http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (6): 1093-1103, 2010
ISSN 1812-5638
© 2010 Asian Network for Scientific Information

On Using the Research-Pyramid Model to Enhance Literature Digital Libraries

[1]Sulieman Bani-Ahmad and [2]Gultekin Ozsoyoglu
[1]Department of Information Technology, Al-Balqa Applied University, Salt Campus, Jordan
[2]Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, Ohio

Abstract: We validate the research pyramid model of research evolution. Moreover, we propose and evaluate two algorithms to identify research pyramids. Finally, we improve publication scores in terms of accuracy and separability via publications' research pyramids. Accurately ranking publications enables users to aggregate pertinent results quickly and easily. Studies show that citation-based publication-importance functions, e.g., PageRank and Citation Count, are extremely skewed and have accuracy problems. Based on the notion of research pyramids we propose a priori technique to assign more effective and accurate publication importance scores. We showed that the proposed technique provides more accurate and significantly less skewed publication scores than citation-based techniques. Our experiments showed 16-25% improvement in search outputs accuracy measured for the top-k search results.

Key words: Research-pyramids, the RP-model, citation, graph analysis, topic diffusion, digital libraries

INTRODUCTION

Searching On-Line Literature Digital Libraries (OLDLs) efficiently and effectively is becoming more and more important as the size and use of OLDLs expand at a very high rate. Consider the following three OLDLs as examples from computer life sciences and from electrical engineering fields:

- In computer science, ACM digital library (ACM) has around one million full-text publications collected over fifty years, all available to search and download
- In electrical engineering and computer science, IEEE xplorer (IEEE), is another OLDL that provides its users with access over the web to more than 1,700 selected conferences proceedings
- ScienceDirect (ScienceDirect) is the world's leading scientific, technical and medical information resource that celebrated its billionth article download back in November'06 since, it has been launched and put into service in 1999

From the above stated numbers one may come to a conclusion that providing accurate publication importance scores for search results and ranking publications returned as search results accurately can significantly help OLDL users in reducing the time they spend in searching OLDLs. Furthermore, accurate and effective publication rankings can also be useful for comparative assessments of publications as well as publication venues and research institutes such as universities. Yet more, properly ranking authors' publications may help in comparatively evaluating scientists as well.

At the present time, OLDLs lack effective and accurate publication rankings (Ratprasartporn *et al*., 2007). For instance, the ACM Digital Library returns unexplained rankings of publication search results that make this ranking not useful to users (ACM). Moreover, search output results of OLDLs tend to experience high level of the topic diffusion problem, which is defined as having large number of search results from multiple topics that are not of the current user's interest (Ratprasartporn and Ozsoyoglu, 2007; Voorhees and Buckley, 2002; Lin, 2005).

The topic diffusion problem occurs because keyword-based searches produce a large number of publications over a relatively large number of topics, thereby producing publication importance scores that are non-specific to topics (Ratprasartporn and Ozsoyoglu, 2007; Voorhees and Buckley, 2002; Lin, 2005).

Using social networks or bibliometrics, a number of publication score functions has been defined in literature (Brin and Page, 1998; Kleinberg, 1998; Bani-Ahmad *et al*., 2005a). In Bani-Ahmad *et al*. (2005b), the authors have comparatively evaluated several citation-based publication score functions, including, (1) PageRank proposed Brin and Page (1998), (2) Authorities scores

Corresponding Author: Sulieman Bani-Ahmad, Department of Information Technology, Al-Balqa Applied University, Salt Campus, Jordan

proposed in Kleinberg (1998), both adopted from the www research domain and (3) citation-count scores from the bibliometrics research domain (Chakrabarti, 2003). Bani-Ahmad *et al.* (2005a, b) observed that all those three score functions suffer from the separability problem, that is; none of these scoring functions assigns scores that distribute well over a given scale, e.g., [0, 1]. Instead, scores distributions of the three experimented publication score functions are found to be highly skewed (Bani-Ahmad *et al.*, 2005a, b) and decay very fast (Redner, 2004; Bani-Ahmad *et al.*, 2005a, b), resulting in a much less useful comparative publication assessment capability for users.

This lack of separability is caused by the rich gets richer phenomena identified in (Redner, 2004; Li and Chen, 2003). The rich gets richer phenomena involves observing a very small number of publications with relatively high numbers of in citations. Those highly-cited publications have even higher chances of receiving new citations. Further studies show that, yet, these citation-based scoring functions are also not very accurate, probably caused by topic diffusion in search outputs (Haveliwala, 2002).

The research-evolution model proposed by Aya *et al.* (2005) suggested that citation relationships between research publications produce multiple, small pyramid-like structures, where each pyramid represents a set of publications that are related to a highly specific research topic. A research pyramid is defined (Aya *et al.*, 2005) as a set of publications that represent a highly specific research topic and usually has a pyramid-like structure in terms of its internal citation graph (Aya *et al.*, 2005).

Publications within an individual research pyramid are: (1) motivated by earlier publications in the topic area (e.g., this paper is motivated in part by citations (Ratprasartporn *et al.*, 2007; Aya *et al.*, 2005), or (2) use techniques proposed in publications from other research pyramids (e.g., this study in part uses some of the techniques presented in citations (Brin and Page, 1998; Kleinberg, 1998)). Other reasons for citations may also be observed (Aya *et al.*, 2005).

PROBLEM STATEMENT

In this study, our goals are to (1) provide a solution to the OLDL search output ranking problem due to the topic diffusion problem, by grouping search outputs at the most-specific (detailed) topic level and without identifying the topics themselves, (2) eliminate the low separability problem of score functions and (3) improve the accuracy of three score functions, namely, PageRank, authorities and citation count score functions. Our approach uses the research pyramid (RP-) model to improve the separability and accuracy of publication scores and is based on normalizing publication scores within a limited scope, namely, within individual research pyramids. These improvements come from the fact that publications are now compared to their peers within their peer groups, namely, their own research pyramid publications that are on the same topic.

