This is the preliminary version of the JASIST submission

Topic-based PageRank on Author Co-citation Networks

Ying Ding[1]
School of Library and Information Science, Indiana University, Bloomington, Indiana, USA
1320 E 10th, Bloomington, IN 47405, USA
Tel: +1 812 855 5388
Fax: +1 812 855 6166

Abstract

Ranking authors is vital for identifying a researcher's impact and his standing within a scientific field. There are many different ranking methods (e.g., citations, publications, h-index, PageRank, and weighted PageRank), but most of them are topic-independent. This paper proposes topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm. The Author-Conference-Topic (ACT) model was used to extract topic distribution of individual authors. Two ways for combining the ACT model with the PageRank algorithm are proposed: simple combination (I_PR) or using a topic distribution as a weighted vector for PageRank (PR_t). Information retrieval was chosen as the test field and representative authors for different topics at different time phases were identified. Principal Component Analysis (PCA) was applied to analyze the ranking difference between I_PR and PR_t.

Introduction

Society produces much more data now than at any other time in human history (Lyman & Varian, 2003). Data are no longer isolated but are linked via various relationships, including cited/citing, co-author, co-occur, friend-of, know, vote-for, and favorite-of. Developed by Google, the PageRank algorithm determines the most important webpages by examining the entire linking structure of the Web. It pioneered the notation of weighted votes and the consideration of graph topology for ranking (Brin & Page, 1998). PageRank and its variations have been widely applied in bibliometrics to rank authors (Ding, Yan, Frazho, & Caverlee, 2009), journals (Bollen, Rodriguez, & Van de Sompel, 2006; Leydesdorff, 2009), and articles (Yan & Ding, forthcoming).

Pages that are important in one domain may not be important in another. PageRank computes a ranking value for each node in a network (e.g., author co-citation network, co-authorship network, journal citation

[1] Corresponding author, Email: dingying@indiana.edu

network, etc.) without considering topical features of a node including the research area in which an author is interested, the domain in which a journal is published, or the topics that an article addresses. Various approaches have been proposed to add the number of publications or the h-index values of authors as weighted vectors to PageRank (Ding, 2011 forthcoming). However, how to link topics in PageRank to provide a topic-based ranking (i.e., one that ranks nodes in the graph while simultaneously considering the topical features of nodes) has not been fully-explored in bibliometrics.

Latent Dirichlet Allocation (LDA) captures the topical features of nodes by postulating a latent structure for a set of topics linking words and documents. The LDA method has been shown to be reliable for detecting multinomial word distribution of topics (Blei et al., 2003). As the extended LDA model, the Author-Topic model proposed by Rozen-Zvi and her colleagues (2004) depicts the content of documents and the interests of authors simultaneously. Later, Tang, Jin and Zhang (2008) extended LDA to reveal the topic distribution of authors, conferences, and citations concurrently (Tang, et al., 2008).

Searching, recommending, or ranking authors at the topic level is highly demanded. Although topic-sensitive PageRank was proposed to address this particular issue (Haveliwala, 2002), it was based on topics that were manually pre-defined rather than automatically extracted. This paper uses an information retrieval dataset with 15,370 articles and 341,871 citations covering the period of 1956-2008 as the test data and proposes two ways of combining LDA with PageRank. Results indicated that the suggested topic-based PageRank algorithms can rank authors by considering their research interests therefore providing value-added information to facilitate scientific collaboration.

The contributions of this paper are follows: 1) it proposed two ways of combining author topic distributions with the PageRank algorithm to calculate the topic-based PageRank scores for authors; 2) it utilized the extended LDA model (i.e., the ACT model) to extract author and document topic distributions simultaneously; 3) bibliometrically, it contributed to the field by being able to rank authors based on PageRank at the topic level, which brings finer granularity to ranking experts; and 4) it is among the first to apply topic-based PageRank to rank scholars and evaluate their research impact in the field of Information Retrieval. This paper is structured as follows: Section 2 reviews the current literature on various LDA and weighted PageRank algorithms; Section 3 describes the process of data collection and the method of merging LDA with PageRank; Section 4 discusses results and correlates them with related measures; and Section 5 suggests possible future work.

Related Work

This Section begins with an introduction to available topic models and then shifts to existing efforts on aligning topics with various weighted PageRank algorithms.

Topic Models

Topics can be automatically extracted from a set of documents by utilizing different statistical methods. Figure 1 shows the plate notation for the major topic models, with gray and white circles indicating observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and plates. Plates indicate repeated sampling with the number of repetitions given by the variable in the lower corner (Buntine, 1994). Here, *d* is a document, *w* is a word, a_d is a set of co-authors,

x is an author, z is a topic, α, β and μ are hyperparameters, θ, ϕ and ψ are multinomial distributions over topics, words and publication venues, respectively. Table A5 in the Appendix lists notations for formulas discussed in this subsection.

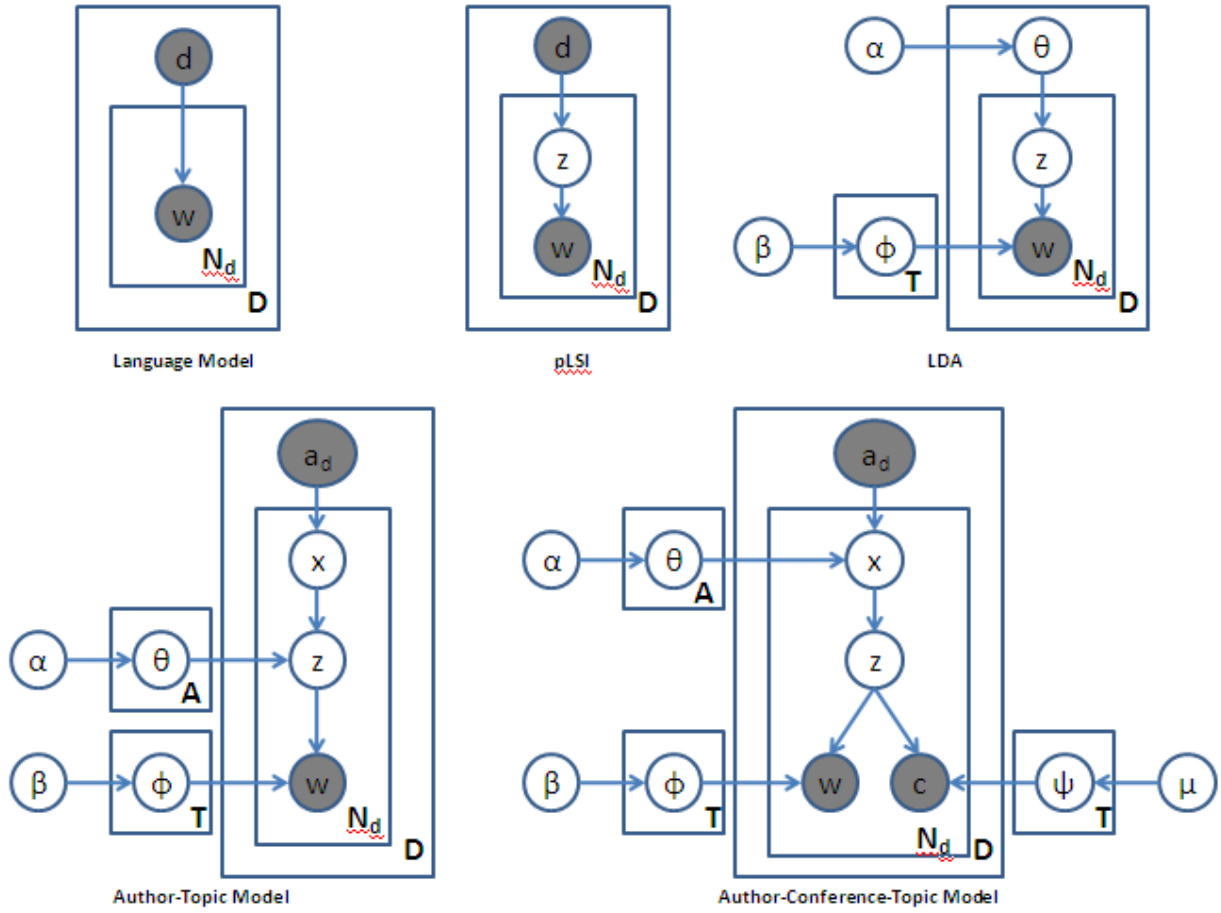

Figure 1: Various LDA models

Language Model (LM)

The Language Model is an early effort in natural language processing and information retrieval to assign a probability distribution to words. There is no latent variable in this model (see Figure 1). For a given query q, the probability between a document and a query word is calculated as (Ponte & Croft, 1998):

$$P(w|d) = \frac{N_d}{N_d + \lambda} \times \frac{tf(w,d)}{N_d} + (1 - \frac{N_d}{N_d + \lambda}) \times \frac{tf(w,D)}{N_D}$$

where *tf(w,d)* is the word frequency of a word w in a document d, N_d is the number of words in the current document, N_D is the number of words in the entire collection, *tf(w,D)* is the frequency of a word w in the collection D, and λ is the Dirichlet smoothing factor and usually set as the average document length in the collection (Zhai & Lafferty, 2001).

Probabilistic Latent Semantic Indexing (pLSI)

Hofmann (1999) proposed the probabilistic Latent Semantic Indexing (pLSI) model by introducing a latent topic layer z between words and documents (see Figure 1). In this model, the probability of generating a word w from a document d is based on the latent topic layer as:

$$P(w|d) = \sum_{z=1}^{T} P(w|z)P(z|d)$$

pLSI does not provide a mathematical grounding for this latent topic layer and is susceptible to severe overfitting (Blei, Ng, & Jordan, 2003).

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) provides a probabilistic model for the latent topic layer (Blei, Ng, & Jordan, 2003). For each document *d*, a multinomial distribution θ_d over topics is sampled from a Dirichlet distribution with parameter α. For each word w_{di}, a topic z_{di} is chosen from the topic distribution. A word w_{di} is generated from a topic-specific multinomial distribution $\phi_{z_{di}}$. The probability of generating a word *w* from a document *d* is:

$$P(w|d, \theta, \phi) = \sum_{z \in T} P(w|z, \phi_z)P(z|d, \theta_d)$$

Therefore, the likelihood of a document collection *D* is defined as:

$$P(Z, W|\Theta, \Phi) = \prod_{d \in D} \prod_{z \in T} \theta_{dz}^{n_{dz}} \times \prod_{z \in T} \prod_{v \in V} \phi_{zv}^{n_{zv}}$$

where n_{dz} is the number of times that a topic *z* has been associated with a document *d*, and n_{zv} is the number of times that a word w_v has been generated by a topic *z*. The model can be explained as: to write a paper, an author first decides what topics and then uses words that have a high probability of being associated with these topics to write the article.

Author-Topic model

Rosen-Zvi, Griffiths, Steyvers, and Smith (2004) proposed the Author-Topic model to represent both document content and author interests. An author is chosen randomly when a group of authors a_d decide to write a document *d* containing several topics. A word *w* is generated from a distribution of topics specific to a particular author. There are two latent variables, *z* and *x*. The formula to calculate these variables is:

$$P(z_i, x_i|z_{-i}, x_{-i}, w, a_d, \alpha, \beta) \propto \frac{C_{mj}^{wT} + \beta}{\sum_{m'}(C_{m'j}^{wT} + V\beta)} \times \frac{C_{kj}^{AT} + \alpha}{\sum_{j'}(C_{kj'}^{AT} + T\alpha)}$$

where z_i and x_i represent the assignments of the *i*th word in a document to a topic *j* and an author *k* respectively, *w* represents the observation that the *i*th word is the *m*th word in the lexicon, z_{-i} and x_{-i} represent all topic and author assignments not including the *i*th word, and C_{kj}^{AT} is the number of times an author *k* is assigned to a topic *j*, not including the current instance. The random variables ϕ (the probability of a word given a topic) and θ (the probability of a topic given an author) can be calculated as:

$$\phi_{mj} = \frac{C_{mj}^{wT} + \beta}{\sum_{m'}(C_{m'j}^{wT} + V\beta)}$$

$$\theta_{kj} = \frac{C_{kj}^{AT} + \alpha}{\sum_{j'}(C_{kj'}^{AT} + T\alpha)}$$

This model can be used to recommend reviewers for peer-reviewed journals. The outcome of this model is a list of topics, each of which is associated with the top-ranked authors and words. Top-ranked authors are not necessarily the most highly cited authors in that area, but are the productive authors who produce the most words for a given topic (Steyvers, Smyth & Griffiths, 2004). Top-ranked words of a topic are those having a high probability of being selected when an author writes a paper on that particular topic.

