5

Sorting Search Results of Literature Digital Libraries: Recent Developments and Future Research Directions

Sulieman Bani-Ahmad
AlBalqa Applied University
Jordan

1. Introduction

An OLDL (Online Literature Digital Library) is a library in which collections, i.e., publications from one or more domains of study, are stored in *digital formats* (as opposed to print, microform, or other media) and accessible by users through the Internet. Examples of well-known OLDLs are IEEE Xplore (IEEE Xplore, 2008), ACM Portal (ACM Digital Library, 2008), CiteSeer (CiteSeer, 2008), Google Scholar (Google Scholar, 2008), and PubMed (PubMed, 2008).

Digital libraries are rapidly growing in popularity. For instance, ScienceDirect (ScienceDirect, 2008), the world's leading scientific, technical and medical information resource celebrated its billionth article download in November'06 since launched in 1999. Besides usage, digital libraries are also rapidly growing in terms of *size* and *diversity of topics*. For instance, (i) in Computer Science, ACM Digital Library (ACM Digital Library, 2008) has close to *one million* full-text publications collected over 50 years, to search and download; (ii) in Electrical Engineering and Computer Science, IEEE Xplore (IEEE Xplore, 2008), another OLDL, provides users with on-line access to more than 1,700 selected conferences proceedings.

These high growth rates introduced several challenges facing the information access capability of OLDLs. Next we list few challenges that probably guides future research related to LDLs.

Challenge 1: Large Sizes and Topic Diversity of Search Output Results. Search outputs of OLDLs tend to suffer from the "topic diffusion" problem, where commonly, keyword-based searches produce a large number of publications over a large number of topics, where not all topics are of interest to the user. One way to solve this problem is to assign scores to search results (i.e., publications). Assigning scores to publications helps OLDLs to present the most important relevant publications to the user first, Citation-based publication score measures (e.g., citation count) are commonly used for ranking publications. At the present time, OLDLs lack effective and accurate publication ranking.

Challenge 2: Lack of Effective Scoring Functions for Publications. At the present time, OLDLs lack effective and accurate publication rankings (Ratprasartporn et al., 2007). Providing accurate publication scores can help users in reducing the time spent in searching OLDLs, and thus enhances the scalability of OLDL usage as users can quickly identify important relevant publications to their topic of interest.

Challenge 3: Lack of Effective Scoring Functions for Search Outputs. In the field of literature digital libraries, citation analysis is employed to order digital library search outputs (e.g., Google Scholar). Examples of citation-based measures are citation-count (Bani-Ahmad & Ozsoyoglu, 2007) and PageRank (Brin & Page, 1998). However, as noticed by (Cho et al., 2005), citation-based measures compute popularity of publications based on the “current” state of a citation graph that continuously changes and evolves. Thus PageRank is effective in capturing the popularity of publications based on the current citation-graph in-hand. In section 4, we show that PageRank may assign inaccurate popularity scores *for both old and recent publications*. And thus PageRank cannot be used to rank OLDL search outputs. We therefore need effective techniques to order search results based on their importance and relevance to users’ interests.

This chapter is organized as follows. After the introduction in section 1, we present and evaluate a set of citation-based score functions for publications. We show that they have problems in both accuracy and separability. To solve these problems, section 3 introduces the *Research-Pyramid Model*, a new model for the evolution of research and citation behavior. For that, we present two algorithms from literature for identifying research pyramid structures in publication citation graphs. We show that this model can help in computing accurate and non-skewed publication scores. In section 4 we propose the notion of *publication’s popularity*. We also present how the temporal popularity of publications, as computed by the PageRank algorithm for instance, varies over time. For that we validate the *publication popularity growth and decay model*. And finally in section 5 we present a number of future research directions related to the topic of this chapter.

The observations preselected in this chapter are based on real experiment conducted on a literature digital collection of around 15,000 publications that we refer to as the AnthP. AnthP. These publications are from the ACM SIGMOD Anthology (ACM SIGMOD Anthology, 2003). For each paper in the AnthP, DBLP bibliography (DBLP, 2003) is used to extract the titles, authors, publication venue (conference or journal), and publication year info. Information extracted about each paper is the paper’s publication venue, the publication year, authors, and citations. The AnthP dataset includes: (a) 106 conferences, journals, and books, (b) 14,891 papers, and (c) 13,208 authors.

2. Evaluating publication scoring functions in digital libraries

This section deals with the issues of defining score functions for publications in digital libraries, and evaluating how good they are. Presently, digital libraries do not assign scores to publications, even though they are potentially useful for (a) providing comparative assessment, or "importance", of papers, and (b) ranking papers returned in search outputs. Using social networks or bibliometrics, one can define a number of publication score functions.

Existing citation-based publication score functions are all based on the notion of prestige in social networks (Wasserman & Faust, 1994) and bibliometry (Chakrabarti, 2003). The well-known PageRank (Brin & Page, 1998) algorithm determines the importance of a publication by the number *and* importances of publications with links to it (i.e. citing papers). The Hyperlink Induced Topic Search (HITS) algorithm (Kleinberg, 1998) is similar to the PageRank algorithm in that HITS involves computing two scores for each publication; hub and authority scores. Authorities represent high-prestige publications, whereas hubs are publications that have links to authorities. Other citation-based score functions can be derived as follows. (a) Use normalized citation count (i.e., how many times a paper is cited by other papers) as the basis for a score function. (b) Revise the score of a paper using the score of its publication venue (conference or journal). (c) Add weights to citations, e.g., citations by an "important" author's work are more significant. (d) Revise the score of a paper using temporal distributions of citations; e.g., citations in the last 10 years are more significant than earlier citations. (e) Revise a paper score using the score of its citation venue; that is, capture the notion of a hub or an authority, e.g., survey journal represents a hub, whereas a research paper represents an authority. (f) Revise a paper score by the score of its author. One can also combine the score functions above. In the next two subsections we present, in more details, and evaluate these citation-based score functions of publications.

2.1 Citation-based publication score functions

In this section we present and evaluate citation-based score functions for publications.

A. PageRank

Importances of papers that cite a particular paper determines its importance. PageRank (Brin & Page, 1998) and HITS (Kleinberg, 1998) were designed based on this assumption. PageRank scores is computed recursively using the formula

$$P_{i+1} = (1-d)M^T P_i + E$$

Where P_i and P_{i+1} are the current and next iteration PageRank vectors respectively. M is a matrix derived from the citation matrix *C* by normalizing all row-sums in *C* to 1. *C*, in turn, is the adjacency matrix of the graph *G* formed as follows; the papers represent the graph nodes, and the citation relationships between these papers represent the edges. *C* is of size *N*x*N*, where *N* is the total number of papers in the system. Finally, *d* and (1-*d*) are the future citation probability. Given that an author *A* who is writing a new paper and already cited paper *u* which in turn cites paper *v*, and let *w* be a paper in *AnthP* selected randomly. The parameter *d* represents the probability that *A* will cite *w*, and (1-*d*) is the probability that *A* will cite *v*.

To guarantee the algorithm convergence, it is assumed to have a hidden link between each pair of the graph nodes. This link is represented by the user-defined parameter *E*. A variation of *E* is simply E_1=*d*. Another variation of *E* that is used in (Brin & Page, 1998) is

$$E_2 = d / N \left[1_N \right] P_i .$$

Where 1_N is a vector of N ones.

B. Hubs and Authorities

Authority score of paper P is computed by summing up the hub scores of the papers citing P. Hub score of P is computed by summing up the authority scores of the papers that P cites. Computation is recursive until results converge after a number of iterations. One difference between HITS and PageRank is that the first one works on papers in the result set of a query, while the latter considers all the papers independent of the query [Cakmak, 2003].

C. Citation Count

A paper, normally, does not cite another paper unless the cited paper is relevant. And, large number of citations to a paper gives an indication that the paper is important. Based on this fact, one can use citation count as a measure for paper importance. For a given paper P, let *CitationCount*(P) be the number of times paper P is cited by other papers. Using the number of citations, paper P is as important as those papers that have the same number of citations and more important than those papers that have fewer citations. We will refer to this paper ranking measure as $P_{Citation\_Count}$.

2.2 Evaluating publication score functions

Figure 1 shows the three score functions, namely, PageRank (P_{PgRank}), Authorities scores of HITS (P_{Auth}) and, the Citation-count (P_{CitCnt}). As it is clear from the figure the three functions are highly skewed, and do not separate scores well over the interval [0, 1]. This figure is based on the AnthP digital collection from the field of data management[1]. More details about AnthP can be found in (Bani-Ahmad & Ozsoyoglu, 2007). In (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 (Render, 2004; Li & Chen, 2003) that distributions of citation-based score functions are also highly skewed and decay very fast. Studies show that the cause is topic diffusion since scores are computed with respect to the full publication set.

