


Article

Citation Oriented AuthorRank for Scientific Publication Ranking

Jinsong Zhang ¹  and Xiaozhong Liu ^{2,*}

¹ School of Maritime Economics and Management, Dalian Maritime University, Dalian 116026, China; jinsong_zhang@dlmu.edu.cn

² School of Informatics, Computing and Engineering, Indiana University Bloomington, Bloomington, IN 47405, USA

* Correspondence: liu237@indiana.edu

Abstract: It is now generally accepted that an article written by influential authors often deserves a higher ranking in information retrieval. However, it is a challenging task to determine an author's relative influence since information about the author is, much of the time, inaccessible. Actually, in scientific publications, the author is an important metadata item, which has been widely used in previous studies. In this paper, we bring an optimized AuthorRank, which is a topic-sensitive algorithm calculated by citation context, into citation analysis for testing whether and how topical AuthorRank can replace or enhance classical PageRank for publication ranking. For this purpose, we first propose a PageRank with Priors (PRP) algorithm to rank publications and authors. PRP is an optimized PageRank algorithm supervised by the Labeled Latent Dirichlet Allocation (Labeled-LDA) topic model with full-text information extraction. We then compared four methods of generating an AuthorRank score, looking, respectively, at the first author, the last author, the most famous author, and the "average" author (of a publication). Additionally, two combination methods (Linear and Cobb–Douglas) of AuthorRank and PRP were compared with several baselines. Finally, as shown in our evaluation results, the performance of AuthorRank combined with PRP is better ($p < 0.001$) than other baselines for publication ranking.



Citation: Zhang, J.; Liu, X. Citation Oriented AuthorRank for Scientific Publication Ranking. *Appl. Sci.* **2022**, *12*, 4345. <https://doi.org/10.3390/app12094345>

Academic Editors: Luigi Di Caro and Claudio Schifanella

Received: 10 February 2022

Accepted: 22 April 2022

Published: 25 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: AuthorRank; publication ranking; PageRank with Priors; citation context analysis

1. Introduction

With the rapid increase in online scientific literature, the task of publication ranking has become increasingly important for scholarly information retrieval and recommendation, which aims at enhancing the efficiency of research. Generally speaking, the article with more influence should be ranked higher in the publication ranking. While citation analysis plus graph mining is a commonly used bibliometric method of performing this task. As contrasted with the traditional method of simple citation counting [1], PageRank [2] is a global page ranking algorithm based on the intuition that links from important pages are more significant than links from other pages. It is an effective means of ranking web pages and scientific publications for addressing user information needs. The relative "importance" of a node is calculated by back-links with transition authority from other nodes, thereby determining which nodes are more important in the network.

However, the original PageRank algorithm is based solely on links, independent of any particular search query, and regardless of other metadata in the content, e.g., title, keywords, abstract, and link location. Because of this limitation, some previous studies have tried to improve PageRank by employing topic-sensitive methods [3], which take the topic into account in determining the rank. Then, it can be widely used for discovering important papers [4].

As one contribution of this paper, we used an optimized topic-sensitive PageRank algorithm, PageRank with Priors (PRP), to determine the topical ranking scores of scientific publications and authors. This method has been proposed in our previous work [5],

which combined PageRank and a supervised topic modeling algorithm Labeled-LDA [6] via full-text extraction with two priors: a publication topic probability and a citation transition probability, and, then, publication ranking results and author ranking results were presented by the PRP algorithm.

On the other hand, the idea behind AuthorRank [7] is that content created by more popular authors should rank higher than content created by less popular authors. It was proposed by Google's Agent Rank patent in 2005, but cannot be implemented in ranking web pages, because it is difficult and expensive to assess the topical reputation or importance of many Web page creators. Actually, it is much easier and economical in the scientific literature domain. For scientific literature, the creator, as an author, is a common and important metadata item, easy to extract from a digital library, and widely studied in previous research [8]. So, the other contribution of this paper is to generate a topical AuthorRank score, which is based on full-text extraction and the Labeled-LDA model, according to topic-sensitive author ranking results by four methods: first author, last author, the most famous author, and average author.

Finally, to test whether topical AuthorRank can replace PageRank, or how it can make the results more accurate for publication ranking [9], we will compare publication ranking results by using topical AuthorRank plus PRP to validate the results. In order to confirm that AuthorRank positively influences publication ranking, two combination methods will be validated, the linear method and the Cobb–Douglas method.

