Social Networking, 2017, 6, 61-79
http://www.scirp.org/journal/sn
ISSN Online: 2169-3323
ISSN Print: 2169-3285

Bibliometry-Aware and Domain-Specific Features for Discovering Publication Hierarchically-Ordered Contexts and Scholarly-Communication Structures

Sulieman Bani-Ahmad

Department of Computer Information Systems, School of Information Technology, Al-Balqa Applied University, Salt, Jordan
Email: sulieman@bau.edu.jo

How to cite this paper: Bani-Ahmad, S. (2017) Bibliometry-Aware and Domain-Specific Features for Discovering Publication Hierarchically-Ordered Contexts and Scholarly-Communication Structures. *Social Networking*, **6**, 61-79.
http://dx.doi.org/10.4236/sn.2017.61005

Received: December 7, 2016
Accepted: January 13, 2017
Published: January 16, 2017

Abstract

Discovering publication hierarchically-ordered contexts is the main task in context-based searching paradigm. The proposed techniques to discover publication contexts relies on the availability of domain-specific inputs, namely a pre-specified ontology terms. A problem with this technique is that the needed domain-specific inputs may not be available in some scientific disciplines. In this paper, we propose utilizing a powerful input that is naturally available in any scientific discipline to discover the hierarchically-ordered contexts of it, namely paper citation and co-authorship graphs. More specifically, we propose a set of domain-specific bibliometry-aware features that are automatically computable instead of domain-specific inputs that need experts' efforts to prepare. Another benefit behind considering bibliometric-features to adapt to the special characteristics of the literature environment being targeted, which in turn facilitates contexts membership decision making. One key advantage of our proposal is that it considers temporal changes of the targeted publication set.

Keywords

Digital Libraries, Bibliometrics, Hierarchically-Ordered Contexts, Scholarly-Communication Structures, Citation Graphs, Co-Authorship Graphs

1. Introduction

In this paper, we aim at enhancing the accuracy of search results, *i.e.* finding relevant publications to a given keyword query by better capturing the notion of "publication importance". Due to the vast amount of literature work in all disciplines, keyword-based searching of digital libraries usually returns large number of relevant publications. User studies show that users usually view the first few results before rewording the keywords to obtain more documents that are relevant/more relevant documents [1]. Consequently, it was anticipated that ranking and sorting search results in terms of relevancy and quality to be useful as they.

Despite their relative success in web search engines, link-based ranking (or citation-based ranking in publications) approaches did not find acceptance in ranking publication for digital libraries [2]. The key reason may is that web gets larger with no quality control as the case in publications. Yet, publication citation-count, one basic citation-based ranking measure—is widely used in practice by academicians as an indicator of its influence to aid in tenure decisions [3].

Most of the well-known digital libraries, like ACM Portal [4] and Google Scholar [5] in computer science, and PubMed [6] in medical sciences, order their search results according to either

- The text-based relevancy score only, e.g., ACM Portal.
- Text-based relevancy and citation-based scores e.g., Google Scholar.
- The pre-assigned document ID as the case in PubMed.

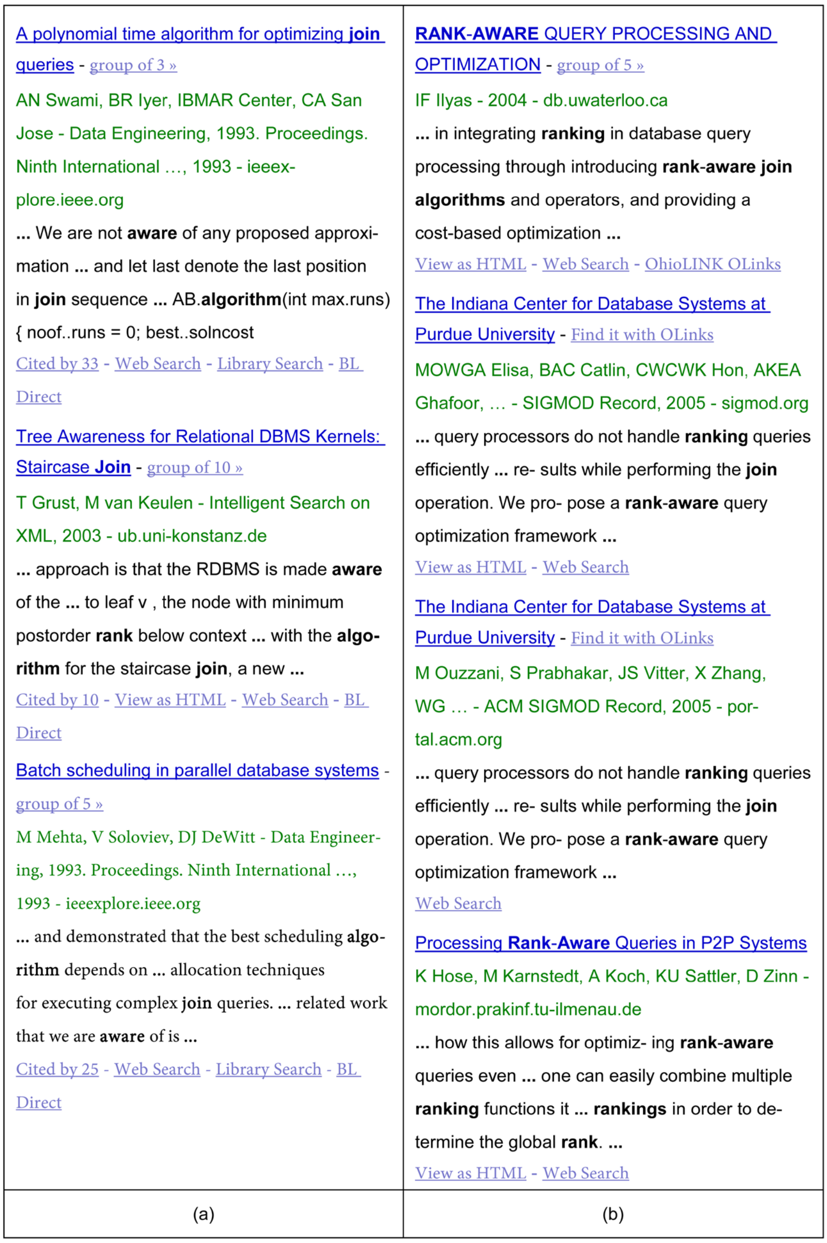
Practically, ranking publications in terms of citation-based scores faces accuracy-related problems that, if solved, will make it a standard in digital libraries design [2]. We believe that the reason behind the unsuccessfulness of citation-based ranking of publications is the complexity and special characteristics of literature environment. For instance, there are a number of quality indicators of publications need to be considered in the process of ranking publications, such as the time distribution of its in-citations. In addition, the bibliometric features of the field of study being targeted need to be considered when making raking decisions.

In this paper, we address the problem of ranking publications and propose techniques that help toward better ranking publications within hierarchically-ordered contexts. We start with an example that illustrates a problem that we refer to as the global ranking bias. After that, we illustrate the need for assigning publications to contexts to obtain scores that are more accurate and that considerably reduce the global ranking bias effect.

Utilization of citation networks is a common starting point among the proposed publication scoring measures [7] [8] [9] [10]. Variations in citation graph characteristics of different publication sets or subsets may negatively affect accuracy of assigned scores. The following example highlights this observation in Google Scholar.