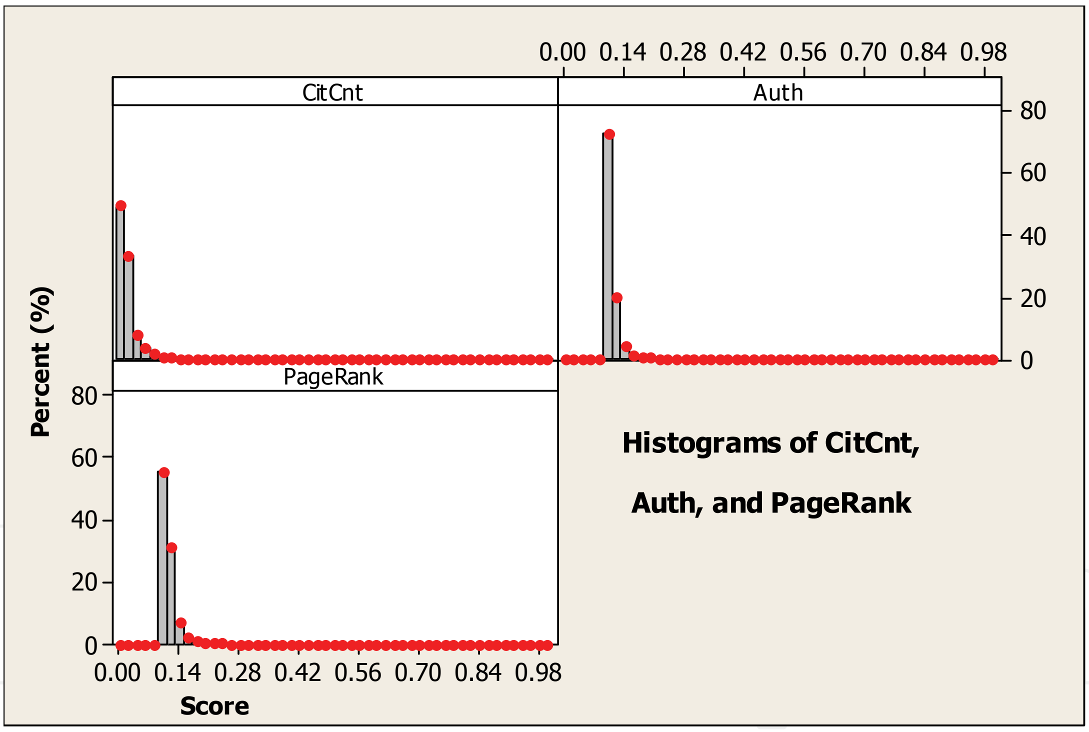


Fig. 1. Skewness of Score distribution of the three main citation-based publication score functions.

In (Bani-Ahmad* et al., 2005), the authors compared and evaluated several publication score functions, including *PageRank* (Brin & Page, 1998) and *Authorities scores* (Kleinberg, 1998), both adopted from the www search domain, and *citation-count scores* from the bibliometrics domain (Chakrabarti, 2003). The authors observed the *separability* problem with all of these

[1] This experimental dataset includes: (a) 106 conferences, journals, and books, (b) 14,891 papers, and (c) 13,208 authors. These papers are obtained from ACM SIGMOD Anthology.

functions which is that none of these scoring functions assigns scores that distribute well over a given scale, e.g., [0, 1]. Instead, distributions of existing publication score functions are highly skewed, and decay very fast (Render, 2004), resulting in a much less useful comparative publication assessment capability for users. This lack of separability is caused by the "rich gets richer" phenomena (Render, 2004; Li & Chen, 2003), i.e., a very small number of publications with relatively high numbers of in-citations have even higher chances of receiving new citations. Yet, these scoring functions are still not very accurate, probably caused by topic diffusion in search outputs (Haveliwala, 2002).

In the following section, and by using the research-pyramid model proposed in (Aya et al., 2005), the authors in (Bani-Ahmad & Ozsoyoglu, 2007) normalize scores of publications within their (the publications) own research pyramids, which allows for a fair comparative assessment of publications as publications are compared to their peers in their own research pyramids.

3. Improved publication scores via research pyramids

Providing accurate publication scores for search results and ranking publications returned as search results accurately can help users in reducing the time spent in searching OLDLs. And, better publication rankings are also useful for comparative assessments of publication venues and scientists as well.

At the present time, OLDLs lack effective and accurate publication rankings (Ratprasartporn et al., 2007). For instance, ACM Digital Library returns rankings of publication search results that are unexplained and not useful to users (ACM Digital Library, 2008). Moreover, search outputs of OLDLs tend to suffer from the "topic diffusion" problem, where commonly, keyword-based searches produce a large number of publications over a large number of topics, thereby producing scores that are nonspecific to topics.

The research evolution model proposed in (Aya et al., 2005) suggests that citation relationships between research publications produce multiple, small *pyramid-like* structures, where each pyramid represents publications 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 citation graph (Aya et al., 2005). Publications within an individual research pyramid are (i) *motivated by* earlier publications in the topic area (e.g., our paper (Bani-Ahmad & Ozsoyoglu, 2007) is motivated in part by citations (Ratprasartporn et al., 2007), and (Aya et al., 2005)), or (ii) *use techniques* proposed in publications from other research pyramids (e.g., our paper (Bani-Ahmad & Ozsoyoglu, 2007) in part uses some of the techniques presented in citations (Brin & Page, 1998) and (Kleinberg, 1998)). Other "reasons" for citations may also be observed (Aya et al., 2005).

In this section, our goals are to (a) 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, (b) eliminate the low separability problem of score functions, and (c) improve the accuracy of three score functions, namely, PageRank, Authorities and Citation Count score functions. The research pyramid (RP-) model is used 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.

In (Bani-Ahmad & Ozsoyoglu, 2007), two approaches to identify research pyramids are presented and evaualted. 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 & Widom, 2002), to compute similarities between publications, and then restructures the k-most-similar publications into a research pyramid.

3.1 Properties of research pyramid model

In (Bani-Ahmad & Ozsoyoglu, 2007), the authors have observed three properties of research publications in three separate data sets, namely, ACM Anthology which is a collection of 15,000 publications (we refer to this set by the AnthP set in future), and computer sciences and life sciences publication sets, each with 70,000 publications (we refer to these sets by the CSSet and LSSet in future) (Pan, 2006). These properties are utilized in the identification of research pyramids.

Property 1 (*Maximum Citation Age)*. In 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}.

It has been observed in (Bani-Ahmad et al., 2005; Pan, 2006) that, in the *AnthP*, CSSet and the LSSet datasets, most publications receive 90% of their in-citations in 10 years, i.e., $C_{AgeMax}=10$. Figure 2 presents the citation age distributions in AnthP. 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.

In rare cases, publications may cite works older than C_{AgeMax}. It is found (Ahmed et al., 2002) that a great proportion of these citations are for historical reasons, which we interpret as: old cited works (a) have coarse similarity to citing papers, and (b) do not belong in the RP of the citing publication.

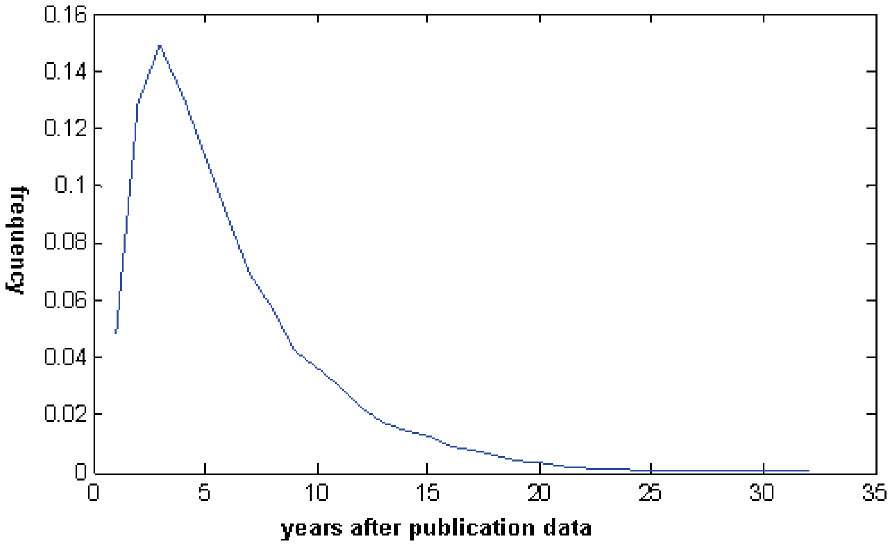


Fig. 2. Citation age distribution curve of AnthP

Property 2 (*Topic Specificity Over Time*). Scientific research publications quickly become very topic-specific over time, usually referable via a highly specific topic.

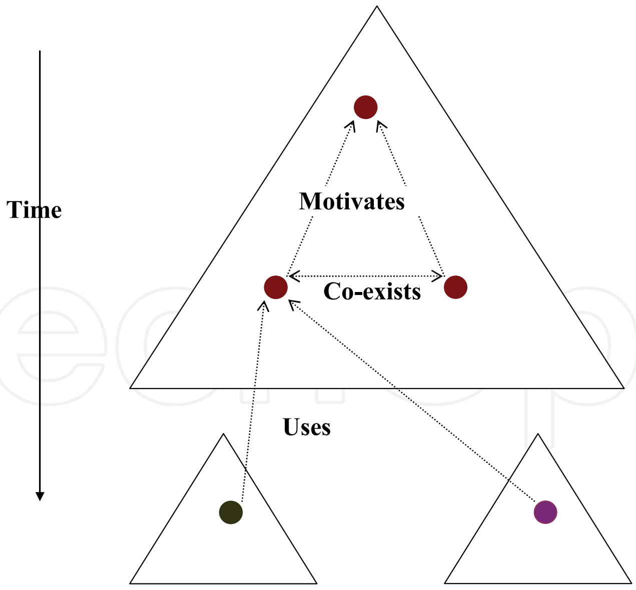


Fig. 3. The RP-Based Model

As illustrated in Figure 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.

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 Figure 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 the authors in (Bani-Ahmad & Ozsoyoglu, 2007) 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.

From Figure 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 the authors in (Bani-Ahmad & Ozsoyoglu, 2007) to build RPs in the experimental results section such that the maximum citation age within an RP is 5 years.*

The two characteristics that identify a *research pyramid RP* are.