This study proposes and empirically evaluates two approaches to identify research pyramids. The first, called LB-IdentifyRP, uses link-based research pyramid identification, which captures research pyramids by identifying pyramid-like structures from the citation graph of the publication set. The second approach, called PB-IdentifyRP, uses proximity-based research pyramid identification, utilizes a graph-based proximity measure, namely SimRank (Jeh and Widom, 2002), to compute similarities between publications and then restructures the k-most-similar publications into a research pyramid.

This study's contributions are:

- Validate the research pyramid model of research evolution
- Propose and evaluate two algorithms to identify research pyramids
- Improve publication scores in terms of accuracy and separability via publications' research pyramids

As a testbed, we have utilized AnthP, a publication set of 14,891 publications from the ACM SIGMOD Anthology. Our experimental results show that:

- The complete publication citation graph (of AnthP) is highly clustered
- Each cluster of the complete publication set has a pyramid-like structure in terms of the citation graph of the cluster
- Each cluster represents a highly specific research topic. Note that the above three findings validate the research pyramid model proposed by Aya *et al.* (2005)
- Topic similarities decay over both the citation age and citation paths

We used the two topic similarity decay curves to guide the RP construction:

- Within RP citation graphs, the average number of in citations per paper varies, pointing to the importance of comparative publication scores within RPs
- Publication scores within RPs are accurate, due to our approach where each publication is compared only to its peer (research pyramid paper) group

CITATION-BASED PUBLICATION SCORES

Existing citation-based publication score functions are all based on the notion of prestige in social networks (Wasserman and Faust, 1994) and bibliometry (Chakrabarti, 2003). In this study, as publication score functions we use:

PageRank (Brin and Page, 1998) algorithm: PageRank score $P_{PageRank}$ of a publication P is recursively computed as the normalized sum of PageRank scores of documents citing P.

Authority score of the HITS (hyperlink induced topic search) algorithm (Kleinberg, 1998): Each document P gets two scores, namely hub and authority scores. Hub score of P is computed by summing up authority scores of the publications that P cites and the Authority score of P, denoted by P_{Auth}, is computed by summing the hub scores of publication citing P.

Normalized citation count score: For a particular paper P that receives C_P citations, the normalized citation count $P_{CitCount}$ is the ratio of C_P to the number C_{Pmax} of in-citations of the most cited paper in the publication set.

Figure 1a-c show that the three score functions, namely, $P_{PageRank}$, P_{Auth} and, $P_{CitCount}$ are highly skewed and do not separate scores well. Notice that the papers that are cited the most have the score of 1.0. Those papers are very few (less than one percent). The majority of scores cluster around the 0.1 value. This is because that, in the publication set used, 73.2% of the papers have received two citations or less. Thus, the majority of the publication set papers has received low scores that cluster around the 0.1 value.

Pan (2006), the author observed the skewness and inseparability of these functions independently in computer science and life sciences publications (70,000 documents in each) as well. And, it is shown (Redner, 2004; Li and Chen, 2003) that distributions of citation-based score functions are also highly skewed and decay very fast. We think that the cause is topic diffusion since scores are computed with respect to the full publication set. By using the research-pyramid model proposed by Aya *et al.* (2005), we normalize scores of publications within their own research pyramids, which allows for a fair comparative assessment of publications as publications are compared to their peers in their own research pyramids.

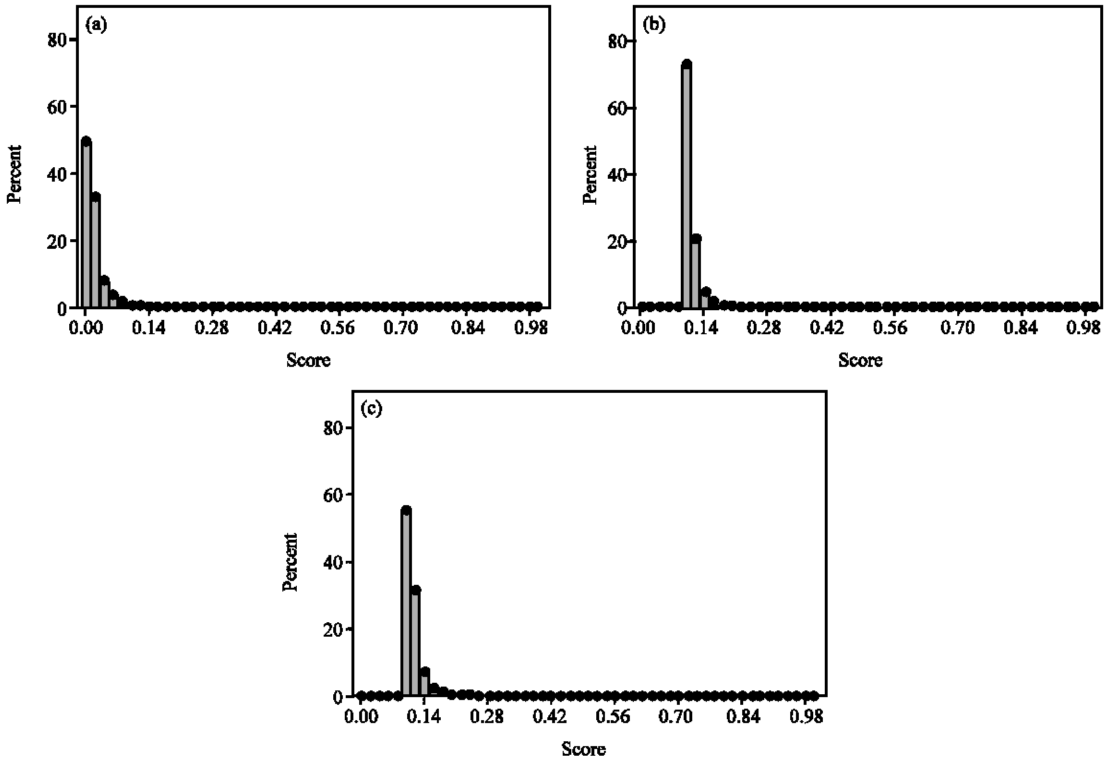

Fig. 1: Histograms of (a) CitCnt, (b) Auth and (c) PageRank. Score distribution of the three publication score functions. Publication set used consists of 15,000 publications from ACM Anthology all from the domain of data mining

PROPERTIES OF RESEARCH PYRAMID MODEL

We have observed three properties of research publications in three separate data sets, namely, ACM Anthology (AnthP; 15,000 publications) (Al-Hamdani, 2003) and computer sciences and life sciences publication sets (each with 70,000 publications) (Pan, 2006).