Author-Conference-Topic Model

Tang, Jin and Zhang (2008) proposed the Author-Conference-Topic (ACT) model, an extended LDA used to model papers, authors, and publication venues simultaneously. Conference represents a general publication venue (e.g., journal, workshop, and organization). The ACT model can be interpreted as: coauthors determine the topics for a paper, and each topic generates words and determines a publication venue. The ACT model calculates the probability of a topic for a given author, the probability of a word for a given topic, and the probability of a conference for a given topic. Gibbs sampling is used for inference, and the hyperparameters α, β, and μ are set at fixed values (α=50/T, β=0.01, and μ=0.1). The posterior distribution is estimated on *x* and *z*, and the results are used to infer θ, φ, and ψ. The posterior probability is calculated as:

$$P(z_{di}, x_{di} | z_{-di}, x_{-di}, w, c, \alpha, \beta, \mu) \propto \frac{m_{x_{di}z_{di}}^{-di} + \alpha_{z_{di}}}{\sum_z (m_{x_{di}z}^{-di} + \alpha_z)} \times \frac{n_{z_{di}w_{di}}^{-di} + \beta_{w_{di}}}{\sum_{w_v}(n_{z_{di}w_v}^{-di} + \beta_{w_v})} \times \frac{n_{z_{di}c_d}^{-d} + \mu_{c_d}}{\sum_c (n_{z_{di}c}^{-d} + \mu_c)}$$

After Gibbs sampling, the probability of a word given a topic φ, probability of a conference given a topic ψ, and probability of a topic given an author θ can be estimated as:

$$\phi_{zw_{di}} = \frac{n_{zw_{di}} + \beta_{w_{di}}}{\sum_{w_v}(n_{zw_v} + \beta_{w_v})}$$

$$\psi_{zc_d} = \frac{n_{zc_d} + \mu_{c_d}}{\sum_c (n_{zc} + \mu_c)}$$

$$\theta_{xz} = \frac{m_{xz} + \alpha_z}{\sum_{z'}(m_{xz'} + \alpha_{z'})}$$

A paper d is a vector w_d of N_d words in which each w_{di} is chosen from a vocabulary of size V. A vector a_d of A_d authors is chosen from a set of authors of size A, and c_d represents a publication venue. A collection of papers D is defined by $D = \{(w_1, a_1, c_1), \ldots (w_{D,} a_D, c_D)\}$. The number of topics is denoted as T.

<u>Topic-related PageRanks</u>

The original PageRank algorithm calculates a single PageRank vector based on the overall structure of the Web regardless of the specific features of an individual node. Topic-related PageRank studies generally concentrate on solving information retrieval issues by better aligning search results with user queries. Bharat and Henzinger (1998) defined a set of topics for a corpus based on the concatenation of the first

1,000 words of each document, added weights to each node based on the similarity of document content and the defined set of topics, and calculated the influence scores of articles for each topic based on the weighted HITS (Kleinberg, 1998). Chakrabarti, Dom, Gibson and Kleinberg (1998) applied the weighted HITS algorithm for automatically compiling topic-centered resource lists. Bharat and Mihaila (2001) proposed the Hilltop algorithm which first identified query-topic-specific hub websites and calculated HITS scores for authority websites. Compared with HITS, PageRank can be calculated offline lowering the query-time cost, and is based on the entire Web graph, which is more resistant to link spam. In contrast, HITS is based on sub Web graph and can only be calculated at the run time (Bharat and Mihaila, 2001).

Rafiei and Mendelzon (2000) proposed two methods to calculate the reputation of a page on certain topics. The first method is based on a one-level weight propagation PageRank model, which indicates that a page can acquire a high reputation on a topic because it is pointed to by many pages or high-reputation pages on the topic. The second method is based on a two-level weight propagation HITS model, where a page is deemed an authority if it is pointed to by good hubs on the topic of interest, where a good hub points to good authorities. To optimize the computational cost, the approach is approximated in a practical way by collecting all terms appearing on a page *p*, looking at all incoming links of a page *p*, and collecting all possible terms from those pages using a breadth-first search. Since the calculation is done at run-time, it is not scalable to the entire Web. Richardson and Domingos (2002) achieved reasonably enhanced search ranking by generating a PageRank vector for each possible query term. They proposed a directed surfer model in which an intelligent surfer jumps from page to page based on the content of the page and his query terms. The probability of a jump is based on the query-dependent PageRank score, which must be calculated at run-time. This method requires considerable processing time and storage and is unlikely to scale.

Making PageRank topic-related can avoid the situation where heavily-linked pages get high ranks regardless of whether a topic appears in the content. Havelilwala (2002) proposed a topic-sensitive PageRank to generate query-specific PageRank scores for Web pages regarding different search queries. The results showed that the topic-sensitive PageRank generated more accurate rankings for a given query than the normal PageRank algorithm. Basically, sixteen categories were selected from the Open Directory Project (ODP), where each category contained a list of URLs. For each category, the weighted vector was formed based on whether a given URL in the network belonged to the list of URLs for a selected category. At the query time, the similarity of the query to each of these categories was calculated, and topic-sensitive PageRank vectors were weighted based on similarity. Topic-sensitive PageRank for a page *i* on a topic $z_j$can be defined as:

$$TSPR_z(i) = \lambda \sum_{j:j\to i} \frac{TSPR_z(j)}{O(j)} + \begin{cases} (1-\lambda)\dfrac{1}{|\tau_z|}, & if\ j \in \tau_z \\ 0, & if\ j\ \notin \tau_z \end{cases}$$

While Havelilwala added the topic-specific vector as a personalized vector to the random jump part of the PageRank formula, Pal and Narayan (2005) added the topical differences of nodes in the network part of the PageRank formula. In Pal and Narayan's model, a surfer on a topic *a* favors links leading to pages on the same topic and has a smaller probability of visiting the non-related topic pages. Pal and Narayan's

approach does not consider the random jump part of the PageRank algorithm. Table A6 in Appendix lists notations for the various PageRank formulas mentioned in this subsection.

Richardson and Domingos (2002) proposed an intelligent surfer model to add topic to both parts of the PageRank algorithm. The intelligent surfer model is a query-specific version of PageRank in which the surfer is following links or jumping to other links based on the relevance of links to the query. For a specific query *q*, page *i*'s query-dependent PageRank score is

$$IS_q(i) = (1-\lambda)\frac{r(q,i)}{\sum_{k\in W} r(q,k)} + \lambda \sum_{j:j\to i} IS_q(j)\frac{r(q,j)}{\sum_{l:j\to l} r(q,l)}$$

Similar to HITS, the intelligent surfer algorithm faces the scalability issue of calculating query-specific PageRank at run-time.

Nie, Davison and Qi (2006) proposed a topic model that combined PageRank and HITS without affecting the overall authority score yet still providing a global ranking that could be interpreted in a query- or topic-specific manner. For the random walk, the topical surfer faces three choices: 1) jumping to a random page with a random topic; 2) following a hyperlink to stay in the same topic; or 3) following a hyperlink to jump to another topic. The modified PageRank formula can be shown as

$$A(i) = \lambda\alpha \sum_{j:j\to i} \frac{A(j)}{O(j)} + \lambda(1-\alpha) \sum_{j:j\to i} \frac{C(j)}{O(j)} \sum_{k\in T} A(k) + (1-\lambda)\frac{1}{N} C(i) \sum_{j:j\to i} \sum_{k\in T} A(k)$$

where *C* is the content vector which is a probability distribution representing the content of a node *i*. Their topical random surfer model is similar to a random surfer model of the traditional PageRank algorithm.

Yang, Tang, Zhang and Li (2009) applied LDA to calculate topic distribution for each document. They proposed a topic-level random walk, in which the surfer not only randomly jumps to new pages related to the search topic but also follows the links on the visited pages that are highly related to the search topic. This process adds the topical level to the two parts of the PageRank algorithm. Their weighted PageRank formula is:

$$r[d,z_i] = (1-\lambda)\frac{1}{|D|}P(z_i|d) + \lambda \sum_{d':d'\to d} \left[(1-\lambda)P(d|d',z_i) + \lambda \frac{1}{T} \sum_{j\neq i} P(d,z_i|d',z_j) \right]$$

where λ is the damping factor and *P(z|d)* is the probability of a topic *z* being generated by a document *d*. This method is limited to topic distribution at the document level. In contrast, the present paper addresses this issue at the author level.

It is obvious that most topic-based PageRank algorithms have been achieved by either pre-computing over the entire Web (Haveliwala, 2002) or ranking the subset of neighborhood graphs containing the query words (Kleinberg, 1999). These topics are either pre-defined (Haveliwala, 2002) or limited to a subset of popular pages (Jeh & Widom, 2003). The exception is Yang et al. (2009) who applied LDA to

compute the topic distribution for documents, but not for authors. This paper applied the extended LDA to calculate the topic distributions for authors and added them to the weighted PageRank algorithm.

Methodology

Information retrieval (IR) was selected as the test field. Papers and their citations were collected from the Web of Science (WOS) covering the period from 1956 to 2008. In total, 15,367 papers with 350,750 citations were collected. Citation records contained the first author, year, source, volume, and page number. The entire dataset was divided into four time phases: 1956-1980 (Phase 1), 1981-1990 (Phase 2), 1991-2000 (Phase 3), and 2001-2008 (Phase 4). Details of data collection and the dataset itself are provided in Ding and Cronin (2010 forthcoming). Table 1 provides an overview of the IR dataset.

Table 1: Overview of the IR dataset

	Phase 1 (1956-1980)	Phase 2 (1981-1990)	Phase 3 (1991-2000)	Phase 4 (2001-2008)	Total
No. of Words	1,997	2,151	6,500	9,506	20,154
No. of Papers	1,313	1,173	4,485	8,396	15,367
No. of Authors	1,567	1,485	8,117	14,593	25,762
No. of Citations	10,862	17,874	110,454	211,560	350,750

Note: No. of Words is the number of unique 1-gram words extracted from paper titles excluding stop words.

The PageRank and weighted PageRank were calculated based on author co-citation networks. An author co-citation network is an undirected and weighted graph in which nodes represent authors, edges represent co-citation relationships among authors, and edge weights represent co-citation frequencies among the authors. The network is based on the following assumptions: 1) author co-citation networks count the number of times an author is cited in the form of co-citation frequency, so that highly cited authors have higher scores in co-citation networks; 2) if a paper of author A is cited together with a paper by a highly cited author B, author A may conduct studies that are relevant to author B, and the researches of author A and author B are important to the citing paper. For each phase, the top 100 highly cited authors were chosen, and the author co-citation networks were formed based on the entire citation dataset for each time phase. The original PageRank and topic-based PageRank were calculated based on author co-citation networks with a damping factor of 0.85, 0.5, and 0.15.

Topic Modeling

The Author-Conference-Topic (ACT) model proposed by Tang, Jin and Zhang (2008) was applied to capture both document content and author interests. The output of this algorithm was five extracted topics, each with an associated list of words and authors ranked by their topic distribution probabilities. The ACT model calculated each author's interest distribution (i.e., the probability of a topic for a given author: $P(t|a)$) across the five extracted topics. For example, author A has a probability of 0.2 for topic 1, 0.1 for topic 2, 0.1 for topic 3, 0.4 for topic 4, and 0.2 for topic 5. For author A, the sum of topic distribution across the five topics is 1.0. The average probability for author A for these five topics would be 1/5=0.2. The model also calculates the author's probability distribution for a given topic (i.e., $P(a|t)=P(a)P(t|a)/P(t)$). For example, there are 14,593 authors in Phase 4, and each author will have a topic distribution on topic 1. The sum of the total author-topic distribution over topic 1 would be 1.0. The average probability for each author-topic distribution over topic 1 would be 1/14593, or 0.0000685.