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

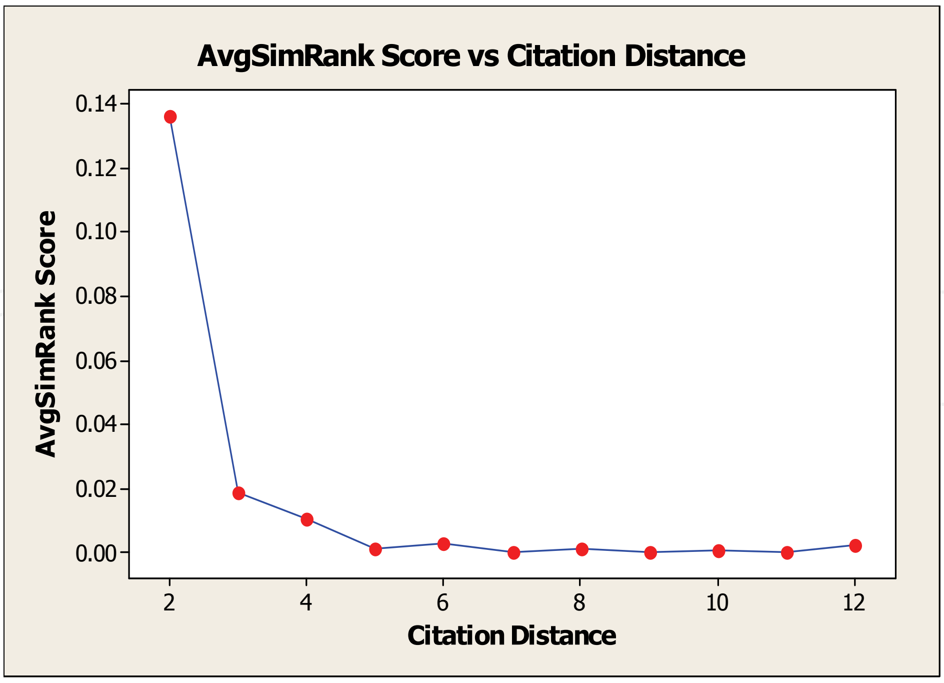


Fig. 4. SimRank score change with citation distance

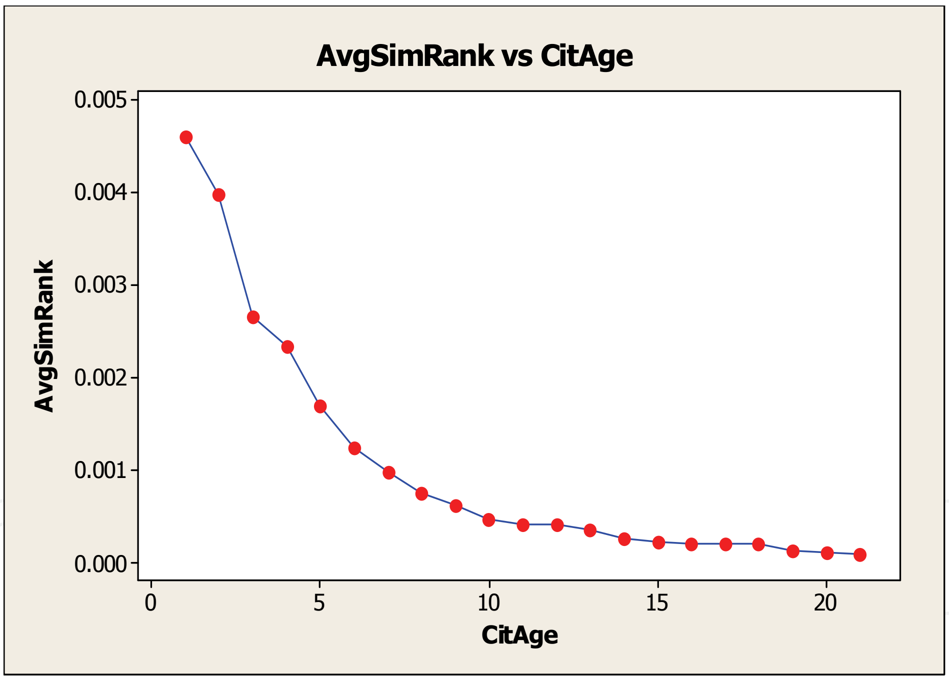


Fig. 5. SimRank score change with citation age

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 & 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).

3.2 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 (see figure 3) (Aya et al., 2005), and we use the four properties 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 $|RP_P| = 1 + C_I + C_I^2 + C_I^3 \approx 15$. 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 6-(a) presents the link-based *LB-IdentifyRP()* procedure. The proximity-based *PB-IdentifyRP()* is similar, except that the function call to *LB-FormRP()* is replaced by the function call *PB-FormRP().* The procedure *LB-IdentifyRP()* (a) 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), (b) calls *LB-FormRP()* to locate the RP set RP_P of P, and (c) eliminates RP_P from the current publication set *CurrAnthP*, and repeats (a)-(c) again, until no more publications are left in *CurrAnthP*.

Note that our approach in this chapter is to create distinct and nonoverlapping research pyramids. An alternative approach, not reported here due to space limitations, is to allow *overlapping research pyramids* as follows: Do not to 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()* (See Figure 6.b) chooses publications that are cornerstone papers, or roots of research pyramids. The function *LB-FormRP()* (Figure 6.c) 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,

1. 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.

2. *T*wo-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*.
3. The above two steps are repeated until *Q* is empty; then RP_P is returned.

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

(a) Procedure LB-IdentifyRP

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

(b) Function ChooseRoot

```
funct LB-FormRP(P, $L_{Max}$)
{Set $RP_p$:={P};   Queue Q;
 Q.Enqueue({P},0);
 while(Q is not empty)
 {<$P_i$,$\ell$>:=Q.Dequeue;
   if($\ell$<$L_{Max}$)then
   {CiterSet=Citers($P_i$,$\ell$, $C_{AgeMax-TopicDecay}$);

   Q.Enqueue(CiterSet,($\ell$+1));
    $RP_p$ = $RP_p$ +CiterSet;
    } } }
 Return $RP_p$}
```

(c) Function LB-FormRP()

```
Funct PB-FormRP(P, $L_{Max}$)
{Set $RP_p$={P}; Queue Q;
 Q.Enqueue(P,0);
 while(Q is not empty)
 {<$P_i$,$\ell$>:=Q.Dequeue;
 if($\ell$<$L_{Max}$) then
 {CiterSet($P_i$):=Citers($P_i$,$\ell$, $C_{AgeMax-TopicDecay}$)

   TopSimSet:=TopSim($P_i$,|CiterSet($P_i$)|, $C_{AgeMax-TopicDecay}$);
   Q.Enqueue(TopSimSet,$\ell$+1);
   $RP_p$= $RP_p$+TopSimSet;
   } }
   Return $RP_p$}
```

(d) Function PB-FormRP()

Fig. 6. Functions of LB- and PB-IdentifyRP algorithms

The function *PB-FormRP()* (figure 6.d) of the proximity-based approach utilizes a graph-based proximity measure, namely *SimRank* (Jeh & 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):

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

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 & 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 & 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 Section 3.5.

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 $|Citers(P_i, \ell, C_{AgeMax\text{-}TopicDecay})|$ 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 illustrated in Figure 3.7 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*.

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).

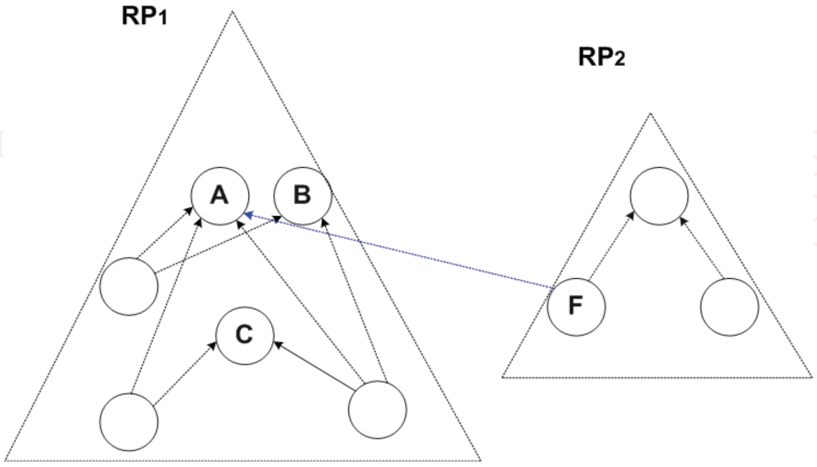


Fig. 7. Examples where *PB-FormRP()* is more successful than *LB-FormRP().*

3.3 Improved publication scores via the RP-Model

In (Bani-Ahmad & Ozsoyoglu, 2007), the authors have applied the two RP-identification algorithms on the AnthP set. After that, they normalized publication scores within the research pyramids identified. The authors observed that, 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 mean that they are sharply left-skewed. They also observed that, 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.

	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

Table 1. The Means, InterQuartile Ranges (IQR), Skewness, and Kurtosis values of the Publication Score Functions applied on the AnthP set.

The above observations on *PageRank* (P_{PgRank} , $P_{PgRank\text{-}LB}$, $P_{PgRank\text{-}PB}$) also apply to *Authorities* scores (P_{Auth} , $P_{Auth\text{-}LB}$, $P_{Auth\text{-}PB}$). 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., 2005).

The authors in (Bani-Ahmad & Ozsoyoglu, 2007) have also performed multiple searches and manually evaluated the accuracy ranking publication via the RP-Model. They observed that research-pyramid-based scores resulted in 16% - 25% more accurate search outputs than the PageRank-based quality scores. Accuracy was measured for the top-k publications in the result sets, where k is 10.