Property 1 (maximum citation age): In online digital libraries (OLDLs), most publications receive most of their in-citations within a fixed number of years after their publication dates. We refer to this value as the Maximum Citation Age and denote it by C_{AgeMax}.

We have observed (Bani-Ahmad *et al*., 2005a, b; Pan, 2006) that, in AnthP and Computer Sciences and Life Sciences OLDLs, most publications receive 90% of their in-citations in 10 years after they get published, i.e., $C_{ageMax} = 10$. Below in Property 4, we give a tighter bound for citation age within which topical similarity within an RP is maintained between citing and cited publications.

Figure 2 presents the citation age distributions in AnthP. We noticed that within ten years after their publishing date, publications receive 90% of their citations, that is, it is highly unlikely that publications receive new citations after 10 years of its publication dates. The Fig. 2 also shows that most papers receive top popularity and awareness levels after 5 years of its publications.

In rare cases, publications may cite works older than C_{AgeMax}. It is found (Ahmed *et al*., 2004; Case and Higgins, 2000) that a great proportion of these citations are for historical reasons, which we interpret as: old cited works (1) have coarse similarity to citing papers and (2) do not belong in the RP of the citing publication.

Property 2 (topic specificity over time): Scientific research publications quickly become very topic-specific over time, usually referable via a highly specific topic.

As shown in Fig. 3, an old research pyramid that covers a certain research topic leads to instantiations of new research topics and thus to creations of new RPs, that use techniques proposed in the publications of parent RP(s). Again, such old citations carry topical similarity between the citing and cited publication at a coarse granularity level. Possible citation exchanges between different RPs also occur and are of type uses, i.e., the citing paper uses techniques proposed by the cited paper.

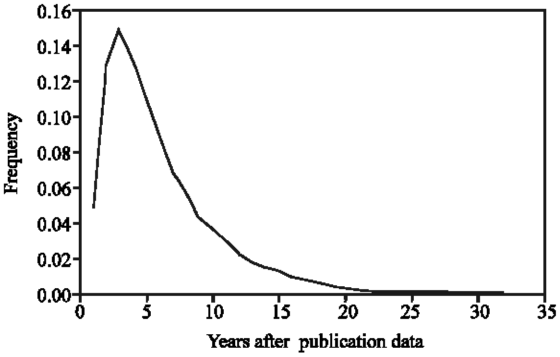

Fig. 2: Citation age distribution curve of AnthP

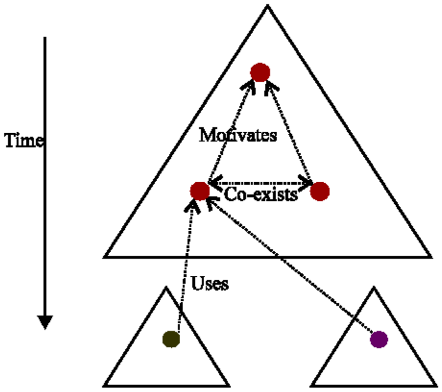

Fig. 3: The RP-based model

Example: Codd's paper E. F. Codd, A Relational Model of Data for Large Shared Data Banks, Commun. ACM 13(6): 377-387(1970) is about the topic relational model and cited around 580 times. A new and more specific topic of 2000's (i.e., citation to Codd's work is 30+ years old), say, rank-aware join algorithms, is coarsely related to the more general topic relational model in that, a publication P in the RP of rank-aware join algorithms and citing Codd's paper uses the techniques proposed in the RP of the relational model.

Property 3 (topic similarity decay over citation path): After very small citation path distances, topical similarity between papers decays significantly.

From Fig. 4, in AnthP, after a citation path of length 3, the topical similarity, as measured by SimRank, significantly decays. We refer to this value by $L_{Max\text{-}TopicDecay}$. This observation led us to build RPs of height at most 3 in the experimental results section.

Property 4 (topic similarity decay over citation age): After a certain citation age, topical similarity between the citing and the cited papers significantly decays.

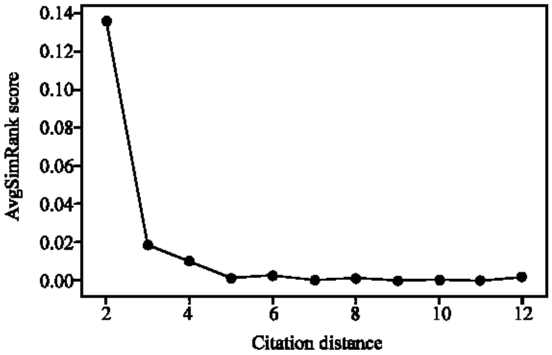

Fig. 4: SimRank score change with citation distance

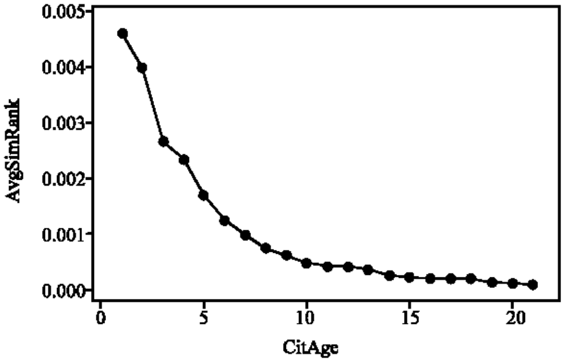

Fig. 5: SimRank score change with citation age

From Fig. 5, in the AnthP set, after a citation age of about 5 years, the topic similarity between the citing and cited papers decays significantly. We refer to this value by $C_{AgeMax\text{-}TopicDecay}$. This observation led us to build RPs in the experimental results section such that the maximum citation age within an RP is 5 years.