In the remainder of this paper, we: (1) review relevant literature on the methodology of publication ranking and AuthorRank; (2) introduce our novel methods for publication and AuthorRank, as well as how to achieve publication ranking with AuthorRank; (3) describe the experimental setting and evaluation results; and (4) discuss the findings and limitations of the study and identify subsequent research steps.

2. Previous Research

2.1. Publication Ranking

Currently, rapid access to digital publications can greatly accelerate research, but the information overload also challenges researchers, especially junior researchers, in finding appropriate citations for their research projects. In previous papers, to deal with this problem, bibliometric methods have focused on ranking publications by citation analysis along with graph mining. A common and easy method based on citation frequency or citation impact has been studied for a long time. For example, the article [10] used citation analysis to assist college chemistry libraries with selecting significant books. In 1972, the volume of publications had increased dramatically, and with the emerging availability of electronic and online resources, it described a method of evaluating scientific journals by citation frequency and impact using data from digital resources [11]. More recently, citation information has been successfully used to enhance information retrieval performance [12].

Nevertheless, various methods were based on a basic assumption, which is straightforward: if paper1 cites paper2, then paper1 and paper2 are related. This assumption is oversimplified because it treats all citations equally, regardless of sentiment, reason, topic, or motivation. Hence, some scholars have considered other factors bearing upon citation analysis results. Citation location was a significant factor [13], noting that a publication cited in the introduction or literature review section and mentioned again in the methodology or discussion sections is likely to make a greater overall contribution to the citing publication than others that have been mentioned only once. Further, citation context [14], a text window containing the target citation tag, may provide detailed and direct information about the nature of a citation, and also can be used to infer semantic information about the cited paper.

The PageRank algorithm, first proposed and used in Google Search in 1998, is a significant method for evaluating node importance via complex graphs analysis, e.g., social networks, web graphs, telecommunication networks, and biological networks. However, PageRank is insensitive to topics and queries. Haveliwala [3] proposed a topic-sensitive

PageRank algorithm, computing a set of PageRank vectors biased using a set of representative topics to capture more accurately the notion of importance with respect to a particular topic by computing topic-sensitive PageRank scores using the topic of the context in which the query appears, and then generating context-specific importance scores for pages using linear combinations of biased PageRank vectors. After that, more and more researchers have used PageRank or the improved PageRank algorithm in the field of bibliometrics for measuring publication and author importance [15–17].

Currently, more and more scholars tried to optimize the PageRank by topic modeling. Latent Dirichlet Allocation (LDA) [18] is the most widely used method among them. Recent work shows that LDA topic modeling can be applied to scholarly network citation analysis. Labeled-LDA [6] is a supervised topic model that constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user tags. The score calculated by topic modeling will incorporate prior probabilities as a relative-rank extension to PageRank [19]. Although some previous studies have used the topic model in conjunction with the PageRank algorithm [20], most only utilize the metadata of title, author, and abstract, but neglect full-text metadata.

In this paper, we combined a Labeled-LDA topic model with PageRank via full-text extraction and incorporated two prior probabilities—publication topic probability and citation transition probability—for ranking the publications in our dataset.

2.2. AuthorRank

AuthorRank was first proposed in 2005 in a Google Agent Rank patent in terms of “using the reception of the content the agents create and their interactions as a factor in determining their rank” [21]. Thus, the author or creator of the content is significant for ranking. In general, publication rank is a function of PageRank, as well as AuthorRank.

In scientific literature, annual citation rates of ecological papers are affected by many factors, including the hypothesis tested, article length, and authors' information [22]. Author metadata is also a common element in publications, and is simple and easy to extract [23]. So, in this paper, we will use the metadata of Authors to assess whether AuthorRank as applied to scientific literature can improve ranking results.

AuthorRank is different from author ranking. Author ranking means ranking authors in a selected dataset according to some factors [24], e.g., publications, research interest, or affiliation. Authors who rank higher should be more important in this field.

Author ranking is usually used in expertise retrieval [25]. The most popular approaches in expertise retrieval are Generative Probabilistic Models, Discriminative Probabilistic Models, Voting Models, and Graph-based Models [26]. PageRank has also been used in expertise retrieval for author ranking [27] by combining a voting model with a graph-based model.

In contrast, AuthorRank determines how the ranking of content is affected by the creator (author). Thus, while AuthorRank ranks publications, it does so based upon author ranking. The article [28] computes a related AuthorRank, applying the PageRank algorithm to citations among authors and quantifying the impact of an author or paper, taking into account the impact of those authors that cite it. The authors [29] use the first author's reputation for generating an AuthorRank score, and verified that the performance of AuthorRank combined with topic-based PageRank is better than other baselines.