Example 1:

Figure 1 shows sample results of querying Google Scholar with the phrase "rank aware join algorithms". Despite the low relevancy between the general "join algorithms" papers **that** appeared first (Figure 1(a)) and the search keywords submitted, the high citation-based scores of "join algorithms" papers



A polynomial time algorithm for optimizing **join** queries - group of 3 »

AN Swami, BR Iyer, IBMAR Center, CA San Jose - Data Engineering, 1993. Proceedings. Ninth International …, 1993 - ieeexplore.ieee.org

... We are not **aware** of any proposed approximation **...** and let last denote the last position in **join** sequence **...** AB.**algorithm**(int max.runs) { noof..runs = 0; best..solncost

Cited by 33 - Web Search - Library Search - BL Direct

Tree Awareness for Relational DBMS Kernels: Staircase **Join** - group of 10 »

T Grust, M van Keulen - Intelligent Search on XML, 2003 - ub.uni-konstanz.de

... approach is that the RDBMS is made **aware** of the **...** to leaf v , the node with minimum postorder **rank** below context **...** with the **algorithm** for the staircase **join**, a new **...**

Cited by 10 - View as HTML - Web Search - BL Direct

Batch scheduling in parallel database systems - group of 5 »

M Mehta, V Soloviev, DJ DeWitt - Data Engineering, 1993. Proceedings. Ninth International …, 1993 - ieeexplore.ieee.org

... and demonstrated that the best scheduling **algorithm** depends on **...** allocation techniques for executing complex **join** queries. **...** related work that we are **aware** of is **...**

Cited by 25 - Web Search - Library Search - BL Direct

(a)

RANK-AWARE QUERY PROCESSING AND OPTIMIZATION - group of 5 »

IF Ilyas - 2004 - db.uwaterloo.ca

... in integrating **ranking** in database query processing through introducing **rank-aware join algorithms** and operators, and providing a cost-based optimization **...**

View as HTML - Web Search - OhioLINK OLinks

The Indiana Center for Database Systems at Purdue University - Find it with OLinks

MOWGA Elisa, BAC Catlin, CWCWK Hon, AKEA Ghafoor, … - SIGMOD Record, 2005 - sigmod.org

... query processors do not handle **ranking** queries efficiently **...** re- sults while performing the **join** operation. We pro- pose a **rank-aware** query optimization framework **...**

View as HTML - Web Search

The Indiana Center for Database Systems at Purdue University - Find it with OLinks

M Ouzzani, S Prabhakar, JS Vitter, X Zhang, WG … - ACM SIGMOD Record, 2005 - portal.acm.org

... query processors do not handle **ranking** queries efficiently **...** re- sults while performing the **join** operation. We pro- pose a **rank-aware** query optimization framework **...**

Web Search

Processing **Rank-Aware** Queries in P2P Systems

K Hose, M Karnstedt, A Koch, KU Sattler, D Zinn - mordor.prakinf.tu-ilmenau.de

... how this allows for optimiz- ing **rank-aware** queries even **...** one can easily combine multiple **ranking** functions it **...** **rankings** in order to determine the global **rank**. **...**

View as HTML - Web Search

(b)

Figure 1. Searching Google Scholar for "rank aware join algorithms" (a) the first matches of the first page and (b) the first matches of the second page.

pushed them up in the result set. On the other hand, the low citation-based scores of the matches reached next (Figure 1(b)) pushed them down in the result set ordering although they are more relevant to the query keywords. This problem occurred due to what we refer to as global ranking bias effect, which results from comparing papers from different contexts together. Results of Figure 1(b) can be classified to the context of "rank-aware join algorithms" which in relatively new, so that they are not compared with the more general context of "join algorithms in relational databases" which has been in the literature for long time.

The scope of ranking measure may result in comparing publications from new subfields, which emerges rapidly, with the overlapping existing subfields. The problem may be more severe for digital libraries that contain publications from different sciences such as biochemistry, biology, etc. as is the case in PubMed.

Therefore, we propose that each paper should be evaluated in terms of importance by taking into account its context and the characteristics of the citation graph of its context(s) [1] [9] [11]. We define the context of paper P as the set of papers that have the same topic as P. Depending on how general or specific the topic is, P may be classified under more than one context in the context hierarchy. Even in the same level in the hierarchy, P may still be classified under more than one context with different degrees of relevancy.

The searching paradigm proposed in [12] reduced the global ranking bias effect by defining paper context utilizing domain-specific ontology terms [13]. Nevertheless, such predefined terms may not always be available. In this paper, we solely rely on relationships revealed from publication set. Citation and co-authorships relationships are examples of relationships naturally available in literature and can be utilized to discover paper contexts and organize the contexts into hierarchical order [8] [14].

Our approach of discovering paper contexts is of two stages. The first captures the author communities of the authors in the target publication set. The output of the first stage is used in the second stage. An author community is a system of scientists or scientist-units interacting frequently about shared topic(s) of research interests [15]. The second stage utilizes the collective paper-to-paper relationship revealed from both citation graph and author communities to discover paper contexts and organize the contexts into a proper hierarchy.

To rank publications within a context, we may imitate what HITS does in the web domain [16]. First, we perform text-based search to find relevant documents to the user's keywords as all search systems do [17]. Next, we analyze the citation graph extracted from the search result. This approach is exactly what HITS does [18]. Still, papers from different research domains are highly likely to appear in the search results for three reasons

1) Research domains of papers may overlap in most of the cases. One cannot put a clear-cut boundary when separating papers into subdomains.

2) Users are usually sensitive to time and efforts spent on finding information [19]. Thus, users usually do not provide enough information of what they have in mind that helps finding relevant papers accurately enough, and (iii) text-based search may return irrelevant papers problems of text search like synonymy, polysemy and context sensitivity results [17] [20].

We consider the different graph structures that can be inferred from the targeted publication set to locate paper contexts, and rank paper in its candidate context(s). Examples of such networks are paper citation graphs and author co-authorship and citation graphs. Paper contexts can be kept large or small depending on the application type. We also propose a technique to find optimal/reasonable size paper contexts. Our main contributions are as follows. We propose.

a) A set of author-author and paper-paper similarity/distance measures.

b) A set of bibliometric features that can be captured from the targeted publication set.

For the sake of evaluating the numerical distribution of the proposed feature formulas, we use three sets of publications set, the first is from the computer science field (around 87,000 articles are selected from ACM, IEEE and VLDB; we refer to this set the CS set). The second is from genomics area in life sciences (around 72,000 articles are selected from PubMed; we refer to this set the LS set), and the third is from data management (around 15,000 articles of ACM Anthology; we refer to this set the DM set). These articles were crawled, downloaded and parsed.

2. Overview of Our Proposal

Current ranking implementations assume large community of papers that can be scored using the same citation infrastructure. This leads to the global ranking bias. Motivated by the fact that citation relationship between papers gives a better clue of paper-paper similarity than text-based similarity, we automatically discover paper contexts and organize the discovered clusters into proper hierarchical order.

Assigning papers to contexts helps in enhancing search performance through better capturing their importance [21]. We refer to paper P score defined in P's context as P's local importance as opposed to global importance. Having papers scores defined within its context(s) reduces the probability of having heavily cited papers from being highly ranked for search queries where they minimal or no authority. This phenomenon is presented in example 1 in the introduction.