Next we present the two characteristics that identify a research pyramid RP.

RP-property 1 (high topic specificity): An RP, usually organizable into a pyramid, is a set of publications that represent a highly specific research topic.

We maintain high topic specificity of RPs by applying properties 3 and 4 and keeping the height of research pyramids low (property 3). Note that we make no attempts to identify the topic associated with an RP, as our approach does not need the topics explicitly. But, in interactive environments, providing topics to users is useful (Ratprasartporn and Ozsoyoglu, 2007).

RP-property 2 (research pyramid construction): RPs are arranged into pyramid structures either directly by using citation graphs (i.e., the link-based approach) (Aya *et al*., 2005) or indirectly using the publication times and close proximity of papers (i.e., the proximity-based approach).

RESEARCH PYRAMID IDENTIFICATION PROCEDURES

Based on the properties of publications and characteristics of RPs, next we propose two offline research pyramid identification procedures, namely, the Link-Based (LB) and the Proximity-Based (PB) RP identification procedures.

Both procedures start by choosing a candidate root node for an RP, called the cornerstone paper. The paper that is located at the root of a research pyramid receives more citations than others as other publications within the research pyramid are motivated by it and directly or indirectly cite it. Thus, our approach is to identify papers with high in-citations as cornerstone papers (i.e., the roots) of RPs to be constructed.

The link-based procedure locates research pyramids by identifying pyramid-like structures in the citation graph of the publication set. In summary, within an individual RP, publications are topically related (Aya *et al*., 2005) and motivated by each other (Fig. 3) (Aya *et al*., 2005) and we use the four properties of section 3 to identify citations within RPs-as summarized next.

In AnthP, the average number of citations to a paper (in-citations), denoted by C_I, is 2.066. Note that, in our experiments, we consider only the AnthP citations that are completely within AnthP; any citation from a paper within AnthP to a paper that is not in AnthP is removed. Using Property 3 and RP-Property 1, we limit RP heights to 3. Thus, the expected number of papers within a research pyramid RP_P with paper P as the root and with height 3 is $|RP_P| = 1 + C_I + C_I^2 + C_I^3 \sim 15$. Of course, the actual identified RP sizes (the number of papers in RP_P) vary. Some RPs may deal with active research topics and, in such cases, the number of in-citations of publications are noticeably higher than C_I, leading to noticeably higher RP sizes as well.

Figure 6a presents the link-based LB-IdentifyRP() procedure that utilizes citation-relationships between publications to identify the research-pyramid structures of the publication set at hand. The procedure LB-IdentifyRP() (1) selects a cornerstone paper P from the existing publication set (originally, say, AnthP) as an RP root, by simply picking the current most-cited publication (only citations that are $C_{AgeMax\text{-}TopicDecay}$ old according to property 4 above), (2) calls LB-FormRP() to locate the RP set RP_P of P and (3) eliminates RP_P from the current publication set CurrAnthP and repeats (a)-(c) again, until no more publications are left in CurrAnthP.

```
(a)

proc LB-IdentifyRP(AnthP, RP-Sets)
{RP-Sets := Ø;
 CurrAnthP := AnthP;
 while (CurrentAnthP = Ø)
   {Root:=ChooseRoot(CurrAnthP);
   $RP_{Root}$:=LB-FormRP(Root,$L_{Max\text{-}TopicDecay}$);
   RP-Sets:=RP-Sets U $RP_{Root}$;
   CurrAnthP:=CurrAnthP-$RP_{Root}$;
   } }

(b)

funct ChooseRoot(CurrAnthP)
  return $TopCited_{TopicDecay}$(CurrAnthP);

(c)

funct LB-FormRP(P, $L_{Max}$)
{Set $RP_P$:={P};   Queue Q;
 Q.Enqueue({P},0);
 while(Q is not empty)
   {<$P_i$,$l$>:=Q.Dequeue;
   if($l$<$L_{Max}$)then
   {CiterSet=Citers($P_i$, $l$, $C_{ageMax\text{-}TopicDecay}$);
   Q.Enqueue(CiterSet, ($l$+1));
   $RP_P$ = $RP_P$ +CiterSet;
   } } }
 Return $RP_P$}

(d)

Funct PB-FormRP(P, $L_{Max}$)
{Set $RP_P$={P}; Queue Q;
 Q.Enqueue(P,0);
 while(Q is not empty)
 {<$P_i$, $l$>:=Q.Dequeue;
 if($l$<$L_{Max}$) then
 {CiterSet($P_i$):=Citers($P_i$, $l$, $C_{AgeMax\text{-}TopicDecay}$)
 TopSimSet:=TopSim($P_i$,|CiterSet($P_i$)|, $C_{ageMax\text{-}TopicDecay}$);
 Q.Enqueue(TopSimSet, $l$+1);
 $RP_P$= $RP_P$+TopSimSet;
 } }
 Return $RP_P$}
```

Fig. 6: Functions of LB- and PB-IdentifyRP algorithms. (a) procedure LB-IdentifyRP, (b) function ChooseRoot, (c) function LB-FormRP() and (d) function PB-FormRP()

Note that our approach in this paper is to create distinct and nonoverlapping research pyramids. An alternative approach is to allow overlapping research pyramids as follows: Do not eliminate any papers from the original publication set (i.e., remove step (c) above); instead, simply color each selected publication and continue until all publications are colored, meaning that, when the algorithm ends, each paper belongs to at least one RP set and possibly more.