In this paper, we use a PageRank with Prior algorithm, combined with Labeled-LDA and full-text extraction, to rank authors in our dataset, and, then, compute an AuthorRank by author ranking based on four methods: First Author, Last Author, Max Author, and Average Author. All these methods will be introduced in the following section.

3. Methodology

3.1. Topic Modeling of LLDA

Labeled Latent Dirichlet Allocation (Labeled-LDA or LLDA) was proposed for training the labeled topic model. It employs a generative probabilistic model in the hierarchical

Bayesian framework and assumes the availability of topic labels and the characterization of each topic by a multinomial distribution, β_{key_i} , overall vocabulary words.

In Labeled-LDA, W is a set of words, w_i , chosen from a document in the training text, $W = \{w_1, w_2, \dots, w_n\}$. n is the number of words. The set of documents in this article, as $P = \{p_1, p_2, \dots, p_d\}$. d is the number of documents. KEY is a set of labels, key_s ; we used keywords as labels from training text, so $KEY = \{key_1, key_2, \dots, key_m\}$. m is the number of labels. Words are chosen in proportion to a label's preference for the word, β_{key_i} , and the publication's preference for the associated label, θ_{p_j} .

$$\beta_{key_i} = \begin{pmatrix} P_{key_1}(w_1) & \cdots & P_{key_1}(w_n) \\ \vdots & \ddots & \vdots \\ P_{key_m}(w_1) & \cdots & P_{key_m}(w_n) \end{pmatrix} \quad (1)$$

$$\theta_{p_j} = \begin{pmatrix} P_{key_1}(p_1) & \cdots & P_{key_m}(p_1) \\ \vdots & \ddots & \vdots \\ P_{key_1}(p_d) & \cdots & P_{key_m}(p_d) \end{pmatrix} \quad (2)$$

The two matrices can be constructed by $P_{key_s}(w_i)$ and $P_{key_s}(p_j)$, representing the co-occurrence probability of label (key_s) and word (w_i), and the co-occurrence probability of occurrence (p_j) and label (key_s), respectively.

$$P_{key_s}(w_i) = \frac{P(w_i|key_s)}{\sum_{t=1}^{|W|} P(key_s|w_t)} \quad (3)$$

$$P_{key_s}(p_j) = \frac{P(key_s|p_j)}{\sum_{x=1}^{|V|} P(key_s|p_x)} \quad (4)$$

Stemming from the LDA method, Labeled-LDA is a supervised topic modeling algorithm, which employs existing topics from scientific metadata. Therefore, in this paper, topic labels were assigned based on author-assigned keywords, and each scientific publication was treated as a mixture of its author-assigned topics (keywords). As a result, both topic labels and topic numbers (the total number of keywords in the metadata repository) are given.

We can therefore infer a possible topic distribution for each paper by LLDA. Figure 1 is an example of 2 topic distributions; to show it clearly, we randomly sampled 50. The horizontal axis shows the possible topics (keywords authorized by the author), and the vertical axis is the topic probability. Content in the paper includes title, abstract, and full-text or citation context in publications. For different content, the topic probability distribution is diverse, while the sum of all values (topic probabilities) for each paper is equal to 1.

The blue line is a conference paper, "Using architectural 'families' to increase FPGA speed and density", which is about narrowing the speed and density gap between FPGAs and MPGAs. So, the highest topic in this paper is "FPGA". The orange line is a paper named "Rerun: Exploiting Episodes for Lightweight Memory Race Recording", published in "the 35th Annual International Symposium on Computer Architecture", which mainly focuses on the field of Multiprocessor deterministic replay. So, the highest topic in this paper is "multi core processor".

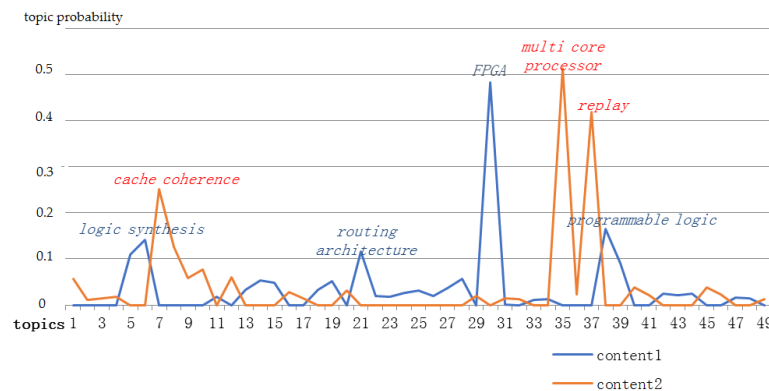


Figure 1. Example of LLDA Topic Model.