Classical documents clustering techniques uses document's features (words) to measure similarity between the documents. In [12] we use domain domain-specific hierarchical ontology terms to organize clusters into proper hierarchical order. In citation graph clustering though, we use three attributes of documents to perform clustering: a) in-citations b) out-citations c) scholarly communication links between papers. Based on these attributes, we propose a set of measures to estimate distances (similarities) between papers. Having done that, we use a properly selected clustering algorithm from the data mining literature to perform clustering, and thus discover paper contexts.

As an intermediate step in discovering paper contexts, we capture the scholarly-communication structure of the paper set in order to discover author communities. An author community is a set of authors that work in a common research domains.

Studying author communities helps:

1) Understanding the growth patterns of scholarly communication in different science disciplines, *i.e.* computer science, data management and medicine,

2) Discovering the relationships among research areas [15], which can be utilized to organize paper contexts into a proper hierarchical order.

One issue is the variance of clusters densities, as well as other network infrastructure properties, which makes cluster membership decision hard to take. The network infrastructure of citation and co-authorship graphs are the main concern of Bibliometrics. Bibliometrics goal is to study the process of written communication and of the nature of development of different disciplines [15]. We utilize a number of bibliometric features in making cluster membership decisions.

3. Experimental Sets and the Corresponding Database Schemas

We use three sets of publications to study the numerical distribution of the proposed features; namely, The (D)ata (M)anagment Set, the (L)ife (S)ciences Set and the (C)omputer (S)ciences Set. The DM Set is a collection of around 15,000 publications from the data management fields. The CS Set is a collection of around 87,000 publications from computer science fields, thus, the CS Set is more heterogeneous compared to the DM set. The LS Set is a collection of 72,000 publications from the genomics area, thus it is homogeneous like the DM set.

The three paper sets where parsed and a group of three databases of the extracted information from them were created.

Figure 2 displays how the number of publications per year changes in the three sets.

Observation 1: the number of publications per year parameter is steadier in the DM field than in the CS and LS sets.

Observation 2: the rate of increase in the publications per year significantly increases after year 1985 in the CS and LS fields.

4. Bibliometric Features of Targeted Publication Sets

In this section, we present a number of bibliometric features that can be utilized to decide on context membership decisions and computing similarity/distance scores between papers and between authors.

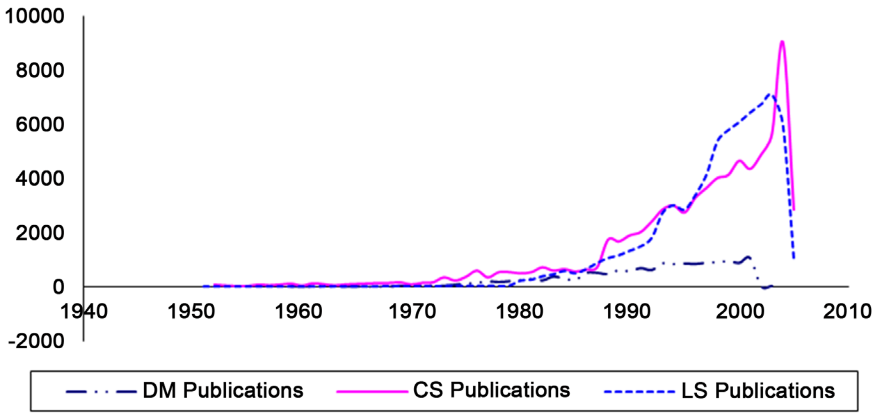


Figure 2. Publication-count-per-year change in the three datasets.

4.1. Paper-Paper and Author-Author Citation Graphs

In this section, we present the bibliometric features that can be extracted from the paper-paper citation curve. We will use the curves and measures presented later to discover paper contexts and author communities.

Different disciplines vary in terms of its nature and rate of development. To capture these two bibliometric features we define the *age of citation* curve. We define the age of citation $C_{P1\to P2}$ from paper *P*1 to *P*2 as the absolute difference between the publication years of *P*1 and *P*2. Citation age distribution graph plots the *age of citation* values vs. frequency of these values. Figure 3 shows the age of citation's distribution for the three paper sets.

Observation 1: In life sciences, authors tend to cite more up-to-date publications than authors in data management field of study.

We may also benefit from self-citation behavior of authors. Self-citation refers to the tendency of authors to cite their own work. One possible measure of *self-citation tendency* of author *A* is the Percentage of self-citations in *A*'s writings according to the following formula $\text{SCA}(A) = P_{A\to A}/P_A$ where $P_{A\to A}$ is the numbers of papers where *A* cites his own work, and C_A is the total number of *A*'s papers. Figure 4 shows the distribution of self-citation percentages for the three paper sets.

Observation 2: life scientists have more tendency to cite their own previous work than data management scientists.

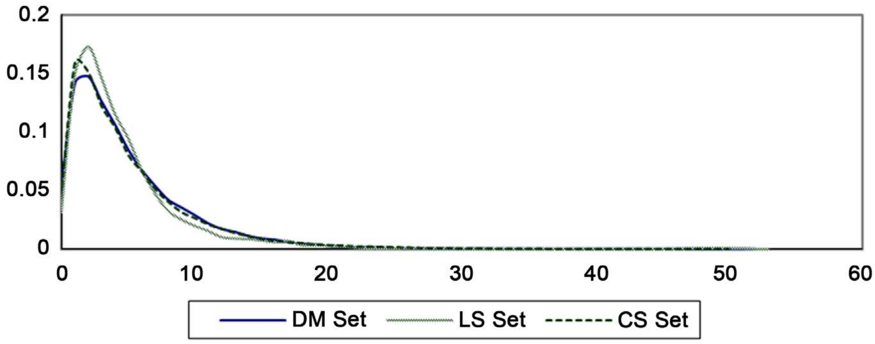


Figure 3. Citation age distribution of the three datasets.

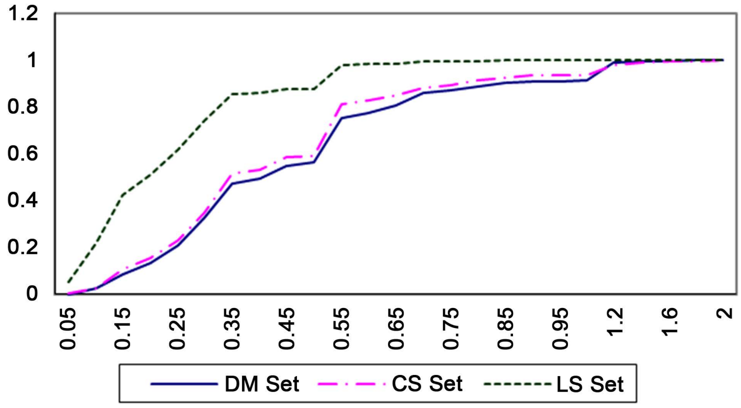


Figure 4. Self citation tendency in the three datasets.

4.2. Author Co-Authorship Graphs

Depending on the rate of growth of technology, and the need to rapidly publish papers in active research areas, authors tend to work jointly. Tendency to work jointly, or *collaborative tendency*, may vary from a discipline to another. One possible measure of collaborative tendency of author A is the size of A's Collaboration Group $\text{CG}(A)$. We define the collaboration group of A as the set of all authors that A has ever published a paper with Figure 5 shows the distribution of collaboration size distribution of the three paper sets.

Observation 3: LS researchers tend to have larger collaboration groups than CS and DM researchers.