Topic-based PageRank I: Simple Combination of LDA and PageRank

PageRank and LDA were calculated separately. The original PageRank with damping factor as 0.85 was calculated based on the author co-citation networks, and the topic distribution was calculated using the ACT model based on publications containing titles and authors. The corresponding PageRank scores (denoted as PR) and topic distributions for each five topics (denoted as I) for the top 100 highly cited authors were selected. Formula 1 shows the simple combination of LDA and PageRank (called Topic-based PageRank I, denoted as I_PR) where $\bar{I}$ represents the average of I and $\overline{PR}$ represents the average of PageRank:

$$I\_PR = \left(\frac{I-\bar{I}}{\bar{I}}\right) * \left(\frac{PR-\overline{PR}}{\overline{PR}}\right) \quad (1)$$

For each phase, I_PR is ranked based on the condition that I should be more than $\bar{I}$.

Topic-based PageRank II: Topic-based Random Walk

A topical random surfer model was proposed in which a surfer has d probability of following the links on current pages or $(1-d)\alpha$ probability of jumping to a new page, where α is the topic distribution of the new page. The topic-based PageRank is represented as

$$PR\_t(i) = (1-d)\frac{t(i)}{\sum_{i=1}^{N} t(i)} + d\sum_{j:j\to i}\frac{PR\_t(j)}{O(j)} \quad (2)$$

where $p_1, p_2, \ldots, p_n$ are the nodes in the network and N is the total number of nodes, $O(j)$ is the number of out-going links on node p_j, $PR\_t(i)$ is the topic-based PageRank on node p_i, and $PR\_t(j)$ is the topic-based PageRank on node p_j.). $t(i)$ is the conditional probability distribution of an author for a given topic ($P(a|t)$). The damping factor d is the probability that a random surfer will follow one of the links on the current page. The damping factor was set to 0.15 (to stress the equal chance of being cited), 0.5 (to indicate that scientific papers usually follow a short path of 2), or 0.85 (to stress the network topology) (Chen, Xie, Maslov, & Redner, 2007). The weighted vector is normalized by the sum of the topic distribution of all nodes, which is $\sum_{i=1}^{N} t(i)$.

Results and Discussion

Topics

Five topics were extracted for each phase (see Table 2). Details about the top 10 ranked words associated with each topic were provided in Table A1-A4 in Appendix. IR research in 1956-1980 focused on data storage, classification, medical and chemical IR, and online IR. Among these, the topic of Classification and Patent IR was most popular. In the period of 1981-1990, Query Processing and Database were the emerging topics, and the topic of Evaluation shifted from system, document, and storage evaluation to search/user/expert evaluation. Among the five topics in the period of 1981-1990, Online IR was the most popular topic. The launch of the WWW in the early 1990s had a strong impact on IR research: Web IR and Multimedia IR emerged during this time period. The focus of interest in databases shifted from relational to object-oriented databases. Similar topics have been identified by Sugimoto and McCain (2010). They found three major topics in IR during the period of 1980-1984: Information Retrieval Systems, Database Management Systems, and Information Storage. In the period of 1991-2000, Evaluation was the most popular topic. Evaluation later became a compulsory part of most IR researches

and disappeared in the period of 2001-2008. Other topics like Medical IR, Multimedia IR, and Database (e.g., object-oriented database) have also been detected by Sugimoto and McCain (2010). It is interesting to see that, in the period of 2001-2008, IR Theory and Model appeared as one of the five extracted topics, in line with the need to adjust traditional IR models and theories to the new web or social web settings. In 2001-2008, the topic of Database and Query Processing was most popular. In the study of Sugimoto and McCain (2010), Internet was identified as the most salient topic and the topic of Database Management Systems was less dominant during the period of 2000-2004. They also found that the topic of Information Retrieval Systems became peripheral and the topic of Digital Library acted as cut points to connect other topics. In general, similar topics were identified by Sugimoto and McCain (2010) for the latest three periods. Sugimoto and McCain (2010) were able to differentiate the hub and periphery of these topics, while this paper was capable to distinguish the popularity of these topics based on the Mean θ of the ACT model.

Table 2: Five extracted topics for each phase

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1956-1980	Thesaurus and Chemical IR	Data Storage and Evaluation	Online IR	Medical IR	**Classification and Patent**
1981-1990	Automatic IR System	**Online IR**	Digital Library	Database and Query Processing	Evaluation
1991-2000	Web IR	Multimedia IR	**Evaluation**	Medical IR	Database and Query Processing
2001-2008	Multimedia IR	**Database and Query Processing**	Medical IR	Web IR and Digital Library	IR Theory and Model

Note: The most popular topic during each phase is highlighted in bold.

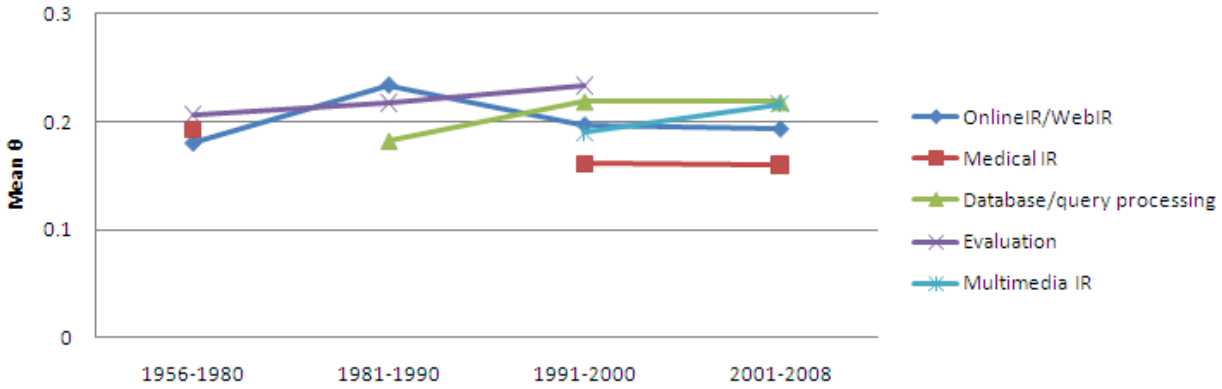

Figure 2: Dynamic changes of topics

Figure 2 shows the dynamic changes of topics across four phases. The Y axis in Figure 2 is the Mean θ which represents the average probability of a topic for a given word. The higher the Mean θ is, the more popular the topic is. Online IR/Web IR remained popular across all four phases. IR researchers showed great interest in Medical IR in Phase 1, then renewed their interest in Phases 3 and 4. Database was a popular topic in Phases 2, 3 and 4. Evaluation was popular in Phases 2 and 3, but declined in Phase 4. Multimedia IR appeared as a popular topic in Phase 3 and its popularity increased in Phase 4 due to the social web effect of sharing photos and videos.

Table 3 shows the shifting focus of each topic in each of the four phases. Online IR became Web IR from Phase 1 to Phase 4, and Database and Query Processing reflected the development of databases: shifting from relational databases (Phase 2), to object-oriented databases (Phase 3), and to XML databases (Phase 4). Evaluation changed from evaluation of storage/system/models to evaluation of online/hypertext/digital library systems. Medical IR developed from an emphasis on storage and systems in Phase 1, to patient management in Phase 3, and health care in Phase 4. And Multimedia IR changed from syntactic analysis in Phase 3 to semantic analysis in Phase 4.

Table 3. Top 10 words associated with each topic in different phases

	1956-1980	**1981-1990**	**1991-2000**	**2001-2008**
Online IR/Web IR	system, online, language, theory, query, computerized, thesaurus, evaluation, semantic, bibliography	online, systems, text, concepts, reference, principles, proceedings, practice, knowledge, services	system, web, knowledge, database, data, query, design, text, management, distributed	web, search, digital, searching, knowledge, system, query, user, model, internet
Database and Query Processing		query, language, query-processing, database, relational, system, distributed, data, database-system, comparison	query, database, databases, data, object-oriented, queries, processing, relational, model, language	query, data, xml, processing, queries, databases, database, efficient, web, querying
Evaluation	system, document, storage, evaluation, data, automatic, model, relevance, indexing, online	systems, document, full-text, model, evaluation, fuzzy, effectiveness, search, user, expert	text, evaluation, systems, searching, search, online, relevance, library, user, hypertext	
Medical IR	system, data, storage, computerized, chemical, medical, literature, biomedical, evaluation, management		database, medical, system, clinical, patient, management, health, identification, automated, optical	database, medical, health, clinical, management, search, design, study, support, knowledge
Multimedia IR			image, content-based, system, indexing, databases, multimedia, images, visual, video, color	image, content-based, learning, images, relevance, color, feedback, video, semantic, similarity

Authors

Author rank for each topic generated by the ACT model is based on how many words an author contributes to the topic. Table 4 shows the top 10 authors associated with topics during each of the four phases. There is overlap between authors in the Online IR/Web IR and Evaluation topics and in the Medical IR and Multimedia IR topics. Some of the authors are prestigious award winners: ASIST Award of Merit (i.e., C. A. Lynch, G. Salton), ASIST Best Book Award (i.e., C. L. Borgman), Gerard Salton Award (i.e. G. Salton, W.S. Cooper), and ASIST Research Award (i.e., W.B. Croft).

Table 4. Top 10 authors associated with each topic in different phases

	1956-1980	1981-1990	1991-2000	2001-2008
Online IR/Web IR	D.T. Hawkins, N.A. Stokolova, E. Eisenbach, K. Yamanaka, T. Radecki, R. Fugmann, J. Eyre, D.H. Kraft, Z. Mazur, K. Hosono	P. Willett, S.P. Harter, C. Batt, D. Ellis, M. Keen, S.E. Hocker, L. Bronars, P.G. Enser, S. Stigleman, B. Vickery	W.B. Corft, H.C. Chen, W. Umstatter, C.A. Lynch, P. Martin, D. Samson, N.J. Santora, C. Womserhacker, N.J. Belkin, R. Wagnerdobler	M. Thelwall, C.C. Yang, A. Spink, P. Jacso, I. Fourie, H.C. Chen, N. Ford, H. Xie, G.G. Chowdhury, B. Hjorland
Database		D.W. Stemple, R.H. Guting,	J. Han, D. Suciu, H.P.	J.Z. Li, F. Bry, H.J. Kim, D.

and Query Processing		A. Sernadas, C. Katzeff, S.Y. Su, W. Perrizo, J.S. Davis, C.T. Yu, B.S. Goldshteyn, I.A. Macleod	Kriegel, S.Y. Su, K.L. Tan, G. Graefe, L. Wong, L. Libkin, J.W. Su, P.Z. Revesz	Papadias, K. Subieta, J. Van den Bussche, D. Taniar, F. Geerts, M. Song, Y.D. Chung
Evaluation	G. Salton, A.G. Pickford, W. Goffman, E. Garfield, G.K. Thompson, W.S. Cooper, K. Janda, F.W. Lancaster, R. Fugmann, P. Willett	C.L. Borgman, T. Radecki, G. Salton, W.B. Croft, J.S. Ro, J. Panyr, D.C. Blair, M.E. Maron, P. Thompson, C.A. Lynch	A. Spink, R.M. Losee, E. Levine, C. Cole, P. Willett, W.R. Hersh, C.T. Meadow, B. Hjorland, E. Garfield, T. Cawkell	
Medical IR	S.J. Martinez, M.G. Manzone, C.M. Bowman, F.A. Landee, J. Frome, I. Berghans, S.L. Visser, H. Skolnik, Y.J. Lee, T.K.S. Engar		S.G. Aiken, I. Soutar, S. Barcza, C.C. Tsai, W. Hersh, S.J. Westerman, H.H. Emurian, L.L. Consaul, H.J. Markowitsch, D. Roberts	R.N. Kostoff, U.J. Balis, G. Eysenbach, R.B. Haynes, G. Nilsson, H. Shatkay, N.L. Wilczynski, C.R. Shyu, J.I. Westbrook, G.O. Babnett
Multimedia IR			H.C. Chen, F. Crestani, A.K. Jain, E. Wilhelm, J.I. Khan, B.S. Manjunath, H.K. Kim, H.M. Wang, S.F. Chang, S. Levialdi	T.S. Huang, H.J. Zhang, G.J. Lu, J. Li, C.C. Chang, E. Izquierdo, J. Lassksonen, H. Burkhardt, C.J. Liu, D. Ziou

Topic-based PageRank for Authors

Using the proposed topic-based PageRank algorithms, it was possible to provide topic-dependent rankings for authors. Across the four phases, I_PR provided interesting rankings for authors who were not only highly cited but also highly productive for a given topic. The PR_t ranks were highly correlated with standard PageRanks (see Appendix).

