reported at the first page are reported in table 9. A closer look at the authors- the third column of the table, shows that the pattern covers all elements of the ground truth. Consequently, a coverage of 100% was calculated for this pattern. On the contrary, a closer look at the search results of the third RGP pattern, {(**R:**"Osindero, Simon"), (**R:** "Teh, Yee-Whye"), (**R:** "Hinton, Geoffrey E"), (**R:** "Bengio, Yoshua")}: (**impact** :306) from the Deep Learning Citation datasets, will explain why its coverage was only 70%. The results reported by Google Scholar when this pattern was used as a keyword are shown in table 10. Looking at the fourth column of the table entitled as: "Paper Title", three papers reported in among the top ten results did not include any of the authors listed in the underlying search pattern among their authors.

TABLE 12. Top 10 results retrieved from GoogleScholar when searching about RSP pattern 2 related to Deep Learning data.

Pattern	CC	Authors	Paper Title	DOI
{(**R:**"Osindero, Simon"), (**R:** "Teh, Yee-Whye"), (**R:** "Hinton, Geoffrey E"), (**R:** "Bengio, Yoshua"}: (***impact*** :306)	689	Salah Rifai Pascal Vincent Xavier Muller Xavier Glorot **Yoshua Bengio**	Contractive auto-encoders: explicit invariance during feature extraction	-
	214	**Yoshua Bengio** Eric Thibodeau-Laufer Guillaume Alain Jason Yosinski	Deep Generative Stochastic Networks Trainable by Backprop	-
	674	Quoc V. Le Jiquan Ngiam Adam Coates Abhik Lahiri Bobby Prochnow Andrew Y. Ng	On optimization methods for deep learning	-
	1429	Dumitru Erhan **Yoshua Bengio** Aaron Courville Pierre-Antoine Manzagol Pascal Vincent Samy Bengio	Why Does Unsupervised Pre-training Help Deep Learning?	-
	34	Jianwen Xie Yang Lu Song-Chun Zhu Ying Nian Wu	**A Theory of Generative ConvNet**	-
	424	**Yoshua Bengio**	Deep Learning of Representations for Unsupervised and Transfer Learning	-
	32	Jifeng Dai Yang Lu Ying-Nian Wu	**Generative Modeling of Convolutional Neural Networks**	-
	674	Hugo Larochelle **Yoshua Bengio** Jérôme Louradour Pascal Lamblin	Exploring Strategies for Training Deep Neural Networks	-
	880	**Yoshua Bengio** Yann LeCun	Scaling Learning Algorithms towards AI	-
	104	Elman Mansimov Emilio Parisotto Jimmy Lei Ba Ruslan Salakhutdinov	**Generating Images from Captions with Attention**	-

CC : Citation Count from GoogleScholar, **DOI:** Document Object Identifier

The titles of these papers are shown in bold-face. The papers were, ''A.

Theory of Generative ConvNet'', ''Generative Modeling of Convolutional Neural Networks'' and ''Generating Images from Captions with Attention'' respectively.

V. CONCLUSION

In ordere to solve the issue of finding academic reviewers for peer review and potential research collaborators, this paper formulated the reviewer recommendation problem as a utility pattern-mining task and contributed a new framework to mine special kinds of patterns that directly recommend reviewers. In essence, two patterns were proposed and mined. A Researcher General topic Pattern(RGP) and Researcher Specific topic Pattern (RSP). Two real-world citation datasets were prepared according the data processing algorithm cotributed by the framework. Thereafter, eight different algorithms were compared and EFIM was selected to mine the patterns. Experimental results showed the patterns mined were of a high quality according to their coverage.

The application of the framework within a big data platform along with a suitable algorithm are among the future work.

ACKNOWLEDGMENT

The authors would like to thank Mr. Erdenebileg Batbaatar for his contribution in writing a Java program that converts the raw citation records into a format suitable for the UP-Miner tool, and Ms. Khishigsuren Davagdorj for drawing figure 1.

REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[2] L. Cagliero, P. Garza, M. R. Kavoosifar, and E. M. Baralis, "Identifying collaborations among researchers: A pattern-based approach," in *Proc. 2nd Joint Workshop Bibliometric-Enhanced Inf. Retr. Natural Lang. Process. Digit. Libraries (BIRNDL)*, 2017, pp. 56–68.

[3] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 46, p. 16569, Aug. 2005, doi: 10.1073/pnas.0507655102.

[4] Y. Liu, W. K. Liao, and A. Choudhary, "A fast high utility itemsets mining algorithm," in *Proc. 1st Int. Workshop Utility-Based Data Mining*, Aug. 2005, pp. 90–99, doi: 10.1145/1089827.1089839.

[5] C. F. Ahmed, S. K. Tanbeer, B. S. Jeong, and Y. K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 12, pp. 1708–1721, Dec. 2009.

[6] V. S. Tseng, B.-E. Shie, C.-W. Wu, and P. S. Yu, "Efficient algorithms for mining high utility itemsets from transactional databases," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1772–1786, Aug. 2013.

[7] P. Fournier-Viger, C. W. Wu, S. Zida, and V. S. Tseng, "FHM: Faster high-utility itemset mining using estimated utility co-occurrence pruning," in *Proc. Int. Symp. Methodol. Intell. Syst.*, 2014, pp. 83–92, doi: 10.1007/978-3-319-08326-1_9.

[8] P. Fournier-Viger, "FHN: Efficient mining of high-utility itemsets with negative unit profits," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2014, pp. 16–29, doi: 10.1007/978-3-319-14717-8_2.

[9] C. J. Chu, V. S. Tseng, and T. Liang, "An efficient algorithm for mining high utility itemsets with negative item values in large databases," *Appl. Math. Comput.*, vol. 215, no. 2, pp. 767–778, 2009.

[10] J. Liu, K. Wang, and B. C. M. Fung, "Direct discovery of high utility itemsets without candidate generation," in *Proc. 12th Int. Conf. Data Mining*, Dec. 2012, pp. 984–989, doi: 10.1109/ICDM.2012.20.

[11] S. Zida, P. Fournier-Viger, J. C.-W. Lin, C. W. Wu, and V. S. Tseng, "EFIM: A highly efficient algorithm for high-utility itemset mining," in *Proc. Mex. Int. Conf. Artif. Intell.*, Dec. 2015, pp. 530–546, doi: 10.1007/978-3-319-27060-9_44.

[12] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, Mar. 2017.

[13] Y.-R. Lin, H. Tong, J. Tang, and K. S. Candan, "Guest editorial: Big scholar data discovery and collaboration," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 1–2, Mar. 2016.

[14] T. H. Davenport, and L. Prusak, *Working Knowledge: How Organizations Manage What They Know*. Boston, MA, USA: Harvard Business Press, 1998.

[15] X. Liu, W. B. Croft, and M. Koll, "Finding experts in community-based question-answering services," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2005, pp. 315–316, doi: 10.1145/1099554.1099644.

[16] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, "ExpertRank: A topic-aware expert finding algorithm for online knowledge communities," *Decis. Support Syst.*, vol. 54, no. 3, pp. 1442–1451, Feb. 2013, doi: 10.1016/j.dss.2012.12.020.

[17] W. Wang, J. Yang, and P. S. Yu, "Efficient mining of weighted association rules (WAR)," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2000, pp. 270–274, doi: 2000.

[18] E. Yan, Y. Ding, and C. R. Sugimoto, "P-Rank: An indicator measuring prestige in heterogeneous scholarly networks," *J. Assoc. Inf. Sci. Technol.*, vol. 62, no. 3, pp. 467–477, Mar. 2011, doi: 10.1002/asi.21461.

[19] M. I. Ishag, S. Han, and K. H. Ryu, "Mapping the knowledge domain of FITAT for better research collaboration and dissemination," in *Proc. 9th Int. Conf. Frontiers Inf. Technol., Appl. Tools (FITAT)*, 2016, pp. 1–5.