Members of an author's collaboration graph may vary in collaboration levels. We define the collaboration level of author B to author A's collaboration group $\text{Cl}(B,A)$ as the ratio between the number of publication of A and B together $P_{A,B}$ and the total number of A's publications P_A, *i.e.* $\text{Cl}(B,A)=P_{A,B}/P_A$.

We may go further and define the Collaboration Level Distribution curve as shown in Figure 6. We may use this curve to check how abnormal the collaboration level between two authors in a particular discipline. Figure 6 shows the collaboration level distribution in the three paper sets.

Observation 4: DM set showed the highest collaboration levels. CS set comes next and the LS set is the lowest.

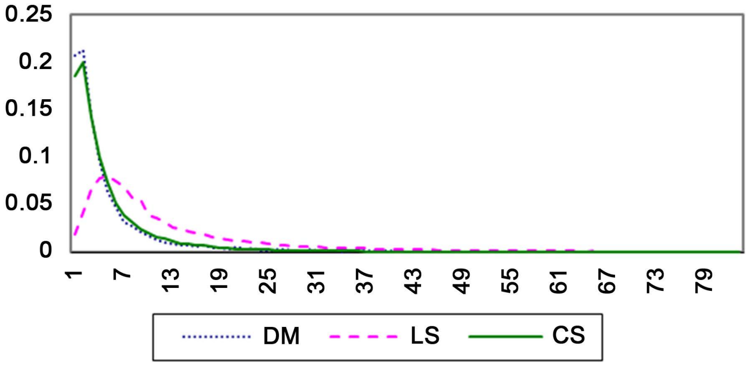


Figure 5. Collaboration set size distribution of the three datasets.

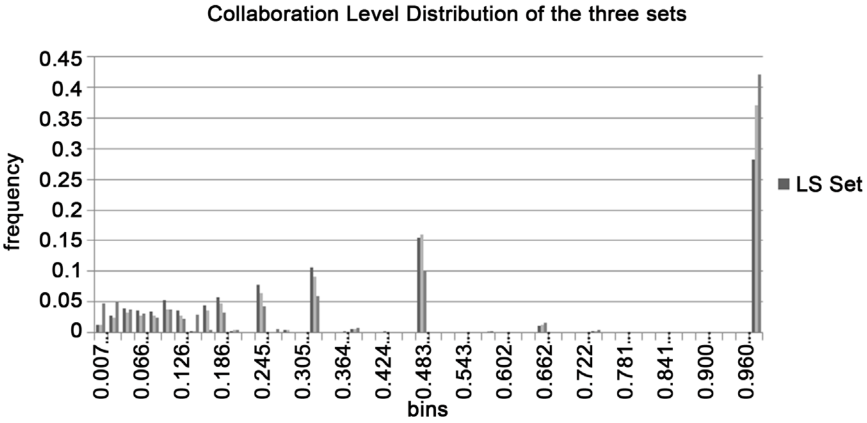


Figure 6. Collaboration level distribution reserved in the three datasets.

4.3. Research Productivity

One bibliometric feature that may vary from discipline to another is the productivity level of authors. One possible indicator of productivity level of authors is *publishing frequency curve*. The publishing frequency curve of author *A* is defined as the distribution of time spans between *A*'s consecutive publications. The time span between consecutive publications *P*1 and *P*2 of author *A* is computed as the absolute difference of *P*1 and *P*2's publication years. Short time spans between *A*'s publications is an indication of his productivity level. Figure 7 illustrates the frequency distribution of time spans in the three papers sets.

4.4. Co-Authorship Relationship

If two authors published common papers, then they probably work in the same research area and thus belong to the same community. Assume authors *A* and *B*, who has published $|P_A|$ and $|P_B|$ papers respectively, has published $|P_A \cap P_B|$ papers in common, then they probably belong to the same community *C* or $(A,B)\in C$. The probability $P((A,B)\in C)$ that these two authors belong to the same community, is directly proportional to *the percentage of common papers* (PCP) between *A* and *B* computed according to the following basic formula,

$$P\left(\left(A,B\right)\in C\right)\propto \mathrm{PCP}\left(A,B\right)=\left|P_A\cap P_B\right|/\left|P_A\cup P_B\right| \qquad (1)$$

To check how unusual the PCP between two particular authors is, or to say how significant the PCP value is, we prepare the PCP distribution as shown in Figure 8. The x-axis in the plots represents the PCP values observed in the corresponding paper set, and the y-axis represents the number of author couples that showed that PCP percentage, normalized by dividing it by the total number of author couples that showed non-zero PCP values.

We observe two types of collaborative couples in any publication set. One involves an advisor with his student, or advisor-student couple. The other involves an author with his college, or college-college couple. The advisor-student collaboration usually involves an unbalanced relationship, *i.e.* the common papers between the student and his advisor is all the student's papers, while they form a subset of the advisor's papers. In the case of college-college pair, the collaborative relationship may also be unbalanced, but usually not perfect.

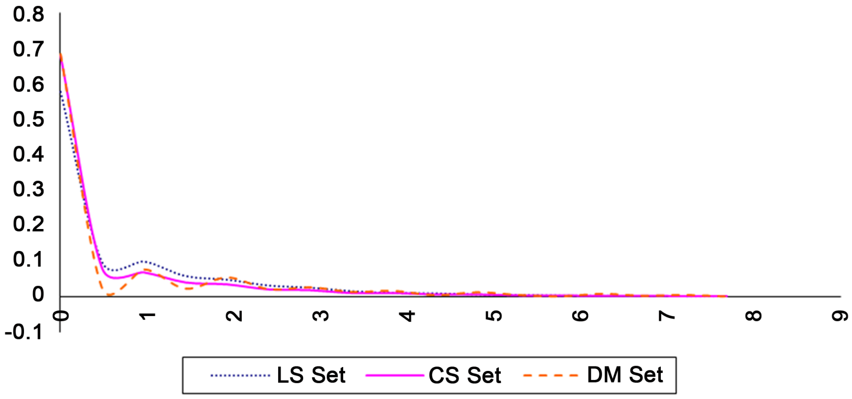
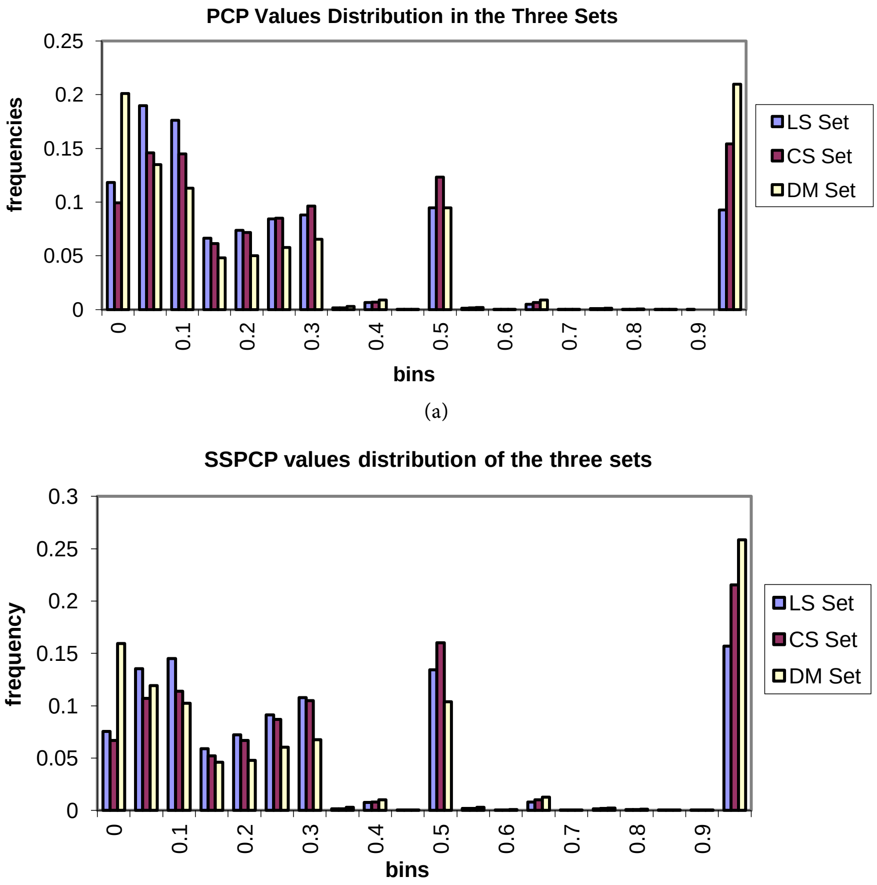