Table 5. Top 10 authors based on topic-based PageRank for different topics

Online IR/Web IR	I_PR	PR_t(.85)	PR_t(.5)	PR_t(.15)
1956-1980	D.T. Hawkins, N.A. Stokolova, R.K. Summit, M.E. Williams, T. Radecki, A. Macleodi, T. Saracevic, R.S. Marcus, R. Fugmann, C.T. Yu	G. Salton, A. Kent, M.E. Williams, F.W. Lancaseter, R.K. Summit, D.T. Hawkins, C.W. Cleverdon, D.B. Mccarn, W.S. Cooper, H. Martint, C.P. Bourne	G. Salton, D.T. Hawkins, M.E. Williams, R.K. Summit, F.W. Lancaster, A. Kent, N.A. Stokolova, R. Fugmann, C.W. Cleverdon, W.S. Cooper	D.T. Hawkins, N.A. Stokolova, R. Fugmann, T. Radecki, G. Salton, R.K. Summit, I.A. Macleod, J. Farradane, M.E. Williams, A.M. Rees
1981-1990	S.E. Robertson, D. Ellis, P. Willett, P. Ingwersen, B.C. Vickery, A.S. Pollitt, D.H. Kraft, H.M. Brooks, A.F. Smeaton, E.A. Fox	G. Salton, A. Bookstein, S.E. Robertson, T. Radecki, W.B. Croft, C.J. Vanrijsbergen, C.T. Yu, W.S. Coopwer, P. Willett, K.S. Jones	G. Salton, P. Willett, S.E. Robertson, A. Bookstein, S.P. Harter, W.B. Croft, T. Radecki, C.J. Vanrijsbergen, C.T. Yu, D. Ellis	P. Willett, S.P. Harter, D. Ellis, S.E. Robertson, G. Salton, A.F. Smeaton, P. Ingwersen, B.C. Vickery, M.J. Bates, A. Bookstein
1991-2000	N.J. Belkin, W.B. Frakes, T. Imielinski, G.W. Furnas, T. Catarci, T. Kohonen, R. Agrawal, S.K. Chang, H.C. Chen, P. Valduriez	G. Salton, N.J. Belkin, S.E. Robertson, S. Abiteboul, T. Saracevic, C.J. Vanrijsbergen, W.B. Croft, M.J. Bates, K.S. Jones, D. Harman	G. Salton, N.J. Belkin, S. Abiteboul, S.E. Robertson, S.K. Chang, T. Saracevic, H.C. Chen, C.J. Vanrijsbergen, W.B. Croft, M.J. Bates	H.C. Chen, N.J. Belkin, G. Salton, S.K. Chang, N. Fuhr, T. Saracevic, S.K.M. wong, M.J. Bates, S. Abiteboul, T. Catarci
2001-2008	A. Spink, T. Saracevic, B. Hjorland, S.E. Roberston, B.J. Jansen, N.J. Belkin, E.M. Voorhees, W.R. Hersh, P. Ingwersen, P. Vakkari	G. Salton, A. Spink, N.J. Belkin, T. Saracevic, S.E. Roberston, Y. Rui, E.M. Voorhees, B.J. Jansen, J.R. Smith, K.S. Jones	A. Spink, T. Saracevic, G. Salton, H.C. Chen, B.J. Jansen, B. Hjorland, N.J. Belkin, S.E. Robertson, P. Vakkari, E.M. Voorhees	A. Spink, H.C. Chen, B. Hjorland, T. Saracevic, B.J. Jansen, P. Vakkari, P. Borlund, S.E. Robertson, F. Crestani, N.J. Belkin
Database and Query Processing	I_PR	PR_t(.85)	PR_t(.5)	PR_t(.15)
1981-1990	C.T. Yu, C.J. Vanrijsbergen, A.L.P. Chen, G. Ozsoyoglu, S.K. Chang, A. Klug, P. Reisner, R. Snodgrass, N. Goodman, S.K.M.Wong	G. Salton, A. Bookstein, S.E. Roberston, T. Radecki, W.B. Croft, C.T. Yu, C.J. Vanrijsbergen, W.S. Cooper, K.S. Jones, D.A. Buell	G. Salton, C.T. Yu, A. Bookstein, T. Radecki, S.E. Robertson, W.B. Croft, C.J. Vanrijsbergen, D.A. Buell, G. Ozsoyoglu, A. Klug	G. Salton, C.T. Yu, G. Ozsoyoglu, I.A. Macleod, A.L.P. Chen, A. Klug, C.J. Vanrijsbergen, T. Radecki, D.D. Chamberlin, W.B. Croft
1991-2000	S. Abiteboul, W. Litwin, J. Paredaens, M.J. Egenhofer, S. Grumbach, C.J. Vanrijsbergen, S.Y. Lee, M. Kifer, R. Snodgrass, H.	G. Salton, S. Abiteboul, N.J. Belkin, S.E. Robertson, T. Saracevic, C.J. Vanrijsbergen, W.B. Croft, M.J. Bates, M. Stonbraker, E.F. Codd	G. Salton, S. Abiteboul, N.J. Belkin, S.E. Robertson, S.K. Chang, E. Bertino, J. Paredaens, R. Snodgrass, C.J. Vanrijsbergen, M.	S. Abiteboul, J. Paredaens, E. Bertino, R. Snodgrass, G. Salton, M.J. Egenhofer, A.U. Tansel, S. Grumbach, S.Y. Lee, M. Gyssens

	Samet		Stonebraker	
2000-2008	H.V. Jagadish, D. Calvanese, M.J. Egenhofer, G. Gottlob, S. Abiteboul, G. Graefe, W.B. Frakes, L. Gravano, R. Baezayates, M. Fernandez	G. Salton, Y. Rui, S.E. Robertson, A. Spink, N.J. Belkin, J.R. Smith, S. Abiteboul, T. Saracevic, E.M. Voorhess, D. Harman	G. Salton, S. Abiteboul, H.V. Jagadish, Y. Rui, J.R. Smith, S.E. Robertson, A. Gupta, G. Gottlob, N.J. Belkin, A. Spink	H.V. Jagadish, G. Gottlob, D. Calvanese, A. Gupta, S. Abiteboul, S. Chaudhuri, M.J. Egenhofer, W.B. Frakes, L. Gravano, M. Fernandez
Evaluation	I_PR	PR_t(.85)	PR_t(.5)	PR_t(.15)
1956-1980	G. Salton, K. Janda, W. Goffman, C.J. Vanrijsbergen, C.W. Cleverdon, F.W. Lancaster, E. Garfield, C.N. Mooers, D.B. Mccarn, C.T. Yu	G. Salton, A. Kent, M.E. Williams, F.W. Lancaster, R.K. Summit, C.W. Cleverdon, D.T. Hawkins, W.S. Cooper, D.B. Mccarn, H. Martint, C.J. Vanrijsbergen	G. Salton, A. Kent, M.E. Williams, F.W. Lancaster, R.K. Summit, C.W. Cleverdon, D.T. Hawkins, W.S. Cooper, D.B. Mccarn, H. Martint	G. Salton, W. Goffman, E. Garfield, W.S. Cooper, C. J. Vanrijsbergen, K. Janda, B.C. Vickery, R. Fugmann, A.M. Rees, C.N. Mooers
1981-1990	G. Salton, T. Radecki, A. Bookstein, W.B. Croft, D.A. Buell, M.E. Maron, S. Miyamoto, R.G. Crawford, D.C. Blair, G.P. Zarri	G. Salton, A. Bookstein, T. Radecki, S.E. Roberston, W.B. Croft, C.J. Vanrijsbergen, C.T. Yu, W.S. Cooper, K.S. Jones, D.A. Buell	G. Salton, T. Radecki, A. Bookstein, W.B. Croft, C.L. Borgman, S.E. Robertson, C.J. Vanrijsbergen, M.E. Maron, D.A. Buell, C.T. Yu	G. Salton, T. Radecki, C.L. Borgman, W.B. Croft, A. Bookstein, M.E. Maron, J. Panyr, D.C. Blair, D.A. Buell, R.G. Crawford
1991-2000	G. Salton, T. Saracevic, S.E. Robertson, M.J. Bates, P. Willett, A. Spink, C.L. Borgman, R. Fidel, K.S. Jones, C. Stanfill	G. Salton, N.J. Belkin, S.E. Robertson, T. Saracevic, C.J. Vanrijsbergen, M.J. Bates, W.B. Croft, S. Abiteboul, A. Spink, K.S. Jones	G. Salton, A. Spink, R.M. Losee, T. Saracevic, S.E. Robertson, N.J. Belkin, M.J. Bates, C.L. Borgman, C.J. Vanrijsbergen, R. Fidel	A. Spink, R.M. Losee, T. Saracevic, C.L. Borgman, P. Willett, M.J. Bates, S.E. Robertson, R. Fidel, G. Salton S.P. Harter
Medical IR	I_PR	PR_t(.85)	PR_t(.5)	PR_t(.15)
1956-1980	A. Kent, J. Frome, R.K. Summit, R.S. Ledley, L.A. Hollaar, A. Fairthorner, V.E. Giuliano, D.J. Hillman, H.M. Kissman, J. Oconnor	G. Salton, A. Kent, M.E. Williams, F.W. Lancaster, R.K. Summit, C.W. Cleverdon, D.T. Hawkins, D.B. Mccarn, J. Frome, W.S. Cooper, H. Martint	G. Salton, A. Kent, F.W. Lancaster, R.K. Summit, M.E. Williams, J. Frome, C.W. Cleverdon, D.T. Hawkins, W.S. Cooper, D.B. Mccarn	J. Frome, A. Kent, R.K. Summit, G. Salton, R.S. Ledley, H.M. Kissman, F.W. Lancaster, R.A. Fairthorne, M.E. Williams, D.T. Hawkins
1991-2000	E. Garfield, W.R. Hersh, S. Ceri, T. Bernerslee, D.R. Swanson, T. Kohonen, W. Kim, J.P. Callan, R. Reiter, P. Buneman	G. Salton, N.J. Belkin, S. Abiteboul, S.E. Robertson, T. Saracevic, C.J. Vanrijsbergen, W.B. Croft, M.J. Bates, K.S. Jones, D. Harman	G. Salton, S. Abiteboul, N.J. Belkin, S.E. Robertson, C.J. Vanrijsbergen, T. Saracevic, S.K. Chang, W.B. Croft, M.J. Bates, M. Stonebraker	G. Salton, S. Abiteboul, W.R. Hersh, E. Garfield, S. Ceri, D.R. Swanson, N.J. Belkin, S.E. Robertson, A.K. Chandra, S.K. Chang
2001-2008	R.N. Kostoff, C. Buckley, S. Berchtold, M.W. Berry, S.F. Chang, R. Fagin, H. Muller, D. Florescu, D.R. Swanson, A.K. Jain	G. Salton, Y. Rui, S.E. Robertson, A. Spink, N.J. Belkin, J.R. Smith, T. Saracevic, E.M. Voorhees, D. Harman, K.S. Jones	G. Salton, R.N. Kostoff, Y. Rui, S.E. Robertson, N.J. Belkin, A. Spink, D.R. Swanson, J.R. Smith, T. Saracevic, S. Abiteboul	R.N. Kostoff, H. Muller, W. Hersh, J. Li, Y. Rui, D.R. Swanson, C. Buckley, S.E. Robertson, N.J. Belkin, W.R. Hersh
Multimedia IR	I_PR	PR_t(.85)	PR_t(.5)	PR_t(.15)
1991-2000	W.B. Croft, A.K. Jain, S.K. Chang, J.K. Wu, S.Y. Lee, W.S. Cooper, H.C. Chen, A. Pentland, A. Gupta, A. Delbimbo	G. Salton, N.J. Belkin, S.E. Robertson, S. Abiteboul, T. Saracevic, C.J. Vanrijsbergen, W.B. Croft, M.J. Bates, S.K. Chang, K.S. Jones	G. Salton, S.K. Chang, S. Abiteboul, N.J. Belkin, S.E. Robertson, W.B. Croft, H.C. Chen, A.K. Jain, C.J. Vanrijsbergen, T. Saracevic	H.C. Chen, A.K. Jain, S.K. Chang, G. Salton, W.B. Croft, J.K. Wu, S.Y. Lee, N. Fuhr, A. Pentland, E.M. Voorhees
2001-2008	Y. Rui, J.P. Eakins, J. Li, S. Chaudhuri, J.R. Smith, A.W.M. Smeulders, J. Han, T. Gevers, N. Vasconcelos, R.M. Haralick	G. Salton, Y. Rui, J.R. Smith, S.E. Robertson, A. Spink, N.J. Belkin, T. Saracevic, E.M. Voorhees, A.W.M. Smeulders, R. Baezayates	Y. Rui, J. Li, G. Salton, J.R. Smith, J.Z. Wang, N. Vasconcelos, A.W.M. Smeulders, J.P. Eakins, W.Y. Ma, S. Chaudhuri	J. Li, N. Vasconcelos, J.Z. Wang, J.P. Eakins, S. Chaudhuri, Y. Rui, W.Y. Ma, T. Gevers, A.W.M. Smeulders, J.R. Smith