3.4 Section summary and conclusions

In this section, The Research-Pyramid model proposed in (Aya et al., 2005) is used to solve the separability and accuracy problems of publication score functions. We showed that (i) normalizing publication scores within their research pyramids provides more accurate and less skewed scores, moreover (ii) 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.

However, as noticed by (Cho et al., 2005), citation-based measures compute popularity of publications based on the "current" state of a citation graph that continuously changes and evolves. Thus PageRank is effective in capturing the popularity of publications based on the current citation-graph in-hand. In the following section, we show that PageRank may assign inaccurate popularity scores *for both old and recent publications*. And thus PageRank cannot be used to rank OLDL search outputs. We therefore need effective techniques to order search results based on their importance and relevance to users' interests.

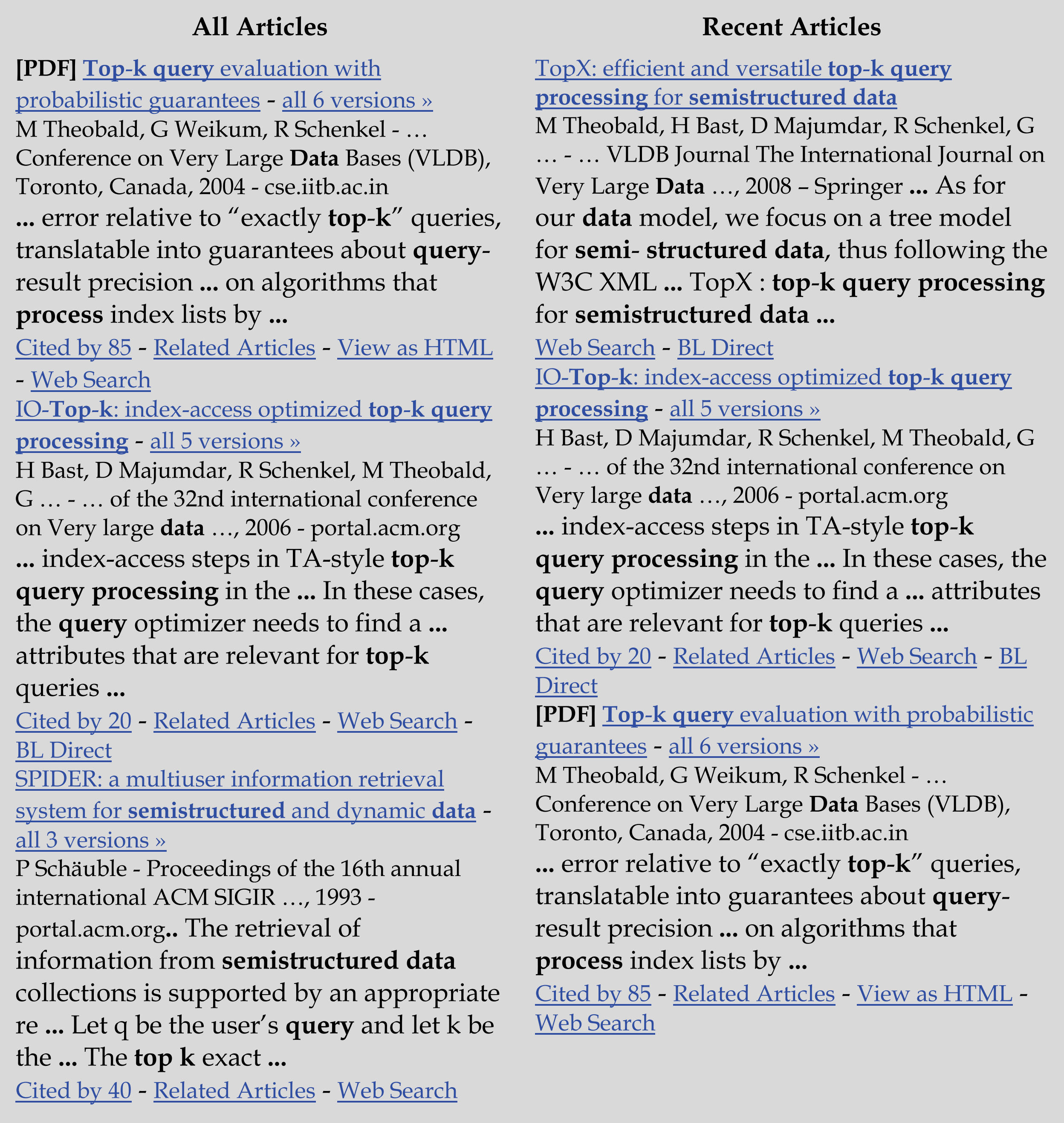
4. On popularity quality: growth and decay phases of publication popularities

4.1 Introduction

In the field of literature digital libraries, citation analysis is employed to evaluate the impact of publications and scientific collections (e.g., journals and conferences). It is also employed to order digital library search outputs (e.g., Google Scholar). Examples of citation-based measures are citation-count (Bani-Ahmad & Ozsoyoglu, 2007) and PageRank (Brin & Page, 1998). However, as noticed by (Cho et al., 2005), citation-based measures compute popularity of publications based on the "current" state of a citation graph that continuously changes and evolves. Next we present two scenarios where usage of such popularity scores becomes problematic.

Example 1 (*Scores for recent publications*; Google Scholar (Google Scholar, 2008)). Figure 8 shows a sample search output from Google Scholar, a digital library search tool by Google (Google Scholar, 2008), for keywords "top-k query processing for semistructured data". On the left-side of figure 8, relevant documents are ordered based on text-based relevancy to query terms **and** the citation-based popularity of the document. On the right-side, documents are ordered based on their publication date. The most relevant document to our query, the one entitled by "TopX: efficient and versatile top-k query processing for semistructured data", is published in 2008, and appears at the top of the right-side search output list (where popularity didn't affect the order of the output list). In comparison, this document is pushed down and appeared on page 5 of the left-side search output list of

Google Scholar. Given that users usually check only a few pages of a returned list of documents (Bani-Ahmad & Ozsoyoglu, 2007), this publication may not even have a chance to develop popularity unless, in time, awareness of readers increases, i.e., it becomes known to users.



All Articles

[PDF] **Top-k query** evaluation with probabilistic guarantees - all 6 versions »
M Theobald, G Weikum, R Schenkel - … Conference on Very Large **Data** Bases (VLDB), Toronto, Canada, 2004 - cse.iitb.ac.in
... error relative to "exactly **top-k**" queries, translatable into guarantees about **query**-result precision **...** on algorithms that **process** index lists by **...**
Cited by 85 - Related Articles - View as HTML - Web Search

IO-**Top-k**: index-access optimized **top-k query processing** - all 5 versions »
H Bast, D Majumdar, R Schenkel, M Theobald, G … - … of the 32nd international conference on Very large **data** …, 2006 - portal.acm.org
... index-access steps in TA-style **top-k query processing** in the **...** In these cases, the **query** optimizer needs to find a **...** attributes that are relevant for **top-k** queries **...**
Cited by 20 - Related Articles - Web Search - BL Direct

SPIDER: a multiuser information retrieval system for **semistructured** and dynamic **data** - all 3 versions »
P Schäuble - Proceedings of the 16th annual international ACM SIGIR …, 1993 - portal.acm.org.. The retrieval of information from **semistructured data** collections is supported by an appropriate re **...** Let q be the user's **query** and let k be the **...** The **top k** exact **...**
Cited by 40 - Related Articles - Web Search

Recent Articles

TopX: efficient and versatile **top-k query processing** for **semistructured data**
M Theobald, H Bast, D Majumdar, R Schenkel, G … - … VLDB Journal The International Journal on Very Large **Data** …, 2008 – Springer **...** As for our **data** model, we focus on a tree model for **semi- structured data**, thus following the W3C XML **...** TopX : **top-k query processing** for **semistructured data ...**
Web Search - BL Direct

IO-**Top-k**: index-access optimized **top-k query processing** - all 5 versions »
H Bast, D Majumdar, R Schenkel, M Theobald, G … - … of the 32nd international conference on Very large **data** …, 2006 - portal.acm.org
... index-access steps in TA-style **top-k query processing** in the **...** In these cases, the **query** optimizer needs to find a **...** attributes that are relevant for **top-k** queries **...**
Cited by 20 - Related Articles - Web Search - BL Direct

[PDF] **Top-k query** evaluation with probabilistic guarantees - all 6 versions »
M Theobald, G Weikum, R Schenkel - … Conference on Very Large **Data** Bases (VLDB), Toronto, Canada, 2004 - cse.iitb.ac.in
... error relative to "exactly **top-k**" queries, translatable into guarantees about **query**-result precision **...** on algorithms that **process** index lists by **...**
Cited by 85 - Related Articles - View as HTML - Web Search

Fig. 8. Searching Google Scholar for "top-k query processing for semistructured data"

Example 2 (*Scores for old publications*; CiteSeer (CiteSeer, 2008)): The two plots in Figure 9 show in-citation counts of two relatively highly cited publications from CiteSeer (CiteSeer, 2008) (the observations made in this example do apply to most of the top-cited papers; check the full list posted by CiteSeer (CiteSeer-Lists, 2008)). Notice that the popularities of the two publications have dropped significantly after 2004. We observe that the probability