Figure 7. Publication frequency distribution of the three datasets.



(a)

(b)

Figure 8. PCP and SSPCP values distribution in the three sets.

To capture the unbalanced relationship of the advisor-student and college-college pairs, we define the Single Sided PCP, or SSPCP between author A and B, once from A's prospective and another from B's prospective. The SSPCP from A's prospective can be computed as

$$\mathrm{SSPCP}_A\left(A,B\right)=\left|P_A\cap P_B\right|/\left|P_A\right| \qquad (2)$$

Similarly, we can compute $\mathrm{SSPCP}_B\left(A,B\right)$ as

$$\mathrm{SSPCP}_B\left(A,B\right)=\left|P_A\cap P_B\right|/\left|P_B\right|.$$

Formula (2) suggests that, a perfect or nearly perfect $\mathrm{SSPCP}_A\left(A,B\right)$ with low $\mathrm{SSPCP}_B\left(A,B\right)$ scores indicate that A and B forms an advisor-student-like couple, with A being the student and B being the advisor. It also indicates the following:

1) B belongs to more than one community with different probabilities.

2) The probability that A belongs to one (or more) of B's candidate communities is very high.

3) A may not alone help us decide upon to which community B belongs most.

In the other hand, the Formula (2) suggests that as the difference between $\mathrm{SSPCP}_A\left(A,B\right)$ and $\mathrm{SSPCP}_B\left(A,B\right)$ scores becomes less than a certain threshold α, this difference gives a clue of how likely author A and B belong to the same community. But still, A may not alone help us decide upon which community B belongs most, or vise versa. We observed that $\alpha = 0.5$ in the three publication sets.

To illustrate more, we discuss three possible scenarios that may occur. The scenarios are presented in the following table:

Case	P_A	P_B	$\lvert P_A \cap P_B \rvert$	A $\mathrm{SSPCP}_A(A,B)$	PCP	B $\mathrm{SSPCP}_B(A,B)$	Observations
$\lvert P_A \rvert \gg \lvert P_B \rvert$	30	5	5	5/30	5/30	5/5	$\mathrm{PCP} = \mathrm{SSPCP}_A(A,B)$ $\mathrm{SSPCP}_A(A,B) = 1$ $\lvert \mathrm{SSPCP}_A(A,B) - \mathrm{SSPCP}_B(A,B) \rvert > 0.5$
$\lvert P_A \rvert > \lvert P_B \rvert$	20	10	4	4/20	4/26	4/10	$\lvert \mathrm{SSPCP}_A(A,B) - \mathrm{SSPCP}_B(A,B) \rvert < 0.5$
$\lvert P_A \rvert \cong \lvert P_B \rvert$	10	9	4	4/10	4/15	4/9	$\lvert \mathrm{SSPCP}_A(A,B) - \mathrm{SSPCP}_B(A,B) \rvert \cong 0.0$

From Figure 8, we notice that the distribution can be divided into three different areas.

- The first is the area where PCP and SSPCP are near perfect. Most of the author couples that lies within this area are of type advisor-student. Notice that in the DM field, more research is conducted in the setting of advisor-student. While in the LS field, research is conducted in variety of settings other than advisor-student, for example, research in LS involves lab technicians and clinicians. This maps to the $\lvert P_A \rvert \gg \lvert P_B \rvert$ case in the above table.
- The second is just in the middle where PCP and SSPCP value = 0.5. This PCP/SSPCP occurs when the common papers are half as much as the total number of both authors or one of the authors. This maps to the $\lvert P_A \rvert \cong \lvert P_B \rvert$ case in the above table.
- The third, which showed the widest distribution of PCP and SSPCP over the interval [0, 0.3]. This maps to the $\lvert P_A \rvert > \lvert P_B \rvert$ case in the above table.

We notice that, as the difference between the author couples becomes less than 0.5, we can safely use SSPCP as an indicator of how likely A and B belong to the same community. However, when the case is and advisor-student case, we need to consider, when computing the final PCP score, the unbalanced relationship between the author couples.

One question that is left is how to compute the final PCP score of authors A and B from $\mathrm{SSPCP}_A(A,B)$ and $\mathrm{SSPCP}_B(A,B)$ scores.

We may think of the relationship between authors A and B as a two dimensional relationship. The strength of this relationship is determined by combining the significance of the SSPCP values of the two authors.

The significance of an SSPCP value, or $\mathrm{Sig}(\mathrm{SSPCP}_{A \text{ or } B}(A,B))$, can be computed based on a set of mapping functions:

The Raw SSPCP Value

In this approach, we use the SSPCP score as it is, in this case the higher SSPCP becomes, the closer the authors becomes to each other. *i.e.*

$$\operatorname{Sig}\left(\text{SSPCP}_{A\text{ or }B}(A,B)\right)=\text{SSPCP}_{A\text{ or }B}(A,B) \tag{3}$$

A problem with this approach is that it does not explicitly consider the bibliometric features of the publication set.

Frequency of SSPCP Value

The frequency of observing the value of SSPCP in the publication set, or $f\left(\text{SSPCP}_{A\text{ or }B}(A,B)\right)$, can be used to infer the significance of, *i.e.*

$$\operatorname{Sig}\left(\text{SSPCP}_{A\text{ or }B}(A,B)\right)=f\left(\text{SSPCP}_{A\text{ or }B}(A,B)\right) \tag{4}$$

The motivation here is that scores that rarely occur are not informative. In this case, SSPCP values within the intervals [0.35, 0.5[and]0.5,1[will be almost zero. This measure suggests that more rare SSPCP values are less significant than common ones.

The P-Value of SSPCP Score

The P-Value of a score v measures the probability of the following random event:

"When randomly selecting author couples A and B from the publication set, what is the probability of observing an $\text{SSPCP}_A(A,B)\geq v$ or higher", *i.e.*

$$\operatorname{Sig}(x=v)=\int_{x=v}^{\infty} f(x)\,\mathrm{d}x \tag{5}$$

where x is a dummy variable that represents the SSPCP values and $f(x)$ is the frequency of observing x in the publication set.

Note: This measure is very useful when the distribution of measure we target (in this case it is SSPCP) follows the Zipf distribution.