Table 5 shows the top 10 authors based on four topic-based PageRank scores for different topics and in different time phases. I_PR and PR_t(.15) tended to provide similar rankings that were different from PR_t(.85) and PR_t(.5). Interestingly, G. Salton was generally ranked top regardless of topics or phases based on PR_t(.85) and PR_t(.5). This is because PR_t(.85) and PR_t(.5) stress network topology and G. Salton is an important network node with a high degree centrality. In this case, I_PR and PR_t(.15) provided better topic-sensitive ranks than PR_t(.85) and PR_t(.5). For example, top ranked authors based on I_PR and PR_t(.15) are diverse and are not dominated by G. Salton: Online IR/Web IR (i.e., D. T. Hawkins in Phase 1, S. E. Robertson and P. Willet in Phase 2, N. J. Belkin and H. C. Chen in Phase 3, A.

Spink in Phase 4), Database and Query Processing (i.e., C. T. Yu and G. Salton in Phase 2, S. Abiteboul in Phase 3, H. V. Jagadish in Phase 4), Evaluation (i.e., G. Salton in all phases, A. Spink in Phase 3), Medical IR (i.e., A. Kent and J. Frome in Phase 1, E. Garfield and G. Salton in Phase 3, R. N. Kostoff in Phase 4), and Multimedia IR (i.e., W. B. Croft and H. C. Chen in Phase 3, Y. Rui in Phase 4).

For the topic-based rankings of the top 100 highly cited authors, the Spearman correlation test (2-tailed) was conducted for each phase to identify correlations among the four different topic-based PageRanks (i.e., I_PR, PR_t(.85), PR_t(.5), and PR_t(.15)). All author rankings of different topics in different phases were highly correlated at a confidence level of 0.01 or 0.05. In order to further understand the correlation among these rankings, Phase 4 (2001-2008) was chosen as an example. Principal Component Analysis (PCA) rotated by Varimax with Kasier normalization was used to extract three major components based on different author rankings in Phase 4. In Table 6, representative variables for each component are highlighted in bold if their absolute loadings are more than 0.4 (Raubenheimer, 2004). Based on these identified representative variables, three components were extracted: Component 1 grouped various PR_t measures, Component 2 highlighted various I_PR measures, and Component 3 showed the various PR_t measures for the first three topics (i.e., Multimedia IR, Database and Query Processing, and Medical IR). Figure 3 displays the projection of three components for each measure. These three components explained 84.37% of the total variance. The result of PCA indicates that the topic-based measures PR_t and I_PR provide different rankings: I_PR provides meaningful ranks by considering both publications and citations, while PR_t ranks are dominated or skewed by highly cited authors.

Table 6. PCA for topic-based PageRanks during the period of 2001-2008

	Component		
	1	2	3
PR_t1_0.85	**.685**	-.089	**.646**
PR_t1_0.5	**.441**	.231	**.801**
PR_t1_0.15	.190	.355	**.809**
I_PR_t1	.140	**.859**	.237
PR_t2_0.85	**.748**	-.024	**.575**
PR_t2_0.5	**.415**	**.439**	**.661**
PR_t2_0.15	.207	**.494**	**.646**
I_PR_t2	.041	**.891**	.244
PR_t3_0.85	**.863**	-.040	**.410**
PR_t3_0.5	**.755**	.308	**.456**
PR_t3_0.15	**.643**	**.416**	.374
I_PR_t3	.204	**.801**	.099
PR_t4_0.85	**.925**	-.036	.249
PR_t4_0.5	**.871**	.339	.188
PR_t4_0.15	**.780**	**.484**	.036
I_PR_t4	.133	**.894**	.189
PR_t5_0.85	**.930**	-.006	.245

PR_t5_0.5	**.850**	.386	.169
PR_t5_0.15	**.718**	**.514**	.079
I_PR_t5	.089	**.918**	.118

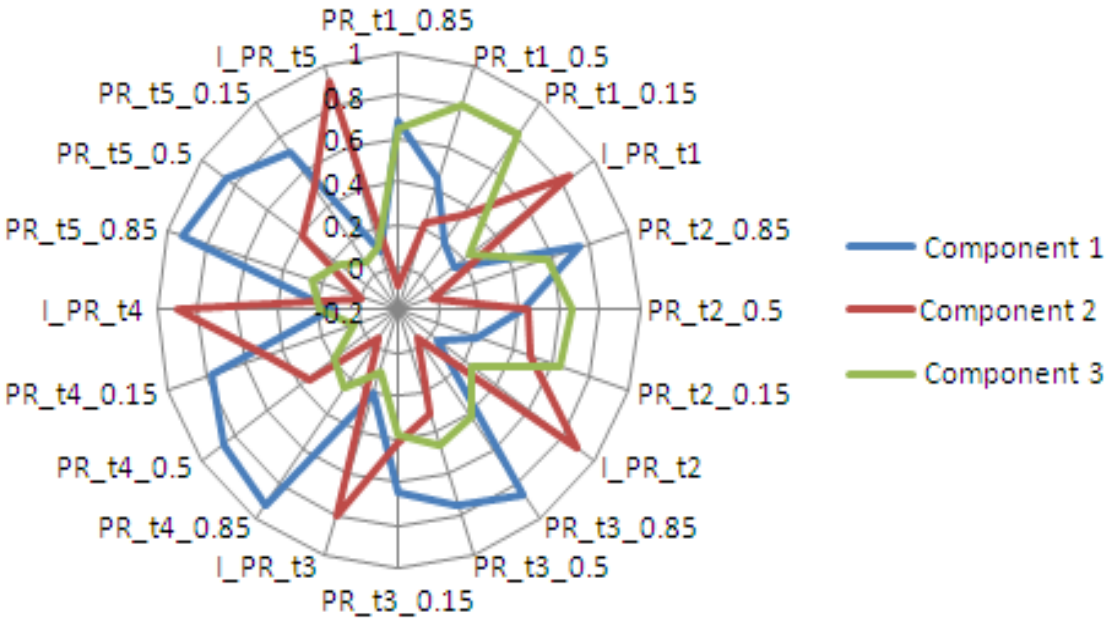

Figure 3. PCA for topic-based PageRanks during the period of 2001-2008

Comparing Topic-based PageRank with Other Measures

Since Phase 4 contains the largest number of publications and citations, this dataset was chosen to compare topic-based PageRanks with other measures for the top 100 highly cited authors, including popular rank (PopRank) and prestige rank (PreRank) from Ding and Cronin (2010 forthcoming), normal PageRank (PR), weighted PageRank on papers (PR_p) or citations (PR_c) (Ding, submitted), H-index, and Impact Factor (IF) index. PCA was applied and five components accounting for 84.96% of total variation were extracted. Based on the identified representative variables which were highlighted in bold (i.e., measures with absolute loading more than 0.4), the five extracted components were: Component 1 contained various PageRank, weighted PageRank and the popularity rank measures, Component 2 collected various topic-based PageRank measures (i.e., I_PR, PR_t(.15), and PR_t(.5)), Component 3 gathered weighted PageRank on paper and topic-based PageRank on Topic 3, 4 and 5, Component 4 represented the h-index measure, and Component 5 indicated the prestige rank. Figure 4 shows the projection of each component for each measure.

Table 7. PCA for 33 different measures for the 2001-2008 dataset

	Component				
	1	2	3	4	5
PR_t1_0.85	**.905**	.256	.100	-.083	.210
PR_t1_0.5	**.657**	**.616**	-.022	.001	.246

PR_t1_0.15	**.403**	**.722**	-.134	.065	.276
I_PR_t1	.002	**.886**	.144	-.163	-.044
PR_t2_0.85	**.864**	.242	.280	.145	.155
PR_t2_0.5	**.459**	**.660**	.224	.375	.186
PR_t2_0.15	.257	**.693**	.128	**.440**	.201
I_PR_t2	-.104	**.874**	.174	.081	-.051
PR_t3_0.85	**.875**	.184	.385	-.088	-.034
PR_t3_0.5	**.641**	**.511**	**.443**	-.005	-.068
PR_t3_0.15	**.490**	**.563**	**.424**	-.018	-.117
I_PR_t3	-.022	**.755**	.271	-.037	-.145
PR_t4_0.85	**.772**	.097	**.543**	-.093	.142
PR_t4_0.5	**.487**	.392	**.714**	-.025	.129
PR_t4_0.15	.280	**.446**	**.769**	-.048	.045
I_PR_t4	-.051	**.876**	.210	-.067	-.085
PR_t5_0.85	**.779**	.115	**.550**	-.048	.124
PR_t5_0.5	**.470**	**.416**	**.705**	.030	.112
PR_t5_0.15	.266	**.485**	**.704**	.022	.088
I_PR_t5	-.126	**.861**	.210	-.078	-.089
PopRank	**.893**	.022	.072	.115	.006
PreRank	.201	-.202	.116	.060	**.842**
PR_0.85	**.937**	.000	.264	-.041	.141
PR_0.5	**.958**	.007	.180	.043	.123
PR_0.15	**.948**	.010	.154	.085	.115
PR_c_0.85	**.943**	-.005	.256	-.041	.134
PR_c_0.5	**.966**	-.005	.149	.051	.093
PR_c_0.15	**.927**	.005	.099	.102	.046
PR_p_0.85	**.795**	.036	**.446**	-.260	.240
PR_p_0.5	**.517**	.200	**.450**	**-.453**	**.412**
PR_p_0.15	.388	.219	**.406**	**-.484**	**.462**
H_Index	.070	.030	-.024	**.781**	.020
IF_Rank	**.488**	.026	-.023	-.300	-.059