[20] A. Heydon and M. Najork, "Mercator: A scalable, extensible Web crawler," *World Wide Web*, vol. 2, no. 4, pp. 219–229, Dec. 1999, doi: 10.1023/A:1019213109274.

[21] U. Yun, H. Ryang, G. Lee, and H. Fujita, "An efficient algorithm for mining high utility patterns from incremental databases with one database scan," *Knowl.-Based Syst.*, vol. 124, no. 15, pp. 188–206, 2017.

[22] N. Paskin, "Digital object identifier (DOI) system," *Encyclopedia Library Inf. Sci.*, vol. 3, pp. 1586–1592, Dec. 2010.

[23] H. S. Shon, S. H. Han, K. A. Kim, E. J. Cha, and K. H. Ryu, "Proposal reviewer recommendation system based on big data for a national research management institute," *J. Inf. Sci.*, vol. 43, no. 2, pp. 147–158, 2017, doi: 10.1177/0165551516644168.

[24] V. S. Tseng, C.-W. Wu, J.-H. Lin, and P. Fournier-Viger, "UP-Miner: A utility pattern mining toolbox," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1656–1659, doi: 10.1109/ICDMW.2015.115.

[25] L. Cagliero, P. Garza, M. R. Kavoosifar, and E. Baralis, "Discovering cross-topic collaborations among researchers by exploiting weighted association rules," *Scientometrics*, vol. 116, no. 2, pp. 1273–1301, Aug. 2018, doi: 10.1007/s11192-018-2737-3.

[26] S. Lin, W. Hong, D. Wang, and T. Li, "A survey on expert finding techniques," *J. Intell. Inf. Syst.*, vol. 49, no. 2, pp. 255–279, Oct. 2017, doi: 10.1007/s10844-016-0440-5.

[27] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise retrieval," *Found. Trends Inf. Retr.*, vol. 6, nos. 2–3, pp. 127–256, Jul. 2012, doi: 10.1561/1500000024.

[28] H. Deng, I. King, and M. R. Lyu, "Formal models for expert finding on DBLP bibliography data," in *Proc. 9th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 163–172, doi: 10.1109/ICDM.2008.29.

[29] I. Soboroff, A. P. de Vries, and N. Craswell, "Overview of the TREC 2006 enterprise track," in *Proc. 15th Text Retr. Conf. (Trec)*, Gaithersburg, MD, USA, 2006, pp. 1–20.

[30] H. Gui, Q. Zhu, L. Liu, A. Zhang, and J. Han. (2018). "Expert finding in heterogeneous bibliographic networks with locally-trained embeddings." [Online]. Available: https://arxiv.org/abs/1803.03370

[31] S. T. Dumais and J. Nielsen, "Automating the assignment of submitted manuscripts to reviewers," in *Proc. 15th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 1992, pp. 233–244, doi: 10.1145/133160.133205.

[32] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre, "Assigning papers to referees," *Algorithmica*, vol. 58, no. 1, pp. 119–136, Sep. 2010, doi: 10.1007/s00453-009-9386-0.

[33] N. Craswell, D. Hawking, A. M. Vercoustre, and P. Wilkins, "P@NOPTIC expert: Searching for experts not just for documents," in *Proc. Ausweb Poster*, Brisbane, QLD, Australia, vol. 15, Apr. 2001, p. 17.

[34] K. Balog and R. M. de Rijke, "Determining expert profiles (with an application to expert finding)," in *Proc. IJCAI*, vol. 7, 2007, pp. 2657–2662.

[35] N. M. Kou, U. L. Hou, N. Mamoulis, and Z. Gong, "Weighted coverage based reviewer assignment," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2015, pp. 2031–2046, doi: 10.1145/2723372.2723727.

[36] B. Li and Y. T. Hou, "The new automated IEEE INFOCOM review assignment system," *IEEE Netw.*, vol. 30, no. 5, pp. 18–24, Sep./Oct. 2016, doi: 10.1109/MNET.2016.7579022.