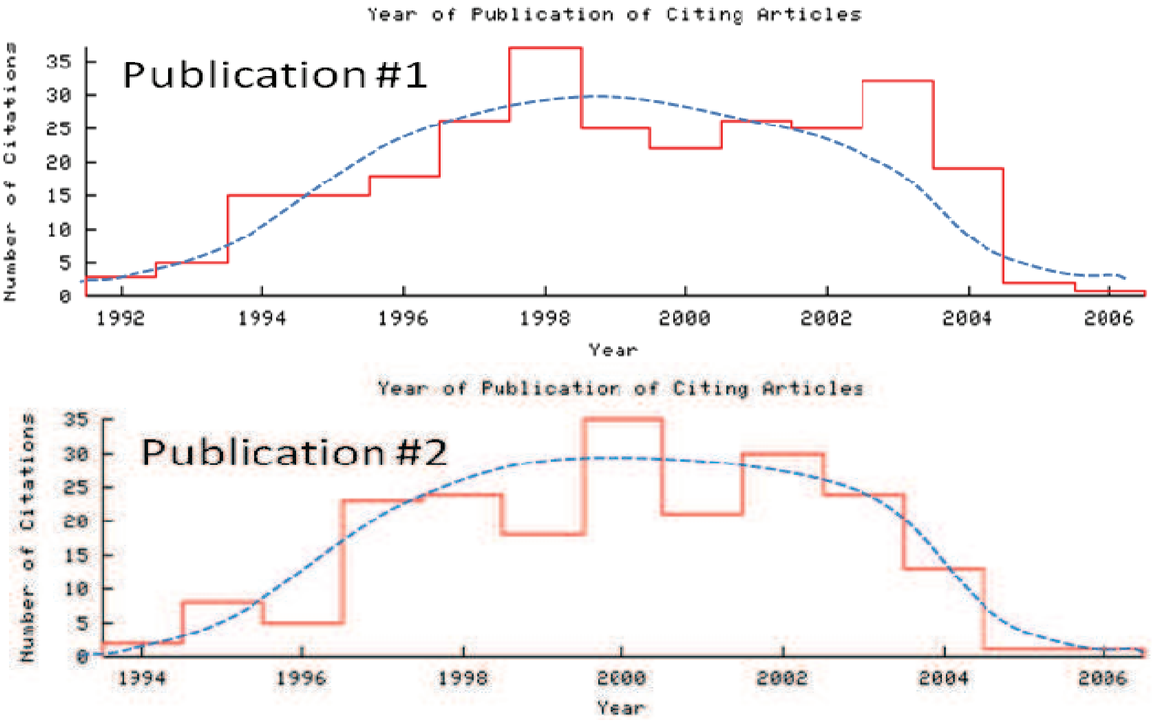
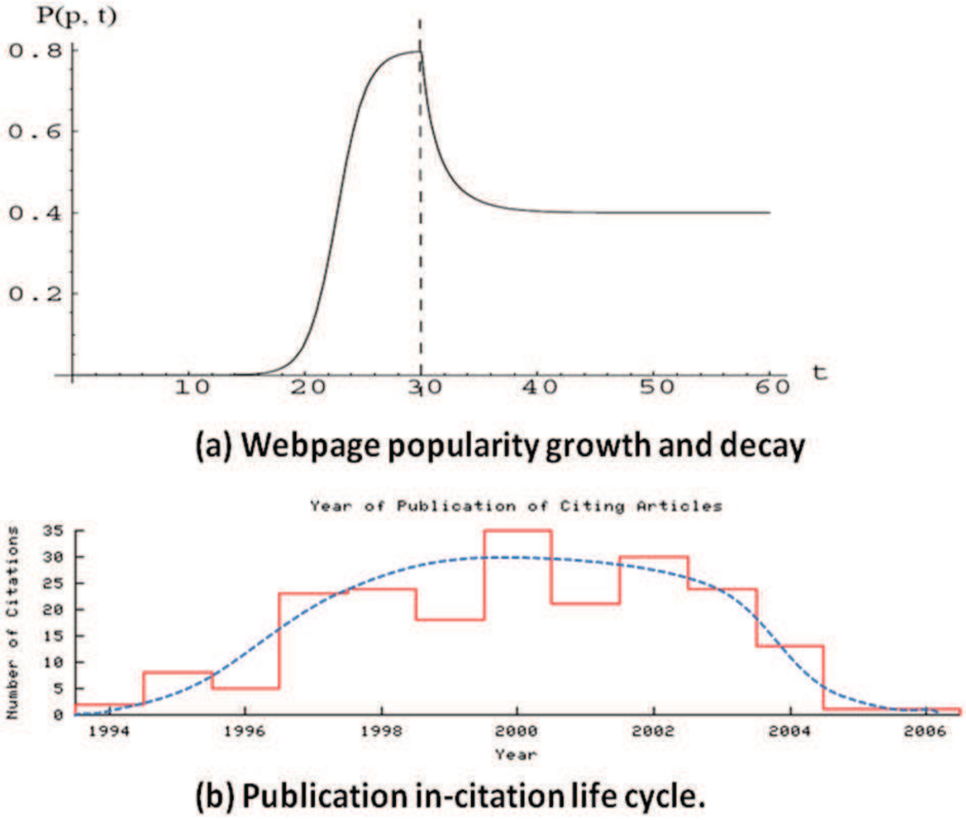


Fig. 9. Citation count per year for two publications that appeared in 1992 and 1994 (from CiteSeer) and cited around 300 times each.



(a) Webpage popularity growth and decay

(b) Publication in-citation life cycle.

Fig. 10. Popularity drop of webpages, as opposed to observed in-citation life cycle of publications.

that a publication receives new citations drops as it gets older. And, we also observe that PageRank scores of old publications reach a certain value, and do not change after that, even when they are not cited anymore. The reason is that citations do not age or disappear, and as we shall shortly explain, citation graphs around old publications minimally change. We thus conclude that PageRank scores of old publications represent their peak popularity (that they achieved in the ***past***), but not their *current* popularity. This means that, even though old publications may in time be of lower interest to present users, their PageRank scores do not change.

Based on the above two examples, we argue that, although PageRank is effective in capturing the peak popularity of publications, PageRank may assign inaccurate popularity scores *for both old and recent publications*.

In (Cho et al., 2005), a web-user model is introduced and a new *popularity growth model* of webpages is presented. Using the growth model, Cho et. al. derived a quality estimator to compute webpage quality as opposed to its peak-time popularity.

In this section, we experimentally validate the popularity growth phase model of publications proposed by (Cho et al., 2005). Moreover, we observe the following differences between publication citation and web-link graphs that the popularity growth model does not take in consideration: (i) publication citations do not ever disappear like web links, (ii) unlike web links, once two papers are published, no new citations between them are added, (iii) also unlike web links, new citations to old papers are very unlikely to occur, and (iv) indirect citations to a publication are of lesser effect on its PageRank score (Desikan et al., 2005). We observe that these differences result in popularity decay for old publications overtime, which we refer to as the *publication popularity decay phase*. In this section, these differences guide us in extending the popularity growth model to accurately capture popularity decay of publications in technology-driven fields of study where authors tend *not* to cite publications that get older, and publication quality becomes less relevant. We demonstrate that our proposal successfully assigns accurate publication scores that are in turn useful for two tasks:

i. **Ranking search results of user queries in literature digital libraries**. Accurate publication scores may help users retrieve new and yet *promising* publications; and new publications may contain undiscovered ideas at the frontiers of the topic of interest for users. Our extended quality estimator identifies high quality papers, presents them to the user, and thus gives new papers a better chance to accumulate awareness more quickly.

ii. **Modeling popularity life cycle of publications**. Coupled with the probabilistic model of researchers' citation behavior, which we discuss in section 5.4, *popularity life cycle* of publications in different publication venues can be modeled. Cho et. al. analytically verified that the quality estimator they propose can successfully be used for pages with changing quality (growth and decay) (see figure 5.3). However, they did not investigate the popularity decay of pages (Cho et al., 2005), probably because of the difficulties in capturing such web data and the complexity of web-link graph dynamics. Studies show that, for literature digital libraries, the popularity decay phase can be successfully modeled and integrated with the popularity growth phase.

Our two-phase publication popularity model, i.e., the popularity growth and decay model, is in heavily different than the webpage popularity model. To illustrate the differences, figure 10 shows two popularity growth and decay curves, one for a webpage (figure 5.3.a from (Cho et al., 2005)) and another for a publication (figure 10.b from CiteSeer (CiteSeer, 2008)). Notice that the popularity of a webpage keeps increasing as the webpage becomes known and those who "like" it place links to it in other pages (Cho et al., 2005). After the webpage reaches the peak, its popularity decays until it reaches a steady-state popularity value (Cho et al., 2005). In comparison, the decay of publication popularity has a much different curve. Studies show that researchers rarely cite old works, especially in fast-moving fields like computer and life sciences. Consequently, we show that, by properly modeling users' citation behavior along with accurate publication quality estimators, we obtain realistic publication popularity growth and decay curves similar to the dashed curves of figure 10.b. Empirically, we observe that the majority of publication "citation count per year" curves conform to this growth and decay model.

4.2 Page quality and webpage popularity evolution model

Cho et. al. (Cho et al., 2005), via a simple user-web model, developed a formula for the popularity growth of webpages, and then used the formula to estimate page quality.

Publication quality, based on the web-user model, is defined as the popularity of the publication given that all possibly interested authors are aware of it and those who like it have cited it.

After getting published, a paper goes through two main phases:

i. a **popularity growth phase** where its popularity increases as more authors become aware of it and cite it. After some time, the publication's popularity reaches to a certain value. During the growth phase of the publication, (i) researchers develop awareness of the publication, i.e., more authors get to know it, and (ii) research problems inspired by the paper get studied by authors. This means that the longer the growth phase of a paper, the better the quality of the paper; and (iii) the authors who *like* the paper cite it in their works.

ii. a **saturation phase**: after the transient growth phase, the publication's PageRank score settles at a certain value, and minimally changes.