The Z Score of SSPCP Value

One technique to isolate extreme scores and reduce their effect on the distribution is to compute the Z scores. We use the following Z score formula from [22],

$$Z(v)=\frac{v-m_{\text{SSPCP}}}{S_{\text{SSPCP}}} \tag{6}$$

where m_{SSPCP} is the mean of the observed SSPCP values, and S_{SSPCP} is the mean absolute-deviation which is defined as follows:

$$S_{[\text{SSPCP}]}=1/n\sum_{x_i\in[\text{SSPCP}]}\left(x_i-m_{\text{SSPCP}}\right)$$

where $[\text{SSPCP}]$ is the vector of all observed SSPCP values.

Back to our question of how to combine the two SSPCP scores into a single PCP score. One possible way to compute $P(A\leftrightarrow B)$ is according to the Pythagorean Theorem, *i.e.*

$$P\left((A,B)\in C\right)=\sqrt{\operatorname{Sig}\left(\text{SSPCP}_A(A,B)\right)^2+\operatorname{Sig}\left(\text{SSPCP}_B(A,B)\right)^2}\Big/\sqrt{2} \tag{7}$$

The $\sqrt{2}$ is used as a normalizing factor which occurs when the both SSPCP are perfect (=1).

One problem of the relying on co-authorship only is that two authors from different disciplines may have common papers. As an example, a database researcher may write a common work in bioinformatics with a professor in the medical school. *A* statistician may publish a common paper with a researcher in nursing or other disciplines where statistical analysis is needed. One way to reduce the effect of this problem is to consider what we refer to as the *angle between authors*.

To illustrate the concept of the *angles between authors*, we discus one possible way to measure the angle between author *A* and *B*. in this way we utilize the citation relationships between authors. Denote the expressions $\mathrm{Sig}\left(\mathrm{SSPCP}_A\left(A,B\right)\right)$, $\mathrm{Sig}\left(\mathrm{SSPCP}_B\left(A,B\right)\right)$ and $\sqrt{\mathrm{Sig}\left(\mathrm{SSPCP}_A\left(A,B\right)\right)^2+f_{\mathrm{pcp}}\left(\mathrm{SSPCP}_B\left(A,B\right)\right)^2}$ by A°, B° and C° respectively. The expression C° is nothing but the length of the third edge opposite to the right angle as shown in Figure 9(a). If we think of the angle between A° and B° as the level of citation relationship between authors *A* and *B*, then we can generalize ($P\left(\left(A,B\right)\in C\right)$.a) to consider the citation relationship between authors as follows:

- If author *A* and *B* are coauthors in a subset of their publications, and they cite each other's works relatively frequently, then they more likely belong to the same community. In this case, the angle between the edges A°, B° will be small and C° will be long indicating higher probability of *A* and *B* belonging to the same community (see Figure 9(c)).
- On the other hand, if authors *A* and *B* are coauthors in a subset of their publications and they cite each other's works relatively rarely, then they more likely belong to two different ICs. In this case, the angle between the edges A°, B° will be large and C° will be short indicating lower probability of *A* and *B* belonging to the same community (see Figure 9(b)).

Consequently, ($P\left(\left(A,B\right)\in C\right)$.a) can be rewritten as follows

$$P\left(\left(A,B\right)\in C\right)=\sqrt{f_{\mathrm{PCP}}\left(\mathrm{SSPCP}_A\left(A,B\right)\right)^2+f_{\mathrm{PCP}}\left(\mathrm{SSPCP}_B\left(A,B\right)\right)^2+2\cdot f_{\mathrm{PCP}}\left(\mathrm{SSPCP}_A\left(A,B\right)\right)\cdot f_{\mathrm{PCP}}\left(\mathrm{SSPCP}_B\left(A,B\right)\right)\cdot\mathrm{Cos}\,\theta_{A,B}}\Big/2 \quad (8)$$

The number 2 in the denominator is used as a normalizing factor. In the case when the both SSPCP are perfect (=1) and the angle $\theta_{A,B}$ is 0, the final score will be 1. Based on the above discussion, we propose the following basic formula to compute $\theta_{A,B}$,

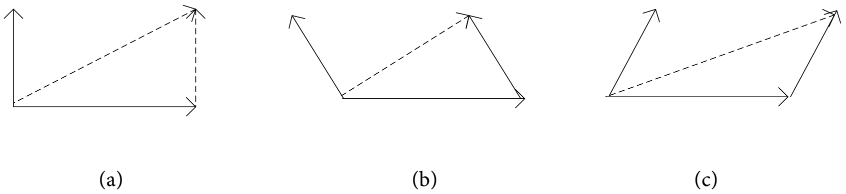

(a) (b) (c)

Figure 9. Three different cases of SSPCP summation.

$$\theta_{A,B} = \mathrm{Max}\left(\left|\mathrm{CS}(A) \cap P_B\right| / \left|P_B\right|, \left|\mathrm{CS}(B) \cap P_A\right| / P_A\right) \cdot \pi \tag{9}$$

where $\mathrm{CS}(A)$ ($\mathrm{CS}(B)$ is similar) is the *citation space* (CS) of *A*, which is the set of papers that A cites in his work.

$\left|\mathrm{CS}(A) \cap P_B\right|$ represents the number of papers written by B are cited by *A*.

We notices that $\theta_{A,B}$ ranges between 0, in the case of perfect relatedness between *A* and *B*, and π when no citation relationship observed between *A* and *B*.

We may also consider the age of citations between authors *A* and *B*. One-way to do this is to utilize the citation age factor $r_{\text{c-age}}$ which we present the definition of in the next subsection.

$$\theta_{A,B} = \left(1 - \mathrm{Max}\left(r_{\text{c-age}}(A \to B), r_{\text{c-age}}(A \leftarrow B)\right)\right) \cdot \pi \tag{10}$$

Other ways to measures the angle between authors A and B are:

The Relative Distance Based on the SSPCP Vectors of the Publication Set

For any author couples *A* and *B*, the higher the difference between $\mathrm{SSPCP}_A(A,B)$ and $\mathrm{SSPCP}_B(A,B)$ becomes, the lower the probability that *A* and B belongs to the same community becomes.

The relative distance between $\mathrm{SSPCP}_A(A,B)$ and $\mathrm{SSPCP}_B(A,B)$ as follows.

$$REDist_{\text{SSPCP}}(A,B) = \frac{\left|\mathrm{SSPCP}_A(A,B) - \mathrm{SSPCP}_B(A,B)\right|}{\text{Euclidian Distance}\left([\mathrm{SSPCP}_A],[\mathrm{SSPCP}_B]\right) / \left\|[\mathrm{SSPCP}_A]\right\|} \cdot \pi \tag{11}$$

where $\text{Euclidian Distance}\left([\mathrm{SSPCP}_A],[\mathrm{SSPCP}_B]\right)$ is the Euclidian Distance between the vector of all observed SSPCP values of *A* prospective ($([\mathrm{SSPCP}_A])$) and B prospective ($([\mathrm{SSPCP}_B])$). We divide it by $\left\|[\mathrm{SSPCP}_A]\right\|$ which represents the number of author couples in either of the SSPCP vectors.

Formula (11) suggests that, as $\left|\mathrm{SSPCP}_A(A,B) - \mathrm{SSPCP}_B(A,B)\right|$ increases, we conclude that Formula (8) is less likely to be a good clue of how related authors *A* and B to each other, and thus gives less weight to the it.