[37] N. M. Kou, N. Mamoulis, Y. Li, Y. Li, and Z. Gong, "A topic-based reviewer assignment system," *Proc. VLDB Endowment*, vol. 8, no. 12, pp. 1852–1855, Aug. 2015, doi: 10.14778/2824032.2824084.

[38] S. Price and P. A. Flach, "Computational support for academic peer review: A perspective from artificial intelligence," *Commun. ACM*, vol. 60, no. 3, pp. 70–79, Mar. 2017, doi: 10.1145/2979672.

[39] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2007, pp. 500–509, doi: 10.1145/1281192.1281247.

[40] H. Peng, H. Hu, K. Wang, and X. Wang, "Time-aware and topic-based reviewer assignment," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, Mar. 2017, pp. 145–157, doi: 10.1007/978-3-319-63579-8_28.

[41] S. Ayaz, N. Masood, and M. A. Islam, "Predicting scientific impact based on *h*-index," *Scientometrics*, vol. 114, no. 3, pp. 993–1010, Mar. 2018, doi: 10.1007/s11192-017-2618-1.

[42] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[43] A. Bai, P. S. Deshpande, and M. Dhabu, "Selective database projections based approach for mining high-utility itemsets," *IEEE Access*, vol. 6, pp. 14389–14409, 2018, doi: 10.1109/ACCESS.2017.2788083.

[44] J. C. W. Lin, S. Ren, and P. Fournier-Viger, "MEMU: More efficient algorithm to mine high average-utility patterns with multiple minimum average-utility thresholds," *IEEE Access*, vol. 6, pp. 7593–7609, 2018, doi: 10.1109/ACCESS.2018.2801261.

[45] N. Lavrač, P. Flach, and B. Zupan, "Rule evaluation measures: A unifying view," in *Proc. Int. Conf. Inductive Logic Program.* Berlin, Germany: Springer, Jun. 1999, pp. 174–185, doi: 10.1007/3-540-48751-4_17.

[46] J. A. Foley, "Peer review, citation ratings and other fetishes," *Springer Sci. Rev.*, vol. 1, no. 2, pp. 5–7, Dec. 2013, doi: 10.1007/s40362-013-0003-x.

[47] L. Cagliero, P. Garza, A. Pasini, and E. M. Baralis, "Additional reviewer assignment by means of weighted association rules," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2018.2861214.

[48] I. Stelmakh, N. B. Shah, and A. Singh. (2018). "PeerReview4All: Fair and accurate reviewer assignment in peer review." [Online]. Available: https://arxiv.org/abs/1806.06237
[49] S. Thurner and R. Hanel, "Peer-review in a world with rational scientists: Toward selection of the average," *Eur. Phys. J. B*, vol. 84, no. 4, pp. 707–711, Dec. 2011, doi: 10.1140/epjb/e2011-20545-7.
[50] W. Thorngate and W. Chowdhury, "By the numbers: Track record, flawed reviews, journal space, and the fate of talented authors," in *Proc. Adv. Social Simulation*. Berlin, Germany: Springer, 2014, pp. 177–188, doi: 10.1007/978-3-642-39829-2_16.
[51] I. Vesper, "Peer reviewers unmasked: Largest global survey reveals trends," *Nature*, Sep. 2018, doi: 10.1038/d41586-018-06602-y.
[52] W. Gan, J. C. W. Lin, P. Fournier-Viger, H. C. Chieh, V. S. Tseng, and P. S. Yu. (2018). "A survey of utility-oriented pattern mining." [Online]. Available: https://arxiv.org/abs/1805.10511
[53] M. Ghasemi and M. Eidiani, "What's going on ResearchGATE," doi: 10.13140/2.1.3351.7125.
[54] A. Noruzi, "Google scholar: The new generation of citation indexes," *Libri*, vol. 55, no. 4, pp. 170–180, Dec. 2005, doi: 10.1515/LIBR.2005.170.
[55] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 990–998, doi: 10.1145/1401890.1402008.
[56] Clarivate Analytics. (2017). *Web of Science Core Collection Help*. Accessed: Apr. 23, 2018. [Online]. Available: https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html
[57] J. F. Burnham, "Scopus database: A review," *Biomed. Digit. Libraries*, vol. 3, no. 1, Dec. 2006, Art. no. 40869, doi: 10.1186/1742-5581-3-1.
[58] Y. Liu, W. Liao, and A. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, May 2005, pp. 689–695, doi: 10.1007/11430919_79.
[59] A. Martín-Martín, E. Orduna-Malea, M. Thelwall, and E. D. López-Cózar, "Google scholar, Web of science, and scopus: a systematic comparison of citations in 252 subject categories," *J. Informetrics*, vol. 12, no. 4, pp. 1077–1160, Nov. 2018, doi: 10.1016/j.joi.2018.09.002.