In this method, the labels are keywords from the sources. These keywords may be author-provided or derived by greedy matching. Greedy matching means loading all possible keywords into memory, and then searching each keyword from the paper title and abstract by using fast matching, with the purpose of expanding the paper topic space.

3.2. PageRank with Priors

There are many different ranking methods (e.g., citations, publications, h-index, and PageRank) in the field of scientific literature. In this paper, we employ a topic-dependent ranking method based on the combination of Labeled-LDA and PageRank with Priors (PRP), an optimized PageRank algorithm for full-text extraction [5]).

In bibliometrics, most previous studies on citation network analysis are based on the simple assumption that *paper₁* and *paper₂* are connected, whenever *paper₁* cites *paper₂*. In this paper, we represent two kinds of prior knowledge in a citation graph by LLDA: a publication topic probability and a citation transition probability.

In this paper, we first create an academic publication network. Among them, the vertex of the network is the academic publication, v_i , the dataset of all publications means V , and the edge of the network is the citation relationship between papers, e_i , the dataset of all citations we used in the network means E .

For each vertex, $v_i \in V$, the publication topic prior vector is $\{p_{z_{key_1}}(v_i), p_{z_{key_2}}(v_i), \dots, p_{z_{key_m}}(v_i)\}$, where $p_{z_{key_m}}(v_i) = P_{key_m}(p_j)$. $P_{key_m}(p_j)$ is the co-occurrence probability of paper (p_j) and label (key_s). The prior probability of vertex v for topic z_{key_m} , is trained by publication metadata (title, abstract, and full-text), and where $\sum_{i=1}^{|V|} p_{z_{key_i}}(v) = 1$.

Each edge, $e_i \in E$, on the graph represents a citation connecting v_i and v_j (v_i cites v_j). The citation topic transitioning probability vector for each edge is $\{p_{z_{key_1}}(v_i|v_j), p_{z_{key_2}}(v_i|v_j), \dots, p_{z_{key_m}}(v_i|v_j)\}$, where $p_{z_{key_s}}(v_i|v_j)$ is the probability of transitioning from vertex v_i to v_j for topic z_{key_s} .

For each topic z_{key_s} , the score of vertex $p_{z_{key_s}}(v)$ and edge $p_{z_{key_s}}(v_i|v_j)$ calculated by Labeled-LDA and author-assigned keywords plus greedy matching results. Hence, graphs for different topics may be different. If topic z_{key_s} does not belong to paper v_i , the publication topic prior $p_{z_{key_s}}(v)$ is 0. If both citing and cited papers do not include topic z_{key_t} , we assign $p_{z_{key_1}}(v_i|v_j)$ a lower score: $p_{z_{key_1}}(v_i|v_j)' = \psi \cdot p_{z_{key_1}}(v_i|v_j)$, where $\psi = 0.1$, because we did not want to totally remove this citation from the academic publication network.

The PageRank with priors algorithm takes into account these two kinds of priors, $p_{z_{key_s}}(v_i)$ and $p_{z_{key_m}}(v_i|v_j)$, to calculate the relative importance of vertices in the citation graph. A hyperlink to a publication counts as a vote of support. For example, if a publication cites only 3 papers and, for a specific topic, the transitioning probabilities to these 3 papers are 0.1, 0.1, and 0.8, then most of the paper's credit on this topic (it is topic author-

ity) goes to the third paper. So, the vertex's (topic relative) importance can be calculated by $I_{key_m}(p_i|PR) = \pi_{key_m}(v)$, and:

$$\pi_{key_m}(v)^{i+1} = (1 - \beta_b) \left(\sum_{u=1}^{d_m(v)} p_{z_{key_m}}(v|u) \pi_{key_m}^{i+1}(u) \right) + \beta_b p_{z_{key_m}}(v) \tag{5}$$

The output, for each vertex (publication), v , is an authority vector $\{A_{z_{key_1}}(v), A_{z_{key_2}}(v), \dots, A_{z_{key_m}}(v)\}$. Each authority score in the vector indicates the publication topic importance with respect to both paper topic and full-text citation priors. We obtain ranking lists as a result ($I_{key_m}(p_i|PR)$).