Citation Exchange between A and B

We may use citation exchange between *A* and B as presented in ($\theta_{A,B}$.a) and ($\theta_{A,B}$.b).

Citation Space Difference between A and B

Citation space of an author *A* is the set of papers that *A* cites in his publications as we stated before. To compute the distance between *A* and B we consider the citations of the papers that are not common between *A* and *B*. *A* basic formula to compute the angle between authors *A* and B based on citation space difference is:

$$CitSD(A,B) = \frac{\left|\left\{\mathrm{CS}(A) \cap \mathrm{CS}(B)\right\} - C\left(P_{A,B}\right)\right|}{\left|\left\{\mathrm{CS}(A) \cup \mathrm{CS}(B)\right\} - C\left(P_{A,B}\right)\right|} \cdot \pi \tag{12}$$

where:

$\mathrm{CS}(A) \cap \mathrm{CS}(B)$ the overlapping between the citation spaces of *A* and *B*.

$C\left(P_{A,B}\right)$ is the citations of the common papers between A and B (excluded).

$\mathrm{CS}(A)\cup\mathrm{CS}(B)$ the total set of citations from the citation spaces of both A and B.

The reason for excluding the citations of the common publications between the two authors is to identify authors who belong to different communities like the case of a researcher from the computer science domain publishing a paper with a researcher from the biomedical science domain when the paper is dealing with a topic from bioinformatics. Excluding the citations of the common papers of bioinformatics, we expect that the computer science researcher cites different papers than those cited by the biomedical specialist.

We may weigh a citation c according to how many times does c appear in the citation space of the author as follows:

$$CitSD(A,B)=\frac{\sum\limits_{C_i\in\left[\{\mathrm{CS}(A)\cap\mathrm{CS}(B)\}-C(P_{A,B})\right]} w(C_i)}{\sum\limits_{C_i\in\left(\{\mathrm{CS}(A)\cup\mathrm{CS}(B)\}-C(P_{A,B})\right)} w(C_i)}\cdot\pi \tag{13}$$

where $\left[\{\mathrm{CS}(A)\cap\mathrm{CS}(B)\}-C\left(P_{A,B}\right)\right]$ is the set of common citations between the citation spaces of A and B excluding the citations of the common publications of A and B. And $\{\mathrm{CS}(A)\cup\mathrm{CS}(B)\}$ is all the citations in the citation spaces of two spaces of A and B.

Second Level of Collaborative Set Difference

The second level collaborative set of author A is defined as the collaborative sets of all authors that collaboratively worked with A. we may use this measure to identify those authors who belong to different communities but still have common publications. A basic formula to measure this parameter is:

$$L2ColSD(A,B)=\frac{\left|\{L2ColS(A)\cap L2ColS(B)\}-\left[L1ColS(A)\cap L1ColS(B)\right]\right|}{\left|\{L2ColS(A)\cup L2ColS(B)\}-\left[L1ColS(A)\cap L1ColS(B)\right]\right|}\cdot\pi \tag{14}$$

where:

$L2ColS(A)$ and $L1ColS(A)$ are the second and first level collaboration set of A.

$\{L2ColS(A)\cup L2ColS(B)\}-\left[L1ColS(A)\cap L1ColS(B)\right]$ is the set of common authors between $L2ColS(A)$ and $L2ColS(B)$ excluding those common authors from the first level.

We may also weigh the second level author x in the collaboration set of A by the number of common publications between x and the first level author(s) as follows.

$$L2ColSD(A,B)=\frac{\sum\limits_{w}\left[L2ColS(A)\cap L2ColS(B)-\left[L1ColS(A)\cap L1ColS(B)\right]\right]}{\sum\limits_{w}\left[L2ColS(A)\cup L2ColS(B)-\left[L1ColS(A)\cap L1ColS(B)\right]\right]}\cdot\pi \tag{16}$$

Another problem of relying on the co-authorship relationship between authors prevents discovering authors who belong to the same community when they have no common publications. To overcome this problem, we utilize another relationship that is based on citation relationship between authors. Details are presented in the next subsection.

4.5. Author-to-Author Citation Relationship

If two authors directly or indirectly cite each other's works, then probably these two authors belong to the same community.

One possible measure of citation relationship strength between authors *A* and *B* is the *Bidirectional Citation Bandwidth* (C_{2BW}). The bidirectional citation bandwidth between authors *A* and *B* is defined , from *A*'s prospective, as the percentage of citation exchange between *A* and *B* (from publications of *A* to *B* and vise versa) to the total citation exchange between *A*'s work and all other authors' work citing or cited by *A*'s work. The following formula clarifies the way to compute $C_{2BW}(A,B)$

$$C_{2BW}(A,B)=\frac{C_{A\to B}+C_{B\to A}}{C_{A\to}+C_{\to A}} \tag{17}$$

where $C_{A\to B}$ and $C_{B\to A}$ are the citation exchange from *A*'s publications to B's publications. $C_{A\to}$ and $C_{\to A}$ are the total in and out citations to and from *A*'s publications.

Similarly, we may compute $C_{2BW}(B,A)$, this time from *B*'s prospective, according to the following formula

$$C_{2BW}(B,A)=\frac{C_{B\to A}+C_{A\to B}}{C_{B\to}+C_{\to B}} \tag{18}$$

where $C_{B\to}$ and $C_{\to B}$ are the total in and out citations to and from *A*'s publications.

We assumed here that citing and the cited works are topically related. However, citation-based relations between papers are often criticized on the ground that citation may not actually represent, due to topic diversity of paper citations, topic-relationship between the source and the destination of citation [8] [15] [23]. To reduce the effect of topic diversity in paper citation we utilize a number of heuristics to weight citations according to the topic-relatedness between the citing and the cited publications.

One possible indicator of the topical relatedness of citations between authors is the *level* 2 *citation relationship strength*. Level 2-citation-relationship strength between authors *A* and B is defined as the overlapping ratio between out citations of *A*'s publications and out citations of *B*'s publications. Denoting *A*'s and *B*'s out citation count by $C_{A\to}$ and $C_{B\to}$ respectively, the level 2 citation relationship strength between *A* and *B* can be computed using the formula $C_{2OL}(A,B)=(C_{A\to}\cap C_{B\to})/\min(C_{A\to},C_{B\to})$. Using the same scenario (3.a) shown above is derived; we derive $P((A,B)\in C)$ based on citation relationship between author *A* and *B* as follows

$$P((A,B)\in C) = \sqrt{f_{C_{2BW}}(C_{2BW}(A,B))^2+f_{C_{2BW}}(C_{2BW}(B,A))^2+2\cdot f_{C_{2BW}}(C_{2BW}(A,B))\cdot f_{C_{2BW}}(C_{2BW}(B,A))\cdot \mathrm{Cos}\,\omega_{A,B}}\Big/2 \tag{19}$$

where $\omega_{A,B}$ is computed as

$$\omega_{A,B} = \left(1 - C_{2OL}\left(A,B\right)\right) \cdot \pi$$

Notice that $\omega_{A,B}$ ranges from 0 to π depending on how strong the level-2 citation relationship between A and B is. The weaker the level-2 citation relationship between authors is, the bigger the angle $\omega_{A,B}$ becomes, and consequently $P\left(\left(A,B\right) \in C\right)$ becomes smaller if the bidirectional citation bandwidth remains unchanged.