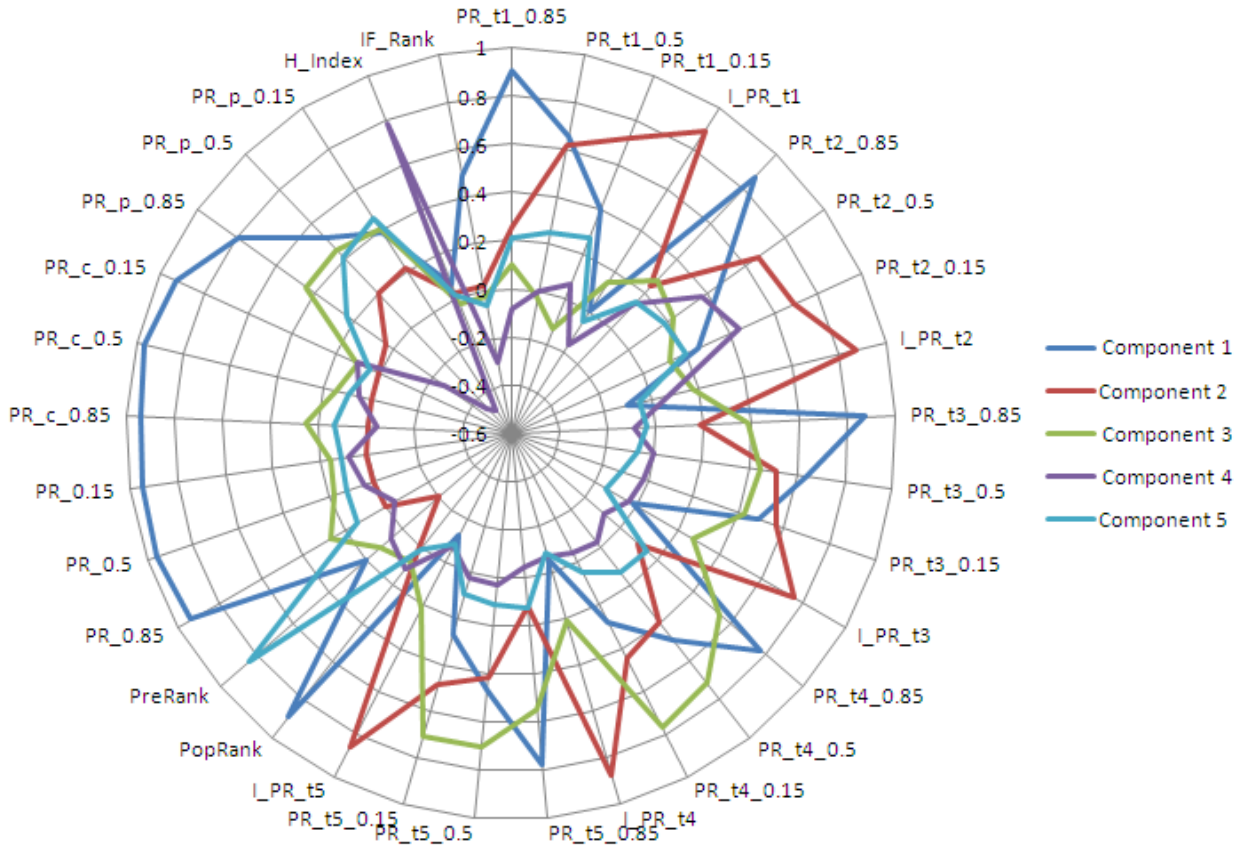

Figure 4. PCA for 33 measures for the 2001-2008 dataset

Conclusion

Ranking researchers is one of the most important topics in bibliometrics. This paper has proposed topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm. The ACT model was used to extract topics and to associate topics with individual authors. The probability of a topic for a given author formed the weighted vector for the PageRank algorithm. The topic model and PageRank algorithms were combined using either simple combination (I_PR) or taken the topic distribution as a weighted vector for PageRank (PR_t). Information retrieval was selected as the test field with 15,367 papers, 350,750 citations and 25,762 authors covering the period from 1956 to 2008. The dataset was divided into four time phases and five topics were extracted for each phase. For the top 10 authors, I_PR and PR_t(.15) provided diverse rankings that were different from those provided by PR_t(.85) and PR_t(.5) where G. Salton dominates the top rank regardless of topics or phases. For the top 100 highly cited authors, the proposed topic-based ranks were compared with other related measures and all of them were highly correlated at a confidence level of 0.01 or 0.05.

The ACT model calculated the probability distribution of author, journal, topic and document simultaneously:

- the probability of a topic for a given author P(t|a) which infers research interest of the given author, the probability of an author for a given topic P(a|t) which can derive the most productive authors for the given topic;

- the probability of a topic for a given journal P(t|j) which infers the topic focus of the given journal, the probability of a journal for a given topic P(j|t) which can derive the most productive journals for the given topic;
- the probability of a topic for a given document P(t|d) which infers the topic distribution of the given document, the probability of a document for a given topic P(d|t) which can derive the most related papers to the given topic; and
- the probability of a topic for a given word P(t|w) which infers the probability distribution of the given word for each extracted topic, the probability of a word for a given topic P(w|t) which can derive the most related words to the given topic.

These probabilities can be used as weighted vectors for the PageRank algorithm. The author will explore possibilities of utilizing these different weighted vectors to generate different diverse rankings at the topic level and to test their convergence or divergence in the future work.

Ranking papers, authors or journals in a domain is important. However, most of the current ranking algorithms cannot rank them at the topic level. An author maybe an expert in topic A, but he may not necessarily be an expert in topic B. The proposed topic-based PageRanks bring finer granularity to ranking experts under various situations by including different contextual information as weighted vectors to PageRank algorithms. For example, including an author's total publications as the weighted vector, PageRank can calculate a contextualized ranking reflecting the scholar's productivity; adding author's expertise as the weighted vector, PageRank can calculate a contextualized ranking reflecting the scholar's domain knowledge and research interest; adding author's academic genealogy as the weighted vector (Russell and Sugimoto, 2009), PageRank can calculate a contextualized ranking reflecting the scholar's educational background. In the future, the author would like to apply weighted PageRank for contextualized ranking. Furthermore, using topic modeling algorithms, semantic associations between any two given nodes in a network can be calculated based on their divergence and entropy. These identified semantic associations can be used to interpret contextualized rankings.

Acknowledgement

The author would like to thank Prof. Jie Tang from Tsinghua University for sharing the ACT code which was applied in this study, and Prof. Cassidy Sugimoto, Prof. Elin Jacob, and two anonymous reviewers for their insightful comments.

References

Bharat, K., & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p104-111, Aug 24-28, 1998, Melbourne, Australia.

Bharat, K., & MIhaila, G. A. (2001). When experts agree: Using non-affiliated experts to rank popular topics. In *Proceedings of the Tenth International World Wide Web Conference*, p597-602, May 1-5, 2001, Hongkong, China.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1033.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the seventh international conference on World Wide Web,* p107-117, April 14-18, 1998, Brisbane, Australia.

Bollen, J., Rodriguez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69(3), 669-687.

Buntine, W.L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2, 159-225.

Chakrabarti, S., Dom, B., Gibson, D., & Kleinberg, J. (1998). Automatic resroucen compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International on World Wide Web Conference*, April 14-18, 1998, Brisbane, Australia.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google. *Journal of Informetrics*, 1, 8-15.

Ding, Y (2011 forthcoming). Applying weighted PageRank on author co-citation network. *Journal of the American Society for Information Science and Technology*

Ding, Y., & Cronin, B. (2010 forthcoming). Popular and/or Prestigious? Measures of Scholarly Esteem, *Information Processing and Management* (DOI:10.1016/j.ipm.2010.01.002).

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11), 2229-2243.

Haveliwala, T. H. (2002). Topic-Sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, p517-526, May 7-11, 2002, Honolulu, Hawaii, USA.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 50-57, Aug 15-19, 1999, Berkeley, CA, USA.

Jeh, G., & Widom, J. (2003). Scaling personalized Web search. In *Proceedings of the 12th International Conference World Wide Web*, p271-279, May 20-24, 2003, Budapest, Hungary.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1), 39-43.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604-632.

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, 60(7), 1327-1336.

Lyman, P., & Varian, H.R. (2003) How Much Information? Accessed (Sept 1, 2010): http://www.sims.berkeley.edu/research/projects/how-much-info-2003

Nie, L., Davison, B.D., & Qi, X. (2006). Topical link analysis for web search. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. P91-98, Aug 6-11, 2006, Seattle, Washington, USA.

Pal, S. K., & Narayan, B. (2005). A web surfer model incorporating topic continuity. *IEEE Transactions on Knowledge and Data Engineering*, 17, 726-729.

Ponte, J. M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p275-281, Aug 24-28, 1998, Melbourne, Australia.

Rafiei, D., & Mendelzon, A. O. (2000). What is this page known for? Computing web page reputations. In *Proceedings of the Ninth International World Wide Web Conference*, May 15-19, 2000, Amsterdam, The Netherlands.

Raubenheimer, J. E. (2004). An item selection procedure to maximize scale reliability and validity. *South African Journal of Industrial Psychology*, 30(4), 59-64.

Richardson, M., & Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in PageRank. In *Proceedings of Advances in Neural Information Processing Systems*, p1441-1448, Dec 9-14, 2002, Vancouver, Canada.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, p487-494, Banff, Canada.

Russell, T., & Sugimoto, C.R. (2009). MPACT Family Trees: Quantifying academic genealogy in library and information science. *Journal of Education for Library & Information Science*, *50*(4), 248-262.

Sayyadi, H., & Getoor, L. (2009). FutureRank: Raning scientific articles by predicting their future PageRank. In *Proceedings of the 9th SIAM International Conference on Data Mining*, p533-544, April 30-May 2, 2009, Nevada, USA.

Steyvers, M., Smyth, P., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. *In Proceeding of the 10th ACM SIGKDD conference on knowledge discovery and data mining*, p306-315, Aug 22-25, 2004, Seattle, Washington, USA.

Sugimoto, C. R. & McCain, K. W. (2010). Visualizing changes over time: A history of information retrieval through the lens of descriptor tri-occurrence mapping. *Journal of Information Science*, 36(4), 481-493.

Tang, J., Jin, R., & Zhang J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of 2008 IEEE International Conference on Data Mining (ICDM2008)*, p1055-1060, Dec 15-19, 2008, Pisa, Italy.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*. p990-998, Aug 24-27, 2008, Las Vegas, USA.

Yan, E., & Ding, Y. (forthcoming). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*.

Yang, Z., Tang, J., Zhang, J., & Li, J. (2009). Topic-level random walk through probabilistic model. In *Proceedings of Joint International Conferences of APWeb/WAIM, 2009*, p162-173, April 2-4, 2009, Suzhou, China.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p334-342, September 9-13, 2001, New Orleans, LA, USA.

Appendix

Table A1: Topic and Author ranks in 1956-1980

Topic 1: Thesaurus and chemical IR		**Topic 2: Data storage and evaluation**		**Topic 3: Online IR**		**Topic 4: Medical IR**		**Topic 5: Classification and Patent**	
WORD	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**
system	0.051034	system	0.027241	system	0.040224	system	0.052891	system	0.026080
language	0.015293	document	0.020445	online	0.034487	data	0.036081	documentation	0.010145
automated	0.014121	storage	0.012516	language	0.013451	storage	0.019271	library	0.008552
automatic	0.008262	evaluation	0.010251	theory	0.010901	computerized	0.010266	classification	0.008020
thesauri	0.007676	data	0.008552	query	0.010901	chemical	0.008465	storage	0.007489
chemistry	0.005332	automatic	0.006853	computerized	0.009626	medical	0.004863	international	0.006958
chemical	0.004746	model	0.006286	thesaurus	0.007076	literature	0.004863	patent	0.006427
thesaurus	0.004746	relevance	0.005720	evaluation	0.005801	biomedical	0.004262	study	0.005365
document	0.004746	indexing	0.004587	semantic	0.005164	evaluation	0.003662	science	0,004833
data	0.004160	online	0.004587	bibliography	0.005164	management	0.003662	control	0.004833
AUTHOR	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**
Y.L. Pokras	0.002562	G. Salton	0.005280	D.T. Hawkins	0.002758	S.J. Martinez	0.001692	H. Beck	0.002363
A.I. Chernyi	0.001630	A.G. Pickford	0.001799	N.A. Stokolova	0.002113	M.G. Manzone	0.001459	C.D. Gull	0.002305
A.N. Kulik	0.001572	W. Goffman	0.001741	E. Eisenbach	0.001878	C.M. Bowman	0.001401	D.J. Foskett	0.001844
G.E. Vleduts	0.001572	E. Garfield	0.001741	K. Yamanaka	0.001878	F.A. Landee	0.001284	I.M. Klempner	0.001786
R.A. Kennedy	0.001514	G.K. Thompson	0.001566	T. Radecki	0.001643	J. Frome	0.001284	B.C. Vickery	0.001786
T.M. Leonteva	0.001514	W.S. Cooper	0.001508	R. Fugmann	0.001585	I. Berghans	0.001225	B.R. Faden	0.001614
V.M. Averbukh	0.001514	K. Janda	0.001450	J. Eyre	0.001585	S.L. Visser	0.001225	H.W. Dillon	0.001556
V.B. Margarit	0.001514	F.W. Lancaster	0.001334	D.H. Kraft	0.001350	H. Skolnik	0.001225	C.P. Bourne	0.001556
V.S. Shetin	0.001397	R. Fugmann	0.001276	Z. Mazur	0.001350	Y.J. Lee	0.001167	S.P. Harter	0.001441
N.M. Sagalovi	0.001397	P. Willett	0.001276	K. Hosono	0.001291	T.K.S. Engar	0.001167	K.E. Marshall	0.001441
AUTHOR	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**
T.M. Leonteva	0.735676	G. Salton	7.546098	D.T. Hawkins	2.102022	A. Kent	1.003435	A. Kent	1.698628
G.E. Vleduts	0.715629	K. Janda	0.494193	N.A. Stokolova	0.745431	J. Frome	0.785387	C.P. Borune	0.868923
D.D. Arnaudov	0.701158	W. Goffman	0.329841	R.K. Summit	0.641113	R.K. Summit	0.696676	J.H. Shera	0.273393
A.I. Serebryanyi	0.653602	C.J. Vanrijsbergen	0.247591	M.E. Williams	0.56327	R.S. Ledley	0.384871	F.W. Lancaster	0.256723
J.W. Perry	0.363009	C.W. Cleverdon	0.242916	T. Radecki	0.36745	L.A. Hollaar	0.122698	S.P. Harter	0.233981
G.L. Mishchenko	0.254444	F.W. Lancaster	0.233755	A. Macleodi	0.341727	A. Fairthorner	0.109559	R.M. Needham	0.132524
A.V. Sokolov	0.245283	E. Garfield	0.200167	T. Saracevic	0.184206	V.E. Giuliano	0.098199	R.S. Taylor	0.131197
A.I. Chernyi	0.118691	C.N. Mooers	0.154252	R.S. Marcus	0.149631	D.J. Hillm,an	0.080544	T.H. Martin	0.110429
Y.I. Shemakin	0.110101	D.B. Mccarn	0.151314	R. Fugmann	0.142107	H.M. Kissman	0.050707	W.S. Cooper	0.104994
D.G. Lakhuti	0.079479	C.T. Yu	0.124452	C.T. Yu	0.093044	J. Oconnor	0.049592	J. Farradane	0.104113
AUTHOR	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**
G. Salton	0.047	G. Salton	0.0595	G. Salton	0.0484	G. Salton	0.0476	G. Salton	0.0493
A. Kent	0.0368	A. Kent	0.0359	A. Kent	0.0355	A. Kent	0.039	A. Kent	0.0395
M.E. Williams	0.0337	M.E. Williams	0.034	M.E. Williams	0.0355	M.E. Williams	0.0342	M.E. Williams	0.0339
F.W. Lancaster	0.031	F.W. Lancaster	0.031	F.W. Lancaseter	0.0314	F.W. Lancaster	0.031	F.W. Lancaster	0.0309
R.K. Summit	0.0273	R.K. Summit	0.0276	R.K. Summit	0.0292	R.K. Summit	0.0286	R.K. Summit	0.0271