Definition:

1. The ***growth region*** of a publication is the time interval during which the publication popularity grows.
2. The ***saturation region*** of a publication is the time interval that starts at the saturation point; and, afterwards, the publication usually does not receive new citations.
3. The ***popularity function*** $\boldsymbol{P}(p,t)$ of publication p, is a function that computes the popularity of p at time t.
4. ***Publication quality*** $Q(p)$ is the intrinsic and (saturation-time popularity) quality of a publication (Cho et al., 2005).

We empirically calculate an estimation $\tilde{Q}(p)$ for the publication quality $Q(p)$ of publication p as the PageRank score at the saturation region. Or, $\tilde{Q}(p) = PR(p, t_{sat})$ where $PR(p, t_{sat})$ is PageRank score of p at the saturation time point t_{sat}.

The popularity *growth* function $\boldsymbol{P}(p,t)$, proposed in (Cho et al., 2005), is derived as:

$$\boldsymbol{P}(p,t) = Q(p)/(1 + C_1.e^{-\beta t}) \tag{1}$$

Note that the function $\boldsymbol{P}(p,t)$ is monotonically increasing with time t. The constant $Q(p)$ is the intrinsic quality of the publication p (that is estimated as p's PageRank score in the saturation region), constant C_1 is the rate of PageRank score growth in Cho's PageRank score growth model. For new publications, $\boldsymbol{P}(p,0) \cong 0$. In time, the exponent component, $e^{-\beta t}$, approaches zero as t increases, and, consequently, $\boldsymbol{P}(p,t)$ converges to $Q(p)$, the intrinsic quality score of the publication, over time.

Remark: The popularity of a publication p at time t is estimated as the p's PageRank score based on the citation graph at time t. Also, the quality of publication p is estimated as the PageRank score at saturation phase (Cho et al., 2005).

The above remark forms a bridge between the PageRank score change curve and Cho et. al.'s popularity growth model (and our model of publication popularity growth and decay model). (Cho et al., 2005) base their model on the fact that the quality of a page is time-invariant and does not change overtime. Thus; $Q(p)$ is assumed to be a constant estimated at any time as the sum of (a) the current popularity or PageRank score of p, and (b) the relative popularity (PageRank) rate of change, i.e.,

$$\tilde{Q}(p) = PR(p,t) + \frac{1}{c}.\frac{d\boldsymbol{P}(p,t)}{dt}.\frac{1}{PR(p,t)} \quad (2)$$

where $0 < c \leq 1$ is a constant which we choose to be 0.1 as in (Cho et al., 2005).

A high quality publication is one with a scientific value, and one can intuitively estimate the quality of a publication based on its impact on other authors. Quantitatively, the quality can be measured as the conditional probability that an author will like the publication (L_p) given that s/he has became aware of it (A_p). Mathematically, $Q(p) = P(L_p|A_p)$, as defined in (Cho et al., 2005).

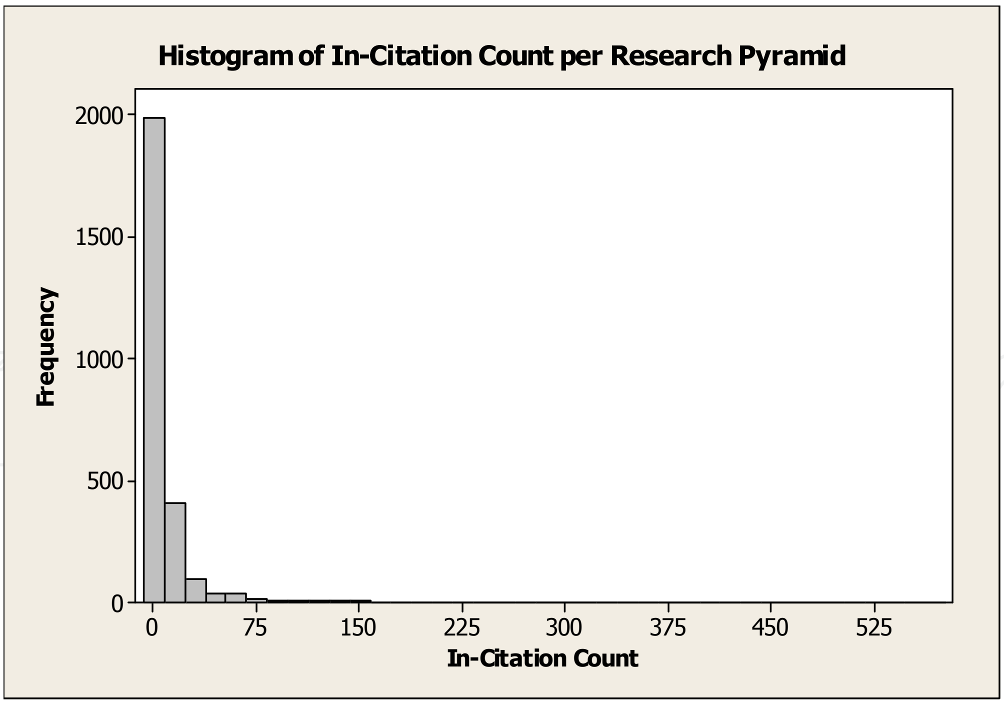


Fig. 11. In-citation per research pyramid (inter-research pyramid citations, i.e. citations from publications outside an RP to ones inside it).

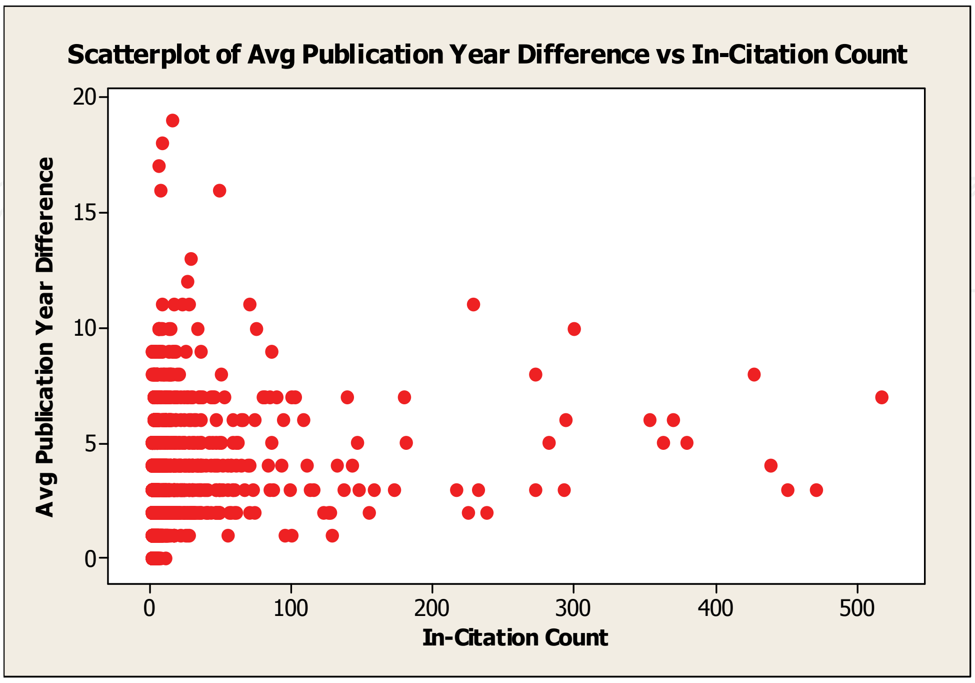


Fig. 12. Inter-pyramid Citation count (x-axis) vs the (average difference in publication dates of publications in a research pyramid).

We argue that we need to distinguish between two measures of quality for a publication.

i. The first measure represents the scientific value of that publication (i.e., how well-written it is, the authors follow a suitable technique to solve the research problem, …, etc). This value is time-invariant and is represented by $Q(p)$ (Cho et al., 2005).

ii. The second measure represents the value of the paper to the user at the time s/he is searching the digital library. This value, in contrast to $Q(p)$, is time-dependent, especially in fast-moving fields of study. We refer to this quality measure as the *Publication Quality with Aging Factor.*

Next in the following subsection, we show that publications go through the popularity growth phase during which publications gain awareness and thus popularity. And in section 5.6, we empirically show the popularity growth curves conform to the “sigmoidal” evolution pattern derived by (Cho et al., 2005). Finally, in section 5.4, we study one aspect of researcher citation behavior, and use it in section 5.5 to propose our notion of *Publication Quality with Aging Factor.*

4.3 Properties of publication citation graphs and research pyramids

In this section we validate Cho et. al.’s popularity growth phase by (i) using PageRank as a popularity indicator, and (ii) utilizing the research-pyramid model of research evolution (Bani-Ahmad & Ozsoyoglu, 2007; Aya et al., 2005), to show that popularity scores of publications converge to a steady-state value that can be estimated by equation (2) above.

We first note one difference between a publication citation graph from a web citation graph: Publication citation graph evolution behavior is to some extent more controlled than web graphs and can be anticipated. A webpage that has been on the web for a relatively long time may still receive new links (citations); old publications, however, are rarely cited (Ahmed et al., 2002; Bani-Ahmad & Ozsoyoglu, 2007; Case & Higgins, 2000). Consequently, publication citation graphs are highly unlikely to face structural changes around relatively old publications. This special characteristic of publication citation graphs allows for developing accurate mathematical models for changes to publication's PageRank scores, and thus better estimation of publication quality. In contrast, a web graph may face abrupt structural changes at any time in any part of the graph. Studies show that, every week, around 8% web pages are replaced and that about 25% new links are created (Ntoulas et al., 2004).