One indicator of topic-relatedness between the citing and the cited papers is the *age of citation*. We define the age of a citation as the absolute difference between the publication years of the citing and the cited papers. The effect citation age on the topic-relatedness clearly appears in disciplines that are technology driven like computer science.

Different disciplines vary in terms of its nature and rate of development. To capture these two bibliometric differences we define the *age of citation* curve. We define the age of citation $C_{P1 \to P2}$ from paper *P*1 to *P*2 as the absolute difference between the publication years of *P*1 and *P*2. Citation age distribution graph $f_{cg}\left(t\right)$ relates the *age of citation* values vs. frequency of these values. Figure 3 shows the age of citation's distribution for the three paper sets.

Notice that the impact of a citation C_i from a work P_A of author A to a work P_B author B to the similarity between A and B is a) inversely proportional to the duration between the two connected works, *i.e.* the publication date of P_A and P_B. b) also inversely proportional to the frequency of having two citations in that paper set $f_{cg}\left(t\right)$, where $t = \left|T\left(P_B\right) - T\left(P_A\right)\right|$, $T\left(P_x\right)$ is the publication date of P_x. And c) directly proportional to the percentage of citations from A to B with duration t or $n_{C_{A \to B/t}}$ to the total number of citations from A to B $n_{C_{A \to B}}$. We refer to this ratio as the *citation-age factor* of related works of authors A and B, which is computed as

$$r_{c-\text{age}} = \frac{n_{C_{A \to B/t}}}{n_{C_{A \to B}}} \times \left(1 - f_{cg}\left(t\right)\right)$$

we involved the frequency of having citations with age t in the targeted publication set as stated in item b). Thus, the probability that A and B belong to the same community, or the relationship strength between A and B based on the citations from A's works to B's works can be computed as

$$\begin{aligned} r_{c-\text{age}}\left(A \to B\right) &= \sum_{\text{all } t_i' \text{ from } A \text{ to } B} \frac{n_{C_{A \to B/t_i}}}{n_{C_{A \to B}}} \times \left(1 - f_{cg}\left(t\right)\right) \\ &= 1/n_{C_{A \to B}} \sum_{\text{all } t_i' \text{ from } A \text{ to } B} n_{C_{A \to B/t_i}} \left(1 - f_{cg}\left(t\right)\right) \end{aligned}$$

The citation age curve $f_{cg}\left(t\right)$ is one of the bibliometric features that depend on the targeted publication set. Similarly, one may compute $r_{c-\text{age}}$ for citations from the opposite direction, *i.e.* from B's works to A's. The case of having two authors citing each other's works will be given more weight based on the similarity measure proposed. We refer to this phenomenon by author's *citation-backward loop*.

5. Conclusion

Discovering publication hierarchically-ordered contexts is a key task in context-based searching paradigm. Discover publication contexts and author communities (*i.e.*, Scholarly-Communication Structures) rely on the availability of domain-specific inputs that need experts' efforts to prepare. However, the needed domain-specific inputs may not be available in some scientific disciplines. In this paper, we proposed utilizing a powerful input that is naturally available in any scientific discipline to discover the hierarchically-ordered contexts of it, namely paper citation and co-authorship graphs. More specifically, we proposed a set of domain-specific bibliometry-aware features that are automatically computable instead of domain-specific inputs that might not be available or difficult to prepare. Another benefit behind considering bibliometric-features to adapt to the special characteristics of the literature environment being targeted, which in turn facilitates contexts membership decision making. Another key advantage of our proposal is that it considers temporal changes of the targeted publication set.

References

[1] Bani-Ahmad, S. and Özsoyoglu, G. (2007) Improved Publication Scores for Online Digital Libraries via Research Pyramids. *The European Conference on Research and Advanced Technology for Digital Libraries* (*ECDL*), Budapest, 16-21 September 2007, 50-62. https://doi.org/10.1007/978-3-540-74851-9_5

[2] Peter's Digital Reference Shelf, A Report on Google Scholar, 2004.

[3] Bauer, K. and Bakkalbasi, N. (2005) An Examination of Citation Counts in a New Scholarly Communication Environment, D-Lib Magazine. https://doi.org/10.1045/september2005-bauer

[4] ACM Digital Library. http://www.acm.org/dl

[5] Google Scholar. http://scholar.google.com

[6] The National Library of Medicine, Entrez PubMed. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[7] Bani-Ahmad, S., Cakmak, A., Özsoyoglu, G. and Al-Hamdani, A. (2005) Evaluating Publication Similarity Measures. *IEEE Data Engineering Bulletin*, **28**, 21-28.

[8] Bani-Ahmad, S., Cakmak, A., Al-Hamdani, A. and Özsoyoglu, G. (2005) Evaluating Score and Publication Similarity Functions in Digital Libraries. *The International Conference on Asian Digital Libraries* (*ICADL*), Bangkok, December 12-15 2005, 483-485. https://doi.org/10.1007/11599517_66

[9] Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, **30**, 107-117. https://doi.org/10.1016/S0169-7552(98)00110-X

[10] Geerts, F., Mannila, H. and Terzi, E. (2004) Relational Link-Based Ranking. *Proceedings of VLDB* 2004, 31 August-3 September 2004. https://doi.org/10.1016/b978-012088469-8.50050-4

[11] Broder, A.A. (2002) Taxonomy of Web Search. *ACM SIGIR Forum*, **36**, 3-10. https://doi.org/10.1145/792550.792552

[12] Ratprasartporn, N., Po, J., Cakmak, A., Bani-Ahmad, S. and Ozsoyoglu, G. (2007) On Context-Based Publication Search Paradigm: Gene-Ontology-Specific Contexts for Searching PubMed Effectively. ICDE.

[13] ACM SIGMOD Anthology. http://www.acm.org/sigmod/dblp/db/antho logy.html

[14] Chakrabarti, S. (2003) Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, Burlington.

[15] Christine, L. (1990) Borgman, Scholarly Communication and Bibliometrics. SAGE Publications, Thousand Oaks.

[16] Cakmak, A. (2003) HITS- and PageRank-Based Importance Score Computations for ACM Anthology Papers. Technical Report, EECS Department, CWRU, Cleveland.

[17] Chakrabarti, S. (2000) Data Mining for Hypertext: A Tutorial Survey. *ACM SIGKDD Explorations Newsletter*, **1**, 1-11. https://doi.org/10.1145/846183.846187

[18] Ding, C., Zha, H., He, X., Husbands, P. and Simon, H. (2003) Hubs and Authorities on the World Wide Web, Technical Report 47847. Technical Report, Lawrence Berkeley National Laboratory (LNBL), Berkeley.

[19] Tang, M.C. and Sun, Y. (2003) Evaluation of Web-Based Search Engines Using User-Effort Measures. *Library and Information Science Research Electronic Journal*, **13**, 44-54.

[20] Bani-Ahmad, S. and Al-Dweik, G. (2010) On Improved Example-Based Search in Digital Libraries via Term Ranking. *International Journal of Theoretical and Applied Information Technology* (*JATIT*), **19**.

[21] Haveliwala, T.H. (2002) Topic-Sensitive PageRank. *Proceedings of the* 11*th International World Wide Web Conference*, Honolulu, 7-11 May 2002, 517-526.

[22] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, Burlington.

[23] Al-Hamdani, A. (2003) Querying Web Resources with Metadata in a Database. PhD Thesis, Case Western Reserve University (CWRU), Cleveland.