C.W. Cleverdon	0.0223	C.W. Cleverdon	0.0231	D.T. Hawkins	0.0271	C.W. Cleverdon	0.0228	C.W. Cleverdon	0.0228
D.T. Hawkins	0.0208	D.T. Hawkins	0.0208	C.W. Cleverdon	0.0222	D.T. Hawkins	0.0215	D.T. Hawkins	0.021
D.B. Mccarn	0.0191	W.S. Cooper	0.02	D.B. Mccarn	0.0198	D.B. Mccarn	0.0193	C.P. Bourne	0.0194
W.S. Cooper	0.0181	D.B. Mccarn	0.0195	W.S. Cooper	0.0181	J. Frome	0.0184	D.B. Mccarn	0.019
H. Martint	0.0175	H. Martint	0,0173	H. Martint	0.0178	W.S. Cooper	0.0177	W.S. Cooper	0.0176
C.P. Bourne	0.0164	C.J. Vanrijsbergen	0.0165	C.P. Bourne	0.0159	H. Martint	0.0177	H. Martint	0.0176

Table A2: Topic and Author ranks in 1981-1990

Topic 1: Automatic IR System		**Topic 2: Online IR**		**Topic 3: Digital Library**		**Topic 4: Database and Query Processing**		**Topic 5: Evaluation**	
WORD	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**
system	0.047685	online	0.026992	online	0.034256	query	0.057781	systems	0.043182
automated	0.015251	systems	0.024298	text	0.023289	language	0.034293	document	0.031683
computerized	0.011800	text	0.015678	software	0.013612	query-processing	0.0248998	full-text	0.013283
language	0.009730	concepts	0.011368	system	0.013612	database	0.022884	model	0.013283
analysis	0.007660	reference	0.009213	library	0.007806	relational	0.020871	evaluation	0.012708
thesaurus	0.006280	principles	0.008135	microcomputer	0.007161	system	0.016845	fuzzy	0.012708
IRS	0.004900	proceedings	0.007597	chemical	0.007161	distributed	0.015502	effectiveness	0.009833
interactive	0.004900	practice	0.007058	directory	0.006516	data	0.009462	search	0.008108
assessment	0.004900	knowledge	0.006519	computerized	0.006516	database-system	0.008120	user	0.007533
effects	0.004210	services	0.006519	bibliography	0.005225	comparison	0.007449	expert	0.007533
AUTHOR	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**
G.L. Mishchenko	0.001679	P. Willett	0.005821	E. Garfield	0.004696	D.W. Stemple	0.001488	C.L. Borgman	0.003359
E.J. Brzezinski	0.001679	S.P. Harter	0.002789	D. Raitt	0.004387	R.H. Guting	0.001488	T. Radecki	0,003298
F. Kucklich	0.001617	C. Batt	0.002462	K. Yamanaka	0.002966	A. Sernadas	0.001364	G. Salton	0.002504
I.F. Pozhariskii	0.001617	D. Ellis	0.001940	D.T. Hawkins	0.002595	C. Katzeff	0.001364	W.B. Croft	0.002076
G.E.G. Russell	0.001492	M. Keen	0.001940	Y. Wolman	0.001915	S.Y. Su	0.001364	J.S. Ro	0.001954
S.B.N. Thompson	0.001492	S.E. Hocker	0.001880	J.R. Schroeder	0.001606	W. Perrizo	0.001364	J. Panyr	0.001893
V.I. Chibisov	0.001492	L. Bronars	0.001880	D.I. Raitt	0.001483	J.S. Davis	0.001302	D.C. Blair	0.001771
R.C. Sinclair	0.001438	P.G. Enser	0.001819	H.G. Fischer	0.001606	C.T. Yu	0.001302	M.E. Maron	0.001771
K.P. Pogorelko	0.001430	S. Stigleman	0.001819	A. Morris	0.001483	B.S. Goldshteyn	0.001302	P. Thompson	0,.001710
V.A. Kopylov	0.001430	B. Vickery	0.001759	N. Audino	0.001421	I.A. Macleod	0.001240	C.A. Lynch	0.001710
AUTHOR	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**
B.W. Kristalnyi	0.606384	S.E. Robertson	2.554820	E. Garfield	1.342955	C.T. Yu	1.181359	G. Salton	7.666796
N.J. Belkin	0.099750	D. Ellis	0.115226	D.T. Hawkins	0.564050	C.J. Vanrijsbergen	0.445802	T. Radecki	3.744411
B. Defude	0.096523	P. Willett	0.404440	C.H. Fenichel	0.276664	CHEN ALP	0.379662	A. Bookstein	2.334190
Z. Ozsoyoglu	0.090829	P. Ingwersen	0.387637	D.H. Kraft	0.136864	G. Ozsoyoglu	0.272104	W.B. Croft	1.936640
M.F. Porter	0.084146	B.C. Vickery	0.289286	P.W. Williams	0.069848	S.K. Chang	0.248238	D.A. Buell	0.509899
P.W. Williams	0.069863	A.S. Pollitt	0.095514	B. Defude	0.037989	A. Klug	0.212135	M.E. Maron	0.493130
S. Abiteboul	0.060936	D.H. Kraft	0.074263	C.T. Meadow	0.022864	P. Reisner	0.148074	S. Miyamoto	0.324912
P. Reisner	0.053376	H.M. Brooks	0.065617	P. Bollmann	0.018644	R. Snodgrass	0.141390	R.G. Crawford	0.223238
P. Bollmann	0.047370	A.F. Smeaton	0.065607	W. Kim	0.017838	N. Goodman	0.129649	D.C. Blair	0.216189
W. Kim	0.045322	E.A. Fox	0.059354	W.G. Waller	0.016226	S.K.M. Wong	0.110384	G.P. Zarri	0.206080
AUTHOR	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**
G. Salton	0.0842	G. Salton	0.0866	G. Salton	0.0844	G. Salton	0.0822	G. Salton	0.0919
A. Bookstein	0.0452	A. Bookstein	0.0462	A. Bookstein	0.0449	A. Bookstein	0.0437	A. Bookstein	0.0493
S.E. Robertson	0.039	S.E. Robertson	0.0419	S.E. Roberston	0.0387	S.E. Roberston	0.0377	T. Radecki	0.0425
T. Radecki	0.036	T. Radecki	0.0357	T. Radecki	0.0355	T. Radecki	0.0352	S.E. Roberston	0.0407
W.B. Croft	0.0316	W.B. Croft	0.0335	W.B. Croft	0.0318	W.B. Croft	0.0308	W.B. Croft	0.0351
C.J. Vanrijsbergen	0.0279	C.J. Vanrijsbergen	0.029	C.J. Vanrijsbergen	0.0276	C.T. Yu	0.029	C.J. Vanrijsbergen	0.0291
C.T. Yu	0.0267	C.T. Yu	0.0258	C.T. Yu	0.0256	C.J. Vanrijsbergen	0.0278	C.T. Yu	0.0258
W.S. Cooper	0.0242	W.S. Cooper	0.0233	W.S. Cooper	0.023	W.S. Cooper	0.0222	W.S. Cooper	0.0239
K.S. Jones	0.0212	P. Willett	0.0225	K.S. Jones	0.0215	K.S. Jones	0.0206	K.S. Jones	0.0221
D.A. Buell	0.0193	K.S. Jones	0.0222	D.A. Buell	0.0191	D.A. Buell	0.0189	D.A. Buell	

Table A3: Topic and Author ranks in 1991-2000

Topic 1: Web IR	**Topic 2: Multimedia IR**	**Topic 3: Evaluation**	**Topic 4: Medical IR**	**Topic 5: Database and Query Processing**