MUSA IBRAHIM MUSA ISHAG was born in North Darfur State, Sudan, in 1983. He received the B.Sc. degree in computer science from the Sudan University of Science and Technology, in 2006, and the master's degree in computer science from Chungbuk National University, South Korea, in 2010, where he is currently pursuing the Ph.D. degree in computer science. Since 2010, he has been a Lecturer with the College of Computer Science and Information Technology, Sudan University of Science and Technology, where he has been on unpaid leave, since 2013. He is a member of the IEEE and ACM.

KWANG HO PARK received the B.S. degree in biochemistry from Chungbuk National University, Cheongju, South Korea, in 2015, and the master's degree from the Database / Bioinformatics Laboratory, Chungbuk National University, in 2017, where he is currently pursuing the Ph.D. degree with the Database/Bioinformatics Laboratory. His main research interests include data mining, bioinformatics, and medical informatics.

JONG YUN LEE received the B.E. and M.E. degrees in computer engineering, and the Ph.D. degree from Chungbuk National University, South Korea, in 1985, 1987, and 1999, respectively. He worked as a Research/Project Leader with the Software Research and Development Institute, Hyundai Electronics Industrial Company Ltd., and Hyundai Information Technologies Company Ltd., South Korea, from 1990 to 1996. Also, he worked with BIT Computer Cooperation, in 1989. He had worked for the Department of Information and Communication Engineering, Kangwon National University (Samchek Campus), as an Assistant Professor, from 1999 to 2003. After that, he is currently a Full Professor with the Department of Software Engineering, Chungbuk National University, South Korea. His current research interests include biomedical databases, query processing and optimization techniques in spatiotemporal databases, data mining, and u-learning. He has been the President of the Korea Convergence Society, and the Director of the Korea Association of Computer Education, since 2010. He has served as an editorial board member for an international journal *INFORMATION*, in 2004, and a journal editorial member at the Korea Information Processing Society, from 2003 to 2006.

KEUN HO RYU received the Ph.D. degree in computer science and engineering from Yonsei University, South Korea, in 1988. He has served at the Reserve Officers' Training Corp (ROTC) of the Korean Army. He is also an Honorary Doctorate of the National University of Mongolia. He is currently a Professor with Chungbuk National University, South Korea, as well as the Faculty of Information Technology with Ton Duc Thang University, Vietnam. He has been the Leader of the Database and Bioinformatics Laboratory, South Korea, since 1986. He has worked at the University of Arizona, Tucson, AZ, USA, as a Postdoctoral and a Research Scientist, and also at the Electronics and Telecommunications Research Institute, South Korea, as a Senior Researcher. He is the former Vice-President of the Personalized Tumor Engineering Research Center.

He has published or presented over 1000 referred technical articles in various journals and international conferences, in addition to authoring a number of books. His research interests include temporal databases, spatiotemporal databases, temporal GIS, stream data processing, knowledge-based information retrieval, data mining, biomedical informatics, and bioinformatics. He has been a member of the IEEE and ACM, since 1982 and 1983, respectively. He has served on numerous program committees, including roles as the Demonstration Co-Chair of the VLDB, as the Panel and Tutorial Co-Chair of the APWeb, and as the FITAT General Co-Chair.