3.3. Author Ranking by PRP

As described before, we used PRP for publication ranking by combining Labeled-LDA-based topic information and full-text publication information. We also applied this method of author ranking based on the assumption that if author₁'s paper cites author₂'s paper, then author₁ and author₂ are somehow related. The relation can be characterized on a directed graph with authors as vertices and citations as edges. In the author graph, $G_A = (V', E')$, V' is a set of vertices representing all the authors; and E' is a set of edges representing author relationships as generated from the citation network.

In most cases, one author has multi-publications, so v'_i , the vertex for a given author is a set of papers, i.e., $v'_i = \{p_1, p_2, p_3, \dots, p_s\}$. Similarly, the number of edges between two vertices $\langle v'_i, v'_j \rangle$ is always more than one, each edge e'_l is expressed as $\langle v'_i, v'_j \rangle = \{ \langle p_1, p_2 \rangle, \langle p_1, p_3 \rangle, \dots, \langle p_l, p_k \rangle \}$, only when $p_l \in v'_i$ and $p_k \in v'_j$.

Therefore, the publication topic prior can be formulated as:

$$p_{z_{key_m}}(v') = \sum p_{z_{key_m}}(v) = \sum_d \frac{P(key_m|p_d)}{\sum_{x=1}^{|V|} P(key_m|p_x)} \tag{6}$$

where $p_d \in v'$, and the transitioning probability score for an edge can be calculated as:

$$p_{z_{key_m}}(v'_i|v'_j) = \sum p_{z_{key_m}}(v_i|v_j) = \sum \frac{P(key_m|citation_{j,i})}{\sum_{x=1}^{d_{out}(v_j)} P(key_m|citation_{j,x})} \tag{7}$$

where $citation_{j,i}$ is the citation context from v'_i to v'_j .

So, the (topic-relative) importance of an author (vertex) can be calculated by the same formula as: $I_{key_m}(v'|R) = \pi_{key_m}(v')$. For output, we can obtain ranking lists for each specific topic as a result.

Author topical ranking by PRP may be more accurate than paper ranking, as each vertex is represented by a list of papers from an author, and more textual information results in more accurate topic inference.

3.4. AuthorRank by Author Ranking

We introduced the concept of AuthorRank and its underlying assumptions in the previous section. Generally speaking, author ranking may influence publication ranking. Thus, a famous or important author on a specific topic will have his or her publications ranked higher, which will thereby have a greater chance to be recommended for the topic. To validate this hypothesis, in this section we propose to use AuthorRank with a topical author ranking score.

Since AuthorRank is a query-independent criterion similar to PageRank, it should be calculated as offline processing, where content authority is measured by the authority accumulated from links, regardless of the query. The PRP algorithm based on the author-ranking method has been proposed before. It is effective for allowing an author to have a

different rank for each topic. So, in this section, we propose a method for paper ranking (AuthorRank) via author ranking, which is also an offline and topic-based method.

The following formula defines the relationship between paper and author.

$$p_i = \{a_1, a_2, a_3, \dots, a_s\} \tag{8}$$

where p_i is the paper, always created by at least one author a_i . How the author affects the publication ranking score, $I_{key_m}(p_i|AR)$, can be calculated in at least four different ways: First Author, Last Author, Max Author, and Average Author.

First Author: The ranking score of a publication (AuthorRank) depends only on the first author’s ranking. In other words, the publication’s topical importance, $I_{key_m}(p_i)$, is based on the score of only the first author ($\pi_{key_m}(a_1)$).

$$I_{key_m}(p_i|AR) = \pi_{key_m}(a_1) \tag{9}$$

Last Author: The AuthorRank score of a publication, p_i , depends only on the last author’s ranking, where a_s is the last author if p_i is co-authored by multiple (s) authors. Otherwise, as is decided by the unique author.

$$I_{key_m}(p_i|AR) = \pi_{key_m}(a_s) \tag{10}$$

Max Author: The AuthorRank score of a publication is decided by the most popular author’s score among all the authors.

$$I_{key_m}(p_i|AR) = \max\{\pi_{key_m}(a_1), \pi_{key_m}(a_2), \dots, \pi_{key_m}(a_s)\} \tag{11}$$

Average Author: The AuthorRank score of a publication is determined by the average scores of all the publication’s authors.

$$I_{key_m}(p_i|AR) = \frac{\sum_{j=1}^n \pi_{key_m}(a_j)}{|s|} \tag{12}$$

Example: To understand these methods better, we provide a simple example. Assume that there are two papers: $paper_1$ and $paper_2$, where $paper_1$ is authored by $author_1$ and $author_2$, as $p_1 = \{a_1, a_2\}$; and where $paper_2$ is authored by $author_1$, $author_3$, and $author_4$, as $p_2 = \{a_1, a_3, a