Next we describe the research-pyramid (RP-) model (Bani-Ahmad & Ozsoyoglu, 2007; Aya et al., 2005) of publications that also suggests time-dependent growth patterns in publication citation graphs. The RP-Model is based on the observation that citations between research publications produce multiple, small pyramid-like structures, where each pyramid represents publications related to a highly specific research topic (Aya et al., 2005). A research pyramid is defined as a set of publications that represent a highly specific research topic, and usually has a pyramid-like structure in terms of its citation graph (Aya et al., 2005; Bani-Ahmad & Ozsoyoglu, 2007).

The RP-Model suggests that publication citation graphs evolve in a time-controlled manner through the stimulation of most-specific research topics from one another as follows. A publication that deals with a new specific research problem appears, and proposes the first solution for it. More publications appear after that publication, addressing the same problem and proposing enhanced or refined solutions to that problem. In time, the research problem (i) is either solved, (ii) settles down with "good-enough" solutions, or (iii) subdivided into more specific research problems (i.e., new research pyramids) (Bani-Ahmad & Ozsoyoglu, 2007).

Publications within an individual research pyramid are (i) motivated by earlier publications in the topic area, or (ii) use techniques proposed in publications from other research pyramids. We have observed that citations between different research pyramids conform to a highly left-skewed distribution, (figure 13), which indicates that as research pyramids of a particular research topic is formed and new research pyramids are instantiated, the RPs already formed receive few external citations from other research pyramids.

Consequently, publication citation graphs are highly unlikely to face structural changes within an already constructed research pyramid because (i) citations do not disappear like web links, (ii) once two papers are published, no new links between them are added, (iii) new citations to old paper are less likely to occur, and (iv) indirect citations to a publication are of lesser effect on its PageRank score (Desikan et al., 2005). Structural changes affect only the developing (i.e., recent) research pyramids. Thus, popularity (or PageRank scores) of publications are expected to converge over time to a steady-state value, which is the essence of the popularity growth model (Cho et al., 2005).

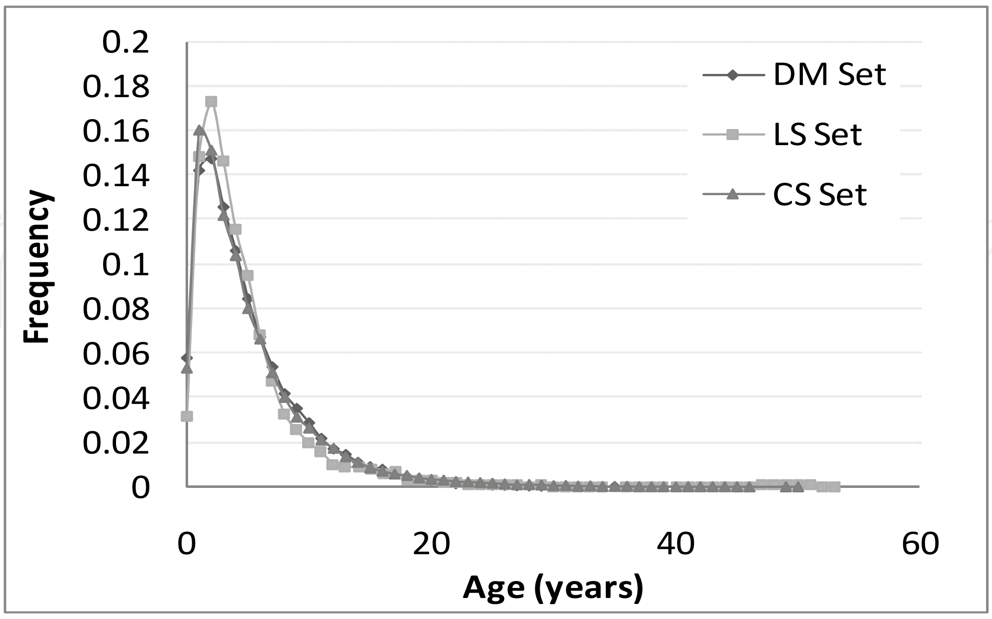


Fig. 13. Empirical citation-age probability distribution curves (i.e., citation age vs frequency of citations with that age) of publications in three datasets (i) Data management (2) life sciences and (iii) computer science.

4.4 The user citation behavior model

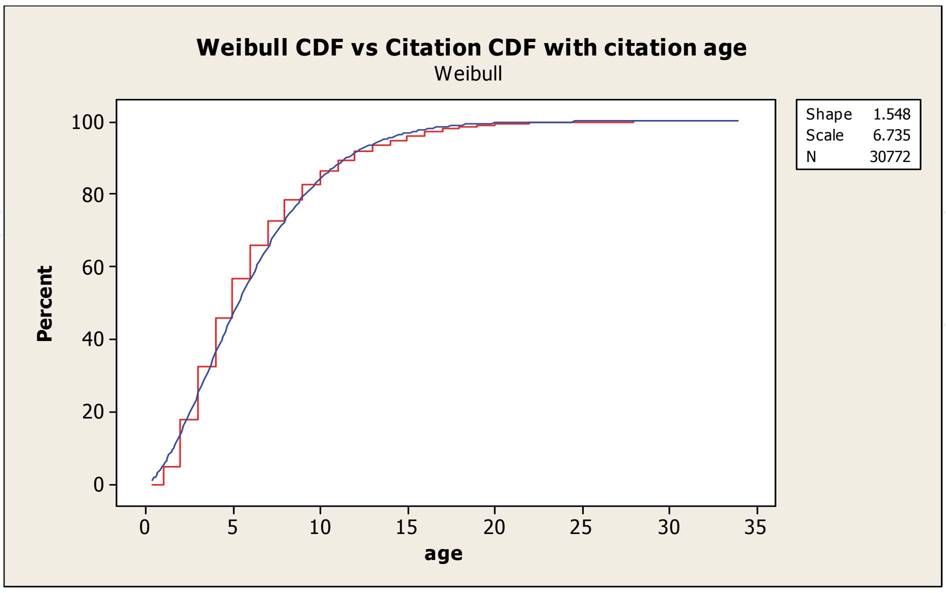


Fig. 14. (a) Age vs frequency of citations of publications. (b) Weibull distribution CDF

Figure 13 shows that user interest in citing a particular paper significantly decays over time. The best probabilistic distribution that fits the citation-age PDFs of figure 13 is the Weibull distribution (Mathworks, 2008). Figure 14 contains the cumulative distribution function (CDF) of the Weibull distribution, and the empirical CDF of the citation-age distribution for the data-management dataset. The two CDF curves show a high match. Using Minitab, 2008 software (Minitab, 2008), we have observed that the citation age curve (figure 14) conforms to the Weibull distribution with the estimated parameters shape (γ)=1.548 and Scale (α)=6.735. Thus, the probability $P(u \rightarrow v)$ of the citation from u to v to occur, is computed as

$$P(u \rightarrow v) = f_{weibull}(|age(u,v)|;\ \gamma, \alpha) \tag{3}$$

where $|age(u,v)|$ is the absolute time difference (in years) between the publication years of u and v. The probability density function of Weibull distribution is given by

$$f_{weibull}(x;\ \gamma, \alpha) = \frac{\gamma}{\alpha}.(x/\alpha)^{\gamma-1}e^{-(\frac{x}{\alpha})^{\gamma}} \tag{4}$$

assuming that $f_{weibull}(x;\ \gamma, \alpha)$ = 0 for x < 0 (which is true in our case as a publication will not receive any citation if it is not published). In section 5.5, we use this formula in estimating the publication quality considering the aging factor.

4.5 Publication quality with aging factor

Assume that a user issues a search query at time ***t***. Viewing the user as a potential author of an upcoming publication, the user will probably follow the Weibull distribution in his/her citations. i.e., the user cites a relevant publication v with probability equal to $f_{weibull}(t - t_{Year}(v))$ where $t_{Year}(v)$ is the publication year of v.

Thus, we argue that considering both the publication quality and the aging factor together leads to a better search output ranking. One possible way to order user search query results is to consider three factors: (i) text-based relevancy of v and the query terms, (ii) the publication quality, (iii) the probability that the user will cite the publication given the ages of relevant publications. Thus, for a given search query term w, and output (publication) v, one possible form of combining the three factors is as follows

$$final_{score(v)} = Sim(w,v) * \ \widehat{\boldsymbol{P}}(p,t) \tag{5}$$

where $Sim(w,v)$ is the text-based similarity between w and v, and $\widehat{\boldsymbol{P}}(p,t)$ is the *temporal popularity of the publication* at time t which is computed as

$$\widehat{\boldsymbol{P}}(p,t) = f_{weibull}(t - t_{Year}(v);\ \gamma, \alpha) * \boldsymbol{P}(p,t)$$

Definition: The ***temporal popularity*** of a publication p at time t, $\widehat{\boldsymbol{P}}(p,t)$ represents users' expected interest in p at t.

4.7 Section summary

In this section, we have (i) experimentally validated the popularity growth phase of publications (Cho et al., 2005), (ii) proposed a probabilistic model for domain-specific publication citation behavior, and (iii) extended *the popularity growth phase* to capture publication popularity decay phase.

5. Chapter summary and future research directions

5.1 Chapter summary

In this chapter, we have introduced a number of recent techniques for ranking the search results of online digital libraries.

Evaluating citation-based score measures of publications.