WORD	PROB	WORD	PROB	WORD	PROB	WORD	PROB	WORD	PROB
system	0.019890	image	0.046076	text	0.013861	database	0.018236	query	0.042509
web	0.014922	content-based	0.019150	evaluation	0.010767	medical	0.011277	database	0.028761
knowledge	0.012352	system	0.013481	systems	0.010031	system	0.010868	databases	0.021575
database	0.011667	indexing	0.011532	searching	0.009736	clinical	0.007798	data	0.018294
data	0.009611	databases	0.008875	search	0.009442	patient	0.003909	object-oriented	0.014076
query	0.009440	multimedia	0.008875	online	0.009147	management	0.003704	queries	0.013763
design	0.008754	images	0.008344	relevance	0.008558	health	0.003704	processing	0.011420
text	0.008069	visual	0.008344	library	0.007388	identification	0.003500	relational	0.011264
management	0.007727	video	0.007989	user	0.006349	automated	0.003500	model	0.010639
distributed	0.007555	color	0.006926	hypertext	0.006349	optical	0.003295	language	0.010483
AUTHOR	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**
W.B. Corft	0.000301	H.C. Chen	0.000348	A. Spink	0.000756	S.G. Aiken	0.000363	J. Han	0.000529
H.C. Chen	0.000301	F. Crestani	0.000348	R.M. Losee	0.000744	I. Soutar	0.000328	D. Suciu	0.000529
W. Umstatter	0.000278	A.K. Jain	0.000337	E. Levine	0.000687	S. Barcza	0.000304	H.P. Kriegel	0.000460
C.A. Lynch	0.000255	E. Wilhelm	0.000337	C. Cole	0.000573	C.C. Tsai	0.000304	S.Y. Su	0.000449
P. Martin	0.000255	J.I. Khan	0.000325	P. Willett	0.000561	W. Hersh	0.000293	K.L. Tan	0.000449
D. Samson	0.000255	B.S. Manjunath	0.000313	W.R. Hersh	0.000515	S.J. Westerman	0.000281	G. Graefe	0.000380
N.J. Santora	0.000255	H.K. Kim	0.000302	C.T. Meadow	0.000435	H.H. Emurian	0.000269	L. Wong	0.000380
C. Womserhacker	0.000243	H.M. Wang	0.000302	B. Hjorland	0.000412	L.L. Consaul	0.000269	L. Libkin	0.000368
N.J. Belkin	0.000243	S.F. Chang	0.000290	E. Garfield	0.000401	H.J. Markowitsch	0.000269	J.W. Su	0.000368
R. Wagnerdobler	0.000243	S. Levialdi	0.000290	T. Cawkell	0.000389	D. Roberts	0.000258	P.Z. Revesz	0.000357
AUTHOR	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**
N.J. Belkin	1.611918	W.B. Croft	0.620038	G. Salton	4.181629	E. Garfield	0.2727	S. Abiteboul	1.686384
W.B. Frakes	0.208332	A.K. Jain	0.553072	T. Saracevic	1.03712	W.R. Hersh	0.244266	W. Litwin	0.197114
T. Imielinski	0.167662	S.K. Chang	0.356876	S.E. Robertson	0.925638	S. Ceri	0.079951	J. Paredaens	0.196647
G.W. Furnas	0.118589	J.K. Wu	0.304342	M.J. Bates	0.570592	T. Bernerslee	0.079585	M.J. Egenhofer	0.175892
T. Catarci	0.101004	S.Y. Lee	0.182347	P. Willett	0.485012	D.R. Swanson	0.073879	S. Grumbach	0.171572
T. Kohonen	0.07921	W.S. Cooper	0.093102	A. Spink	0.413471	T. Kohonen	0.066366	C.J. Vanrijsbergen	0.136199
R. Agrawal	0.075587	H.C. Chen	0.091848	C.L. Borgman	0.224815	W. Kim	0.065877	S.Y. Lee	0.127221
S.K. Chang	0.070374	A. Pentland	0.077482	R. Fidel	0.191618	J.P. Callan	0.061553	M. Kifer	0.122086
H.C. Chen	0.065015	A. Gupta	0.0681	K.S. Jones	0.159335	R. Reiter	0.053147	R. Snodgrass	0.121895
P. Valduriez	0.061602	A. Delbimbo	0.061798	C. Stanfill	0.121081	P. Buneman	0.043532	H. Samet	0.102031
AUTHOR	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**
G. Salton	0.0523	G. Salton	0.0525	G. Salton	0.0567	G. Salton	0.0519	G. Salton	0.0478
N.J. Belkin	0.0287	N.J. Belkin	0.0269	N.J. Belkin	0.0315	N.J. Belkin	0.0271	S. Abiteboul	0.0307
S.E. Robertson	0.0256	S.E. Robertson	0.0256	S.E. Robertson	0.03	S. Abiteboul	0.0255	N.J. Belkin	0.0245
S. Abiteboul	0.0248	S. Abiteboul	0.0239	T. Saracevic	0.0272	S.E. Robertson	0.0255	S.E. Robertson	0.0233
T. Saracevic	0.0217	T. Saracevic	0.021	C.J. Vanrijsbergen	0.023	T. Saracevic	0.0213	T. Saracevic	0.0192
C.J. Vanrijsbergen	0.0208	C.J. Vanrijsbergen	0.0207	M.J. Bates	0.0216	C.J. Vanrijsbergen	0.0207	C.J. Vanrijsbergen	0.0192
W.B. Croft	0.0191	W.B. Croft	0.0199	W.B. Croft	0.0209	W.B. Croft	0.0189	W.B. Croft	0.0173
M.J. Bates	0.0178	M.J. Bates	0.0174	S. Abiteboul	0.0201	M.J. Bates	0.0173	M.J. Bates	0.0156
K.S. Jones	0.0153	S.K. Chang	0.0166	A. Spink	0.019	K.S. Jones	0.0153	M. Stonbraker	0.0156
D. Harman	0.0145	K.S. Jones	0.0153	K.S. Jones	0.017	D. Harman	0.0142	E.F. Codd	0.0146

Table A4: Topic and Author ranks in 2001-2008

Topic 1: Multimedia IR		Topic 2: Database and Query Processing		Topic 3: Medical IR		Topic 4: Web IR and Digital library		Topic 5: IR Theory and Model	
WORD	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**
image	0.063250	query	0.033203	database	0.010424	web	0.023822	document	0.014450
content-based	0.017681	data	0.025732	medical	0.007140	search	0.015858	text	0.010966
learning	0.008809	xml	0.019248	health	0.004982	digital	0.008366	query	0.009878
images	0.008667	processing	0.018614	clinical	0.004513	searching	0.006395	image	0.009587
relevance	0.008383	queries	0.016147	management	0.004325	knowledge	0.006001	relevance	0.008499
color	0.008312	databases	0.012764	search	0.004138	system	0.005764	fuzzy	0.008281
feedback	0.008312	database	0.009733	design	0.004138	query	0.005764	web	0.007991
video	0.007673	efficient	0.009451	study	0.003668	user	0.005528	model	0.006539
semantic	0.007389	web	0.009381	support	0.003575	model	0.005212	system	0.006321
similarity	0.007318	querying	0.008958	knowledge	0.003575	internet	0.004424	cross-language	0.006176

AUTHOR	PROB	AUTHOR	PROB	AUTHOR	PROB	AUTHOR	PROB	AUTHOR	PROB
T.S. Huang	0.000572	J.Z. Li	0.000572	R.N. Kostoff	0.000321	M. Thelwall	0.000628	F. Crestani	0.000573
H.J. Zhang	0.000528	F. Bry	0.000415	U.J. Balis	0.000283	C.C. Yang	0.000545	G.J.F. Jones	0.000554
G.J. Lu	0.000409	H.J. Kim	0.000371	G. Eysenbach	0.000257	A. Spink	0.000457	E. Herrera-wiedma	0.000548
J. Li	0.000365	D. Papadias	0.000364	R.B. Haynes	0.000251	P. Jacso	0.000444	J. Savoy	0.000510
C.C. Chang	0.000358	K. Subieta	0.000358	G. Nilsson	0.000238	I. Fourie	0.000425	M. Lalmas	0.000510
E. Izquierdo	0.000352	J. Van den Bussche	0.000339	H. Shatkay	0.000218	H.C. Chen	0.000393	K. Jarvelin	0.000510
J. Lassksonen	0.000327	D. Taniar	0.000327	N.L. Wilczynski	0.000218	N. Ford	0.000381	N. Kando	0.000466
H. Burkhardt	0.000308	F. Geerts	0.000327	C.R. Shyu	0.000218	H. Xie	0.000368	S.M. Chen	0.000403
C.J. Liu	0.000308	M. Song	0.000320	J.I. Westbrook	0.000218	G.G. Chowdhury	0.000355	N. Fuhr	0.000397
D. Ziou	0.000302	Y.D. Chung	0.000320	G.O. Babnett	0.000212	B. Hjorland	0.000349	I. Ounis	0.000378
AUTHOR	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**	**AUTHOR**	**I_PR**
Y. Rui	1.395121	H.V. Jagadish	0.924168	R.N. Kostoff	1.523479	A. Spink	2.779089	N. Fuhr	0.78954
J.P. Eakins	0.575893	D. Calvanese	0.546725	C. Buckley	0.114193	T. Saracevic	1.57628	C. Zhai	0.491143
J. Li	0.421822	M.J. Egenhofer	0.442505	S. Berchtold	0.060519	B. Hjorland	0.768776	F. Crestani	0.486208
S. Chaudhuri	0.402936	G. Gottlob	0.40447	M.W. Berry	0.059129	S.E. Roberston	0.600389	C.J. Vanrijsbergen	0.480959
J.R. Smith	0.357613	S. Abiteboul	0.390116	S.F. Chang	0.056708	B.J. Jansen	0.531536	J. Savoy	0.354813
A.W.M. Smeulders	0.291388	G. Graefe	0.329796	R. Fagin	0.042499	N.J. Belkin	0.457445	K.S. Jones	0.320696
J. Han	0.284959	W.B. Frakes	0.290321	H. Muller	0.034922	E.M. Voorhees	0.282649	S.E. Robertson	0.211595
T. Gevers	0.201961	L. Gravano	0.225926	D. Florescu	0.0326	W.R. Hersh	0.111195	J.P. Callan	0.198884
N. Vasconcelos	0.17684	R. Baezayates	0.186399	D.R. Swanson	0.031936	P. Ingwersen	0.106922	R.R. Yager	0.198019
R.M. Haralick	0.168773	M. Fernandez	0.171795	A.K. Jain	0.031742	P. Vakkari	0.104115	D. Hawking	0.191976
AUTHOR	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**	**AUTHOR**	**PR_t(.85)**
G. Salton	0.0469	G. Salton	0.0486	G. Salton	0.0496	G. Salton	0.0504	G. Salton	0.0529
Y. Rui	0.0319	Y. Rui	0.0256	Y. Rui	0.0256	A. Spink	0.034	S.E. Robertson	0.027
J.R. Smith	0.0248	S.E. Robertson	0.0232	S.E. Robertson	0.024	N.J. Belkin	0.0273	A. Spink	0.0248
S.E. Robertson	0.0218	A. Spink	0.0218	A. Spink	0.0233	T. Saracevic	0.0193	N.J. Belkin	0.0237
A. Spink	0.0207	N.J. Belkin	0.0215	N.J. Belkin	0.0229	S.E. Roberston	0.0258	Y. Rui	0.0228
N.J. Belkin	0.0205	J.R. Smith	0.021	J.R. Smith	0.0203	Y. Rui	0.0219	T. Saracevic	0.0194
T. Saracevic	0.017	S. Abiteboul	0.0184	T. Saracevic	0.0193	E.M. Voorhees	0.0188	E.M. Voorhees	0.0191
E.M. Voorhees	0.016	T. Saracevic	0.0179	E.M. Voorhees	0.0176	B.J. Jansen	0.0188	N. Fuhr	0.0187
A.W.M. Smeulders	0.0157	E.M. Voorhess	0.169	D. Harman	0.0157	J.R. Smith	0.018	J.R. Smith	0.0181
R. Baezayates	0.0151	D. Harman	0.0152	K.S. Jones	0.0154	K.S. Jones	0.0163	K.S. Jones	0.0172

Table A5: Notations for various LDA formulas in Related Work_Topic Modeling (LDA) section.

Notations	Meaning
d	document
w	word
x	author
z	topic
c	publication venue
N_d	the number of words in the current document
N_D	the number of words in the entire collection of documents
a_d	the set of co-authors
α	hyperparameter for generating Θ from Dirichlet Distribution
β	hyperparameter for generating φ from Dirichlet Distribution
μ	hyperparameter for generating Ψ from Dirichlet Distribution
Θ	a multinomial distribution over topics
φ	a multinomial distribution over words
Ψ	a multinomial distribution over publication venues
D	collection of documents
A	collection of authors
T	collection of topics

C_{mj}^{WT}	The number of times assign mth word in lexicon to topic j
C_{dj}^{DT}	The number of times assign dth document to topic j
C_{aj}^{AT}	The number of times assign ath author to topic j
C_{cj}^{CT}	The number of times assign cth conference to topic j
z_{-di}	All word-topic assignment not include current situation (assign word *i* in document *d* to a random topic in current instance)
x_{-di}	All word-author assignment not include current situation (assign word *i* in document *d* to a random author in current instance)
m_{xz}	The number of times assign topic *z* to author *x*
n_{zv}	The number of times assign word *v* to topic *z*
n_{zc}	The number of times assign conference *c* to topic *z*

Table A6. Notations for various PageRank formulas in Related Work_Topic Related PageRank section

Notations	Meaning
λ	The damping factor: probability of a random jump in the random surfer model
W	The set of web pages
N	The number of pages in W
T	The number of topics
z	The *z*th Topic
$I(v)$	In-degree of page *v*
$O(v)$	Out-degree of page*v*
$A(v)$	Authority score of page *v*
$H(v)$	Hubness score of page *v*
$p \rightarrow q$	There is a hyperlink on page *p* that points to *q*
τ_j	The set of URLs within topic z_j
$TSPR_t(k)$	Topic-sensitive PageRank for page *k* on topic *t*
$IS_q(i)$	For a specific query *q*, page *i*'s query-dependent PageRank score
$r(q, i)$	Query *q*'s relevance to page *i*
α	When following a link, the probability for surfer to stay on the same topic to maintain topic continuity
$C(i_z)$	The content vector of topic z in page i