The two main functions of the link-based LB-IdentifyRP() procedure are ChooseRoot() and LB-FormRP(). ChooseRoot() (Fig. 6b) chooses publications that are cornerstone papers, or roots of research pyramids. The function LB-FormRP() (Fig. 6c) forms the RP_P of a root publication P by adding direct citers of P (i.e., level-1 citers) into RP_P and indirect citers of P at a level up to the L_{Max}; in experiments, we choose L_{Max} as 3, by following the property 3. The function citers (P, l, $C_{AgeMax\text{-}Topic\text{-}Decay}$) returns the set of publications that cite P at a level l (which is at most L_{Max}) where the citation age of the citing paper with respect to P is less than the maximum citation age $C_{AgeMax\text{-}Topic\text{-}Decay}$, (Properties 1 and 4). In more detail:

- Paper-id pid_P of root P along with its level 0 is inserted into RP_P and the queue Q, which holds paper-ids for future expansions and their distances to the root paper P
- Two-tuple <P_i, l> in Q is dequeued and expanded by locating direct or indirect citers of P_i so long as their levels with respect to P is at most $L_{Max\text{-}TopicDecay}$ (i.e., 3) and their citation age with respect to P (the root) is less than the maximum citation age $C_{ageMax\text{-}TopicDecay}$ (i.e., 5). All expanded publications and their level info with respect to P are inserted into the queue Q
- The above two steps are repeated until Q is empty; then RP_P is returned

The proximity-based PB-IdentifyRP() is similar to the link-based, except that the function call to LB-FormRP() is replaced by the function call PB-FormRP(). The function PB-FormRP() (Fig. 6d) of the proximity-based approach utilizes a graph-based proximity measure, namely SimRank (Jeh and Widom, 2002), to compute similarities between publications. It captures RP_P of the root publication by locating publications that are most similar to P and yet (a) are linked to P with a citation path length of at most $L_{Max\text{-}TopicDecay}$ and (b) have a citation time distance less than $C_{AgeMax\text{-}TopicDecay}$. SimRank iteratively computes similarity scores between nodes in a graph G following the rule that two nodes are similar if they are linked with similar nodes. In other words, the SimRank similarity between two nodes a and b, S(a, b), is iteratively computed using the formula (until the similarity scores converge):

$$S(a,b) = \left[C/|I(a)||I(b)|\right]^* \sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|} S(I_i(a),I_j(b)) \quad (1)$$

where, I (a) and I (b) are sources of in-links of a and b, respectively. C is the decay factor between 0 and 1. We choose C = 0.8 (Jeh and Widom, 2002). If |I (a)|or|I (b)| = 0 then S(a, b) = 0 by definition, in the case where a = b, S (a, b) = 1. The space complexity of the naive SimRank algorithm is O(N^2) where N is the graph size (the citation graph in publication domain). We prune as in Jeh and Widom (2002) by considering node pairs that are near each other in the range of radius r. We choose r = 6, which is twice the value of the expected research pyramid height as also explained in earlier.

PB-FormRP() receives as input the root P, the maximum level L_{Max} from root and utilizes the maximum citation age $C_{AgeMax\text{-}TopicDecay}$ (as 5) and returns the RP set RP_P of publication P following the same main steps of LB-FormRP() with one main difference: the way the two-tuple <P_i, l> dequeued from Q is expanded, as follows:

- Top |Citers($P_{i\ ,l,CAgeMax\text{-}TopicDecay}$)| similar papers, based on SimRank, to P_i are identified. The number of citers of P_i is used to capture the density of the RP being identified and thus to expand RP at P_i accordingly
- The identified similar papers are added to RP_P and also enqueued to Q for further expansion, this time with the level increased by 1. Similar to LB- FormRP() a maximum level of $L_{Max\text{-}TopicDecay}$ (which is 3) is employed

Advantage of PB-FormRP() over LB-FormRP() is that it successfully captures co-existing members of RP as well as those that are not reachable through any citation path from RP's root (as shown in Fig. 3 above). We give an example.

Example: Figure 7 shows two RPs; RP_1 and RP_2. RP_1 contains two co-existing roots A and B. Such a case occurs when two researchers work on the same problem simultaneously. At some point of our RP identification process, A will probably be recognized as a root of a new RP, say RP_3, as it has more in-citations than B. And, since B is not reachable through any path from A, LB-FormRP() will fail to identify B as a member of RP_3. PB-FormRP() will succeed to place both A and B into RP_3 in this case as B is very similar to A. A similar problem will be observed with paper C that is not reachable through any path from the root. Furthermore, LB-FormRP() may incorrectly identify F, that probably uses a technique proposed in A, as a member of RP_3 when F is really a member of RP_2 which co-exists with RP_3. PB-FormRP() successfully repels F from RP_3 as F is not similar to A or any of RP_3's members, based on SimRank.

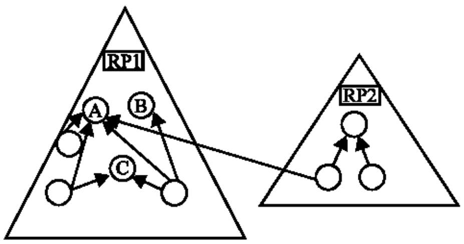

Fig. 7: Examples where PB-FormRP() is more successful than LB-FormRP()

We observe here that PB-FormRP() may capture pyramid-like structures, but not exactly pyramid structures. SimRank computes similarity between two papers P_1 and P_2 by averaging the similarity of the citers of both. However, note that similar papers to a member of an RP will be the other members of the same RP since members of an RP are usually cited by each other (as they are motivated by each other).

EMPIRICAL EVALUATIONS OF SCORE FUNCTIONS

AnthP, utilized as the OLDL testbed here, is a publication set of 14,891 publications from the ACM SIGMOD Anthology. After eliminating citations to papers outside AnthP, the average in-citations per AnthP paper is 2.066.

The three citation-based publication score functions (PageRank, Authorities and Citation count) have separability (high skew) and accuracy problems. We have observed that 99% of AnthP publications have scores below 0.1. This is because in-citations conform to the power law distribution, which describes the scale invariance found in many natural phenomena including publication citation graphs. As for low accuracy (probably due to topic diffusion problem (Haveliwala, 2002)), different research topics differ in their citation graph densities. Thus, a paper P's chances of receiving new citations depends on how dense the citation graph of the research topic of P is.

Observation: AnthP RPs (that represent specific research topics) have an almost normal distribution in the average in-citations received by members of an RP (Fig. 8a, b).