In section 2 of this chapter, we compared and evaluated several publication score functions; including PageRank (Brin & Page, 1998), Authorities (Kleinberg, 1998) and citation-count scores (Chakrabarti, 2003). We observed the separability problem with all of these functions, which is defined as the scoring functions producing scores that do not distribute well over a given scale, e.g., [0, 1]. Instead, distributions of the existing publication score functions are highly skewed, and decay very fast (Render, 2004), resulting in a much less useful comparative publication assessment capability for users. This lack of separability is caused by the "rich gets richer" phenomena (Render, 2004; Li & Chen, 2003), i.e., a very small number of publications with relatively high numbers of in-citations have even higher chances of receiving new citations. Yet, these scoring functions are still not very accurate, probably due to topic diffusion in search outputs (Haveliwala, 2002).

Improved publication scores via research-pyramids

In section 3, we observed that (a) the complete publication citation graph (of AnthP) is highly clustered, (b) each cluster of the complete publication set has a pyramid-like structure in terms of the citation graph of the cluster, and (c) each cluster represents a highly specific research topic. These three observations validated the research pyramid model proposed by (Aya et al., 2005).

We also found that topic similarities decay over both citation ages and citation paths. We used two topic similarity decay curves to guide the research-pyramid construction, and proposed and validated two algorithms to identify research pyramid structures in citation graphs.

Within research-pyramid citation graphs, we noticed that the average number of in-citations per paper varies, pointing to the importance of comparative publication scores within research pyramids. We then observed that normalizing publication scores within research-pyramids produces accurate and nearly normally distributed scores of publications.

Popularity Growth and Decay of Publications

In section 4, we proposed new definitions for popularity growth and decay for publications by coupling Cho et. al.'s model of popularity growth with our probabilistic publication citation behavior model, which we referred to as *the publication quality with aging factor*. In detail, we (i) experimentally validated the popularity of publications change over time and follow the logistic growth equation (Cho et al., 2005), (ii) proposed an empirical model for one aspect of researchers' citation behavior in technology-driven fields of study such as computer science (this model captures researchers' tendency not to cite old publications), and (iii) extended the popularity growth model (Cho et al., 2005) to capture publication popularity decay. Our major findings were as follows: **(a)** empirically, the probability of citing any publication conforms to the Weibull distribution (Mathworks, 2008) over the age of that publication. However, the shape and scale parameters of the distribution changes with the quality of publication venues, **(b) w**e showed that the derivative of the popularity growth function accurately represents (i.e., directly proportional to) the temporal publication popularity at any time, **(c) w**e observed that our definition of *publication quality with aging factor* matches the derivative of the popularity growth curve. This provides an analytical foundation for our growth and decay model of publication popularity.

5.2 Future research directions

Advanced Search Interface via Research Pyramids

As future work, one may work on the problem of automatically annotating research pyramids with keywords representing fine-grained research topics. Also, by using the identified research pyramids, we may work on visualization, namely, building a hierarchical structure that places research pyramids into a hierarchical structure. Using RP annotations and the hierarchical structure of RPs, building an advanced query interface that involves pruned searches becomes possible.

Accurate Identification of Research Pyramids

The two RP-identification algorithms proposed in section 2 are very basic, and form the first attempts. As future work, one may find more accurate techniques to identify cornerstone publications within research pyramids. Also, more accurate techniques to identify members of each RP need to be developed.

Publication-venue Specific User Citation Behavior

As future work, one can work on identifying the correlation between the impact of the publication venue on user's citation behavior and publications that appear in prestigious conferences. More specifically, one may attempt to model users' citation behavior for prestigious publication venues. Our hypothesis is that, by understanding users' citation behavior, one can provide users of online digital libraries with higher quality of services.

6. References

ACM Digital Library (2008), http://portal.acm.org/dl.cfm. Viewed in March 2008.

ACM SIGMOD Anthology (2003), http://www.acm.org/sigmod/dblp/db/anthology.html. Viewed in 2003.

Ahmed, T.; Johnson, B.; Oppenheim, C. & Peck, C. (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, 2004.

Aya, S.; Lagoze, C.; & Joachims, T. (2005). Citation Classification and its Applications, International Conference on Knowledge Management.

Bani-Ahmad*, S.; Cakmak A.; Ozsoyoglu, G. & Al-Hamdani, Abdullah (2005). Evaluating Score and Publication Similarity Functions in Digital Libraries. ICADL, 2005.

Bani-Ahmad, S. & Ozsoyoglu, G. (2007). Improved Publication Scores for Online Digital Libraries via Research Pyramids. ECDL 2007.

Bani-Ahmad, S.; Cakmak, A.; Ozsoyoglu, G. & Al-Hamdani A. (2005). Evaluating Publication Similarity Measures, IEEE Data Eng. Bull. 28(4): 21-28, 2005

Brin, S. & Page, L. (1998), The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems.

Cakmak, A. (2003). HITS- and PageRank-based Importance Score Computations for ACM Anthology Papers, Tech. report, EECS Dept, CWRU, 2003.

Case, D. O. & Higgins, D. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication, Jour. Of American Society of Information Science, 51(7):635-645, 2000.

Chakrabarti, S. (2003), Mining the Web, Morgan-Kauffman, 2003.

Cho, J.; Roy, S. & Adams, R. (2005). Page Quality: In Search of an Unbiased Web Ranking, ACM SIGMOD.

CiteSeer (2008), www.citeseer.com. Viewed in March 2008.

CiteSeer-Lists (2008). List of Most cited articles in Computer Science, http://citeseer.ist.psu.edu/articles.html. viewed on March 2008.

DBLP (2003). The DBLP Computer Science Bibliography. http://www.informatik.uni-trier.de/~ley/db. Viewed in April 2003.

Desikan, P.; Pathak, N.; Srivastava, J. & Kumar, V. (2005). Incremental page rank computation on evolving graphs. In the proceedings of WWW conference 2005.

Google Scholar (2008), http://scholar.google.com/scholar. Viewed in March 2008.

Haveliwala, T. H. (2002). Topic-sensitive PageRank, WWW Conference, Hawaii, 2002.

IEEE Xplore (2008), http://ieeexplore.ieee.org. Viewed in March 2008.

Jeh G. & Widom, J. (2002). SimRank A measure of structural-context similarity, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

Kleinberg, J. (1998). Authoritative sources in hyperlinked environments, The 9th ACM-SIAM Symposium on Discrete Mathematics (SODA) Conference, 1998.

Li, X. & Chen, G. (2003). A local-world evolving network model, Physica A 328 (2003) 274 – 286

Mathworks (2008), http://www.mathworks.com/. Viewed in April 2008.

Minitab (2008). Minitab Statistical Software, http://www.minitab.com/. Viewed in April 2008.

Ntoulas, A.; Cho, J. & Olston, C. (2004). What's new on the web?: the evolution of the web from a search engine perspective. WWW '04.

Pan, F. (2006). Comparative Evaluation of Publication Characteristics in Computer Science and Life Sciences, MS Thesis, EECS, Case Western Reserve University, 2006.

PubMed (2008). http://www.ncbi.nlm.nih.gov/entrez/query.fcgi. Viewed in March 2008.

Ratprasartporn, N. & Ozsoyoglu, G. (2007). Finding Related Papers in Literature Digital Libraries, ECDL 2007.

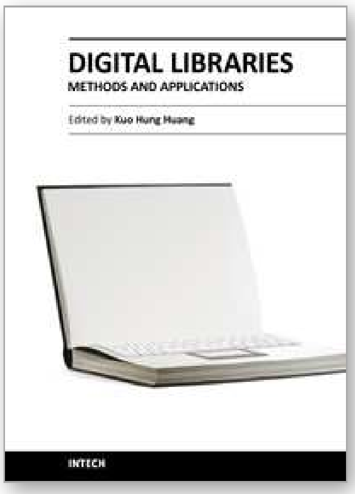
Ratprasartporn, N., Po, J., Cakmak, A., Bani-Ahmad, S., Ozsoyoglu, G., On Context-Based Publication Search Paradigm: Gene-Ontology-Specific Contexts for Searching PubMed Effectively. Technical Report, CWRU 2006.

Ratprasartporn, N.; Po, J.; Cakmak, A.; Bani-Ahmad, S. & Ozsoyoglu, G (2007). Evaluating utility of different ranking functions in context-based environment, DBRank Workshop, Istanbul, Turkey, April 2007.

Redner, S. (2004). Citation statistics from more than a century of physical review. Physics 0407137, 2004.

ScienceDirect (2008). www.sciencedirect.info. Viewed in March 2008.

Wasserman, S. & Faust, K. (1994). Social Network Analysis, Cambridge U. Press, Cambridge, 1994



Digital Libraries - Methods and Applications

Edited by Dr. Kuo Hung Huang

ISBN 978-953-307-203-6

Hard cover, 220 pages

Publisher InTech

Published online 04, April, 2011

Published in print edition April, 2011

Digital library is commonly seen as a type of information retrieval system which stores and accesses digital content remotely via computer networks. However, the vision of digital libraries is not limited to technology or management, but user experience. This book is an attempt to share the practical experiences of solutions to the operation of digital libraries. To indicate interdisciplinary routes towards successful applications, the chapters in this book explore the implication of digital libraries from the perspectives of design, operation, and promotion. Without common agreement on a broadly accepted model of digital libraries, authors from diverse fields seek to develop theories and empirical investigations that to advance our understanding of digital libraries.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Sulieman Bani-Ahmad (2011). Sorting Search Results of Literature Digital Libraries: Recent Developments and Future Research Directions, Digital Libraries - Methods and Applications, Dr. Kuo Hung Huang (Ed.), ISBN: 978-953-307-203-6, InTech, Available from: http://www.intechopen.com/books/digital-libraries-methods-and-applications/sorting-search-results-of-literature-digital-libraries-recent-developments-and-future-research-direc

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821