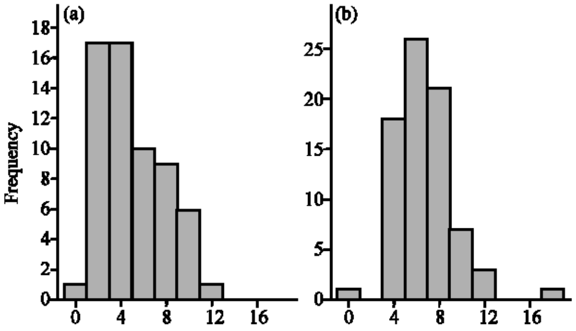

Fig. 8: Variance of citation-graph densities in different topics. (a) LB and (b) PB

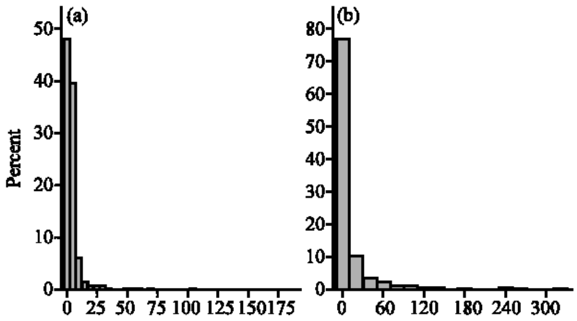

Fig. 9: Observed RP sizes by LP-IdentifyRP and PB-IdentifyRP. (a) LB sizes and (b) PB sizes

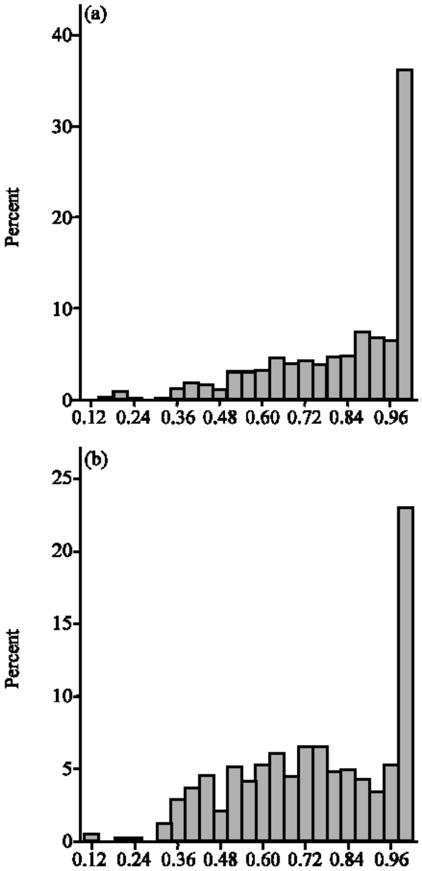

Fig. 10: Score distributions of PageRank normalized within Rps. (a) LB PageRank and (b) PB PageRank

For separability, first we verify the RP model on the AnthP set. We have experimentally observed that only 3.32% of SimRank scores are higher than 0.1, indicating that AnthP is highly clustered.

Observation: Average size of AnthP RP is 15.

Figure 9a and b show the distribution of the observed RP sizes within AnthP. Note that the PB approach identified larger RP sizes as it can identify co-existing RP roots and members that are not reachable through any citation path from the roots.

Figure 10a and b shown that $P_{\text{pgRank-LB}}$ and $P_{\text{PageRank-based}}$ publication scores distribute much better overthe interval [0, 1]. As for the citation-count-based scores, $P_{\text{CitCnt-LB}}$ and $P_{\text{CitCnt-PB}}$, Fig. 11a and b show that they also distribute much better over the interval [0, 1].

Observation: For RP-based scores, the observed skew values (Table 1) range between (-0.05) and (1.88) in the RP-based scores (zero skew indicates that the distribution is symmetric).

In comparison, the original scores showed highly skewed values that range between 8.12 and 13.04, which means that they are sharply left-skewed.

Observation: For RP-based scores, kurtosis values (that measure how sharply peaked a distribution is) range between (-0.26) to (2.65) (near zero Kurtosis values indicate normally peaked data).

In comparison, in the case of globally normalized scores, Kurtosis values range between (113.28) and (291.10). The enhancement of score distribution comes from the fact that publications are being compared to their peer groups, i.e., publications that belong to the same scope and thus have the same chances of receiving new citations.

The above observations on PageRank (P_{pgRank}, $P_{\text{pgRank-LB}}$, $P_{\text{PgRank-PB}}$) also apply to Authorities scores (P_{Auth}, $P_{\text{Auth-LB}}$, $P_{\text{Auth-PB}}$). Here, we report only PageRank-related results as we have observed that P_{Auth} and P_{PgRank} scores are highly correlated with a correlation coefficient of 0.98 and the correlation between P_{PgRank} and P_{citCnt} is 0.74 (Bani-Ahmad *et al.*, 2005a, b).

Observation: Each author in AnthP is identified with (i.e., author papers in) 2.19 and 2.16 LB and PB research pyramids (Fig. 12a, b).

This indicates that publications within an RP are highly related and, thus, the identified RPs are accurate.

We used expert knowledge in the data management field to manually evaluate the accuracy of searching via RPs. For this purpose, we built a prototype keyword-based search system that:

- Sends search keywords to Microsoft's Fulltext Search engine (MsFTS), that indexes the titles of AnthP publications. In turn, MsFTS generates a list of relevant publications (result set) along with rank values (which measures text-based relevancy between the publications and the search keywords)

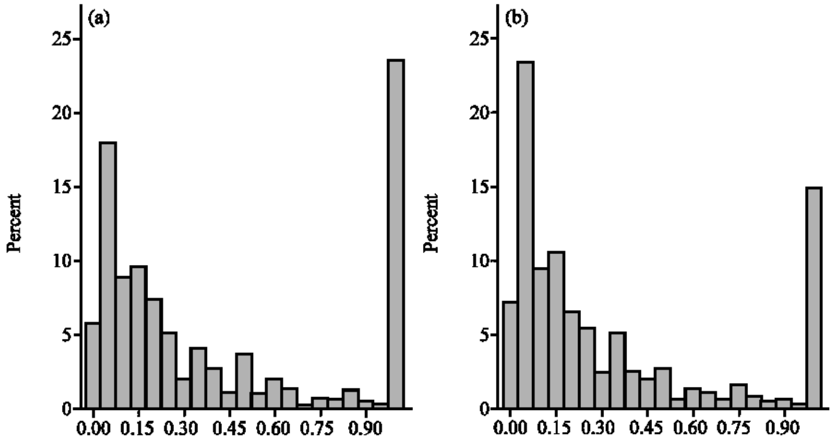

Fig. 11: Score distributions of Citation-count-based normalized within Rps. (a) LB CitCnt and (b) PB CitCnt

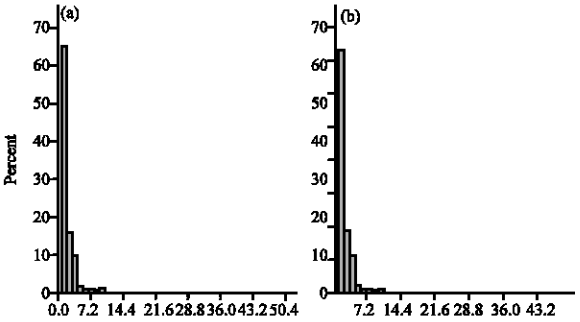

Fig. 12: Distribution of No. of RPs annotated with each author. (a) LB and (b) PB

Table 1: The means, Inter Quartile Ranges (IQR), skewness and kurtosis values of the publication score functions

Functions	Mean	IQR	Skewness	Kurtosis
CitCnt	0.02527	0.01845	8.12	113.28
Auth	0.11352	0.01134	13.04	291.10
PageRank	0.12091	0.01733	8.84	134.65
LBCitCnt	0.55698	0.88462	-0.05	-1.81
LBAuth	0.81266	0.37723	-1.02	-0.26
LBPageRank	0.77649	0.46181	-0.80	-0.84
PBCitCnt	0.20802	0.21910	1.88	2.65
PBAuth	0.62386	0.32036	-0.07	-0.58
PBPageRank	0.55653	0.31615	0.30	-0.60

- For each publication p in the result set, aggregates p's rank value returned by MsFTS with its scores, measured in two ways, namely globally-normalized PageRank and LBPageRank. We refer to this final score as the quality of paper p or Q(p). The quality scores are then used to sort the search output list so that high quality results appear at the top. The idea behind this aggregation is to push down publications that have high PageRank/LBPageRank scores and yet also have low rank values Rank(p), i.e., low relevancy to the search keywords. Q(p) is computed according to the following formula:

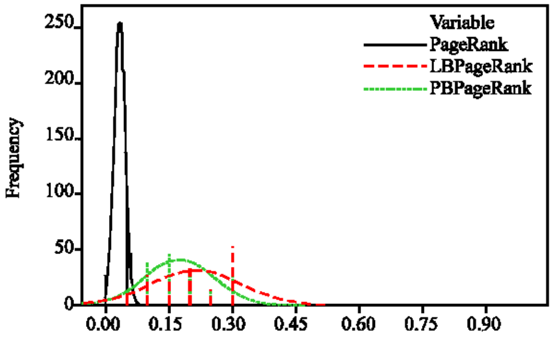

Fig. 13: Quality values distribution of the search results

$$Q(p) = Rank(p) * [LB]PageRank(p) \qquad (2)$$

We performed multiple searches and manually evaluated the accuracy of our system's outputs. We observed that LBPageRank-based quality scores resulted in 16-25% more accurate search outputs than the PageRank-based quality scores.

Observation: Quality scores Q(p) that are calculated computed using RP-based PageRank distribute much better than those computed using the globally-normalized PageRank (Fig. 13).

The accuracy search outputs was measured for the top-k publications in the result sets, where k is 10. In Table 2 and 3, we report our observations on one search experiment for the keywords complexity of join. Each publication in the Table 2 and 3 is evaluated by several domain experts who assigned a score between 0 and 10, where 0 score indicates no relevancy to the search terms and a score of 10 completely relevant. Integer numbers between those two extreme values indicates different levels of relevance.

Table 2: Sample results of the complexity of join query. Quality is computed using RP-based PageRank along with the average relevancy scores as assigned by experts

Quality	Publication title	Relevancy
1	Measuring the complexity of join enumeration in query optimization	9.0
0.487889	On the complexity of testing implications of functional and join dependencies	4.0
0.449827	Distributive join a new algorithm for joining relations	8.5
0.449827	The value of merge join and hash join in Sql server	2.0
0.449827	Multi table joins through bitmapped join indices	4.0
0.351713	Diag join an opportunistic join algorithm for 1 N relationships	8.0
0.339844	Utilizing page level join index for optimization in parallel join execution	4.5
0.315144	Evaluation of main memory join algorithms for joins with set comparison join predicates	8.0
0.287197	Join algorithm costs revisited	10.0
0.287197	Heuristic and randomized optimization for the join ordering problem	9.5
0.287197	Seeking the truth about ad hoc join costs	10.0

Table 3: Sample results of the complexity of join query. Quality is computed using the globally-normalized PageRank along with the average relevancy scores as assigned by experts

Quality	Publication title	Relevancy
0.148119	Measuring the complexity of join enumeration in query optimization	9.0
0.074381	Multiprocessor hash based join algorithms	5.5
0.067604	Efficient processing of spatial joins using R trees	7.0
0.062389	Join processing in database systems with large main memories	7.5
0.061929	On the complexity of testing implications of functional and join dependencies	4.0
0.060843	Join And semi join algorithms for a multiprocessor database machine	6.5
0.060467	Evaluation of main memory join algorithms for joins with set comparison join predicates	8.0
0.059288	Multi table joins through bitmapped join indices	4.0
0.055105	Partition based spatial merge join	2.0
0.053342	Multi step processing of spatial joins	2.0
0.05314	Tradeoffs in processing complex join queries via hashing in multiprocessor database machines	8.0

Observation: Quality scores of search results distribute better when computed based on RP-based publication score functions (Table 2, 3).

The average expert relevancy scores assigned to publications of Samples of Table 2 and those of Table 3 are 7.07 and 5.77 (Table 2). The above observation indicates that searching via RP-based publication scores is more accurate than globally normalized publication scores.

THE CASE EXPLORER PROJECT

The research conducted in this study is part of the CASE EXPLORER project (2003-2008). The project is resumed by Sulieman Bani-Ahmad at Al-Balqa Applied University in Jordan. The CASE EXPLORER is a score-guided searching and querying prototype portal for ACM SIGMOD Anthology, a digital library for the database systems research community, containing about 15,000 papers. CASE EXPLORER has a powerful user interface that allows users to pose score-guided ad hoc queries to search the Anthology, automatically computes the scores of query results from the scores of database objects (papers, authors, publication venues) and returns either the top-k results or results with high scores. CASE EXPLORER database is built by extracting metadata from the Anthology, storing it in a database, deriving multiple scores for papers, authors and publication venues. Propagating database scores to query outputs is achieved by a unique score propagation methodology. A rich set of queries are offered to users using a powerful and innovative user interface that allows users to add arbitrarily many conditions to their queries.

As an extension of the CASE EXPLORER project, Bani-Ahmad resumed the project in Jordan and is currently working on enhancing example-based search in literature digital libraries.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their very valuable comments that have substantially improved the paper in both substance and in presentation. We would also like to thank all the CASE EXPLORER project team at Case Western Reserve University. Namely, Nattakarn Ratprasartporn, Ali Cakmak who helped in enriching the paper by their discussion and comments. The project is financially supported by the National Science Foundation (NSF) grant number 8/2009-8/2012 and partially supported by the school of graduate studies at al-Balqa Applied University in Jordan.

CONCLUSIONS

In this study, we validated the Research-Pyramid model proposed by Aya *et al*. (2005). We proposed two algorithms to identify the research pyramids of a given collection. We also used the research pyramid model and the identified research pyramids to solve the separability and accuracy problems of publication score functions. We showed that normalizing publication scores within their research pyramids provides more accurate and separable (less skewed scores). Moreover, we showed that ranking search results by these scores promises to give higher accuracy compared to ranking by globally normalized publication scores due to reduction of topic diffusion effect.

REFERENCES

Ahmed, T., B. Johnson, C. Oppenheim and C. Peck, 2004. Highly cited old papers and the reasons why they continue to be cited. Part II., The 1953 Watson and Crick article on the structure of DNA. Scientometrics, 61: 147-156.

Al-Hamdani, A., 2003. Querying web resources with metadata in a database. Ph.D. Thesis, EECS Department, Case Western Reserve University Cleveland.

Aya, S., C Lagoze and T. Joachims, 2005. Citation classification and its applications. Proceedings of the 2005 International Conference on Knowledge Management, Oct. 27-28, North Carolina, USA., pp: 287-298.

Bani-Ahmad, S., A. Cakmak and G. Ozsoyoglu, 2005a. Evaluating publication similarity measures. IEEE Data Eng. Bull., 28: 21-28.

Bani-Ahmad, S., A. Cakmak, A. Al-Hamdani and Gultekin Ozsoyoglu, 2005b. Evaluating score and publication similarity functions in digital libraries. Proceedings of the International Conference of Asian Digital Libraries, Dec. 12-15, Bangkok, Thailand, pp: 483-485.

Brin, S. and L. Page, 1998. The anatomy of a large-scale hypertextual web search engine. Comput. Networks ISDN Syst., 30: 107-117.

Case, D.O. and G.M. Higgins, 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. J. Am. Soc. Inform. Sci., 51: 635-645.

Chakrabarti, S., 2003. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, California.

Haveliwala, T.H., 2002. Topic-sensitive PageRank. Proceedings of the 11th International World Wide Web Conference. May 7-11, Honolulu, Hawaii, USA., pp: 1-10.

Jeh, G. and J. Widom, 2002. SimRank: A measure of structural-context similarity. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, Edmonton, Alberta, Canada, ACM., pp: 538-543.

Kleinberg, J.M., 1998. Authoritative sources in hyperlinked environments. J. ACM, 46: 604-632.

Li, X. and G. Chen, 2003. A local-world evolving network model. Phys. A Stat. Mech. Appl., 328: 274-286.

Lin, W.H., 2005. A revisiting the effect of topic set size on retrieval error. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, Salvador, Brazil, pp: 637-638.

Pan, F., 2006. Comparative evaluation of publication characteristics in computer science and life sciences. M.Sc. Thesis, EECS, Case Western Reserve University.

Ratprasartporn, N. and G. Ozsoyoglu, 2007. Finding Related Papers in Literature Digital Libraries. In: Research and Advanced Technology for Digital Libraries, Kovacs, A., N. Fuhr and C. Meghini (Eds.). LNCS., 4675, Springer-Verlag, Berlin, Heidelberg, ISBN: 978-3-540-74850-2, pp: 271-284.

Ratprasartporn, N., S. Bani-Ahmad, A. Cakmak, J. Po and G. Ozsoyoglu, 2007. Evaluating utility of different ranking functions in context-based environment. Proceedings of the 23rd IEEE International Conference on Data Engineering Workshop, April 17-20, Istanbul, Turkey, pp: 261-268.

Redner, S., 2004. Citation statistics from more than a century of physical review. Physics 0407137. http://arxiv.org/abs/physics/0407137v2.

Voorhees, E.M. and C. Buckley, 2002. The effect of topic set size on retrieval experiment error. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 11-15, ACM Press, USA., pp: 316-323.

Wasserman, S. and K. Faust, 1994. Social Network Analysis: Methods and Applications. Vol. 8, Cambridge University Press, Cambridge, ISBN-10: 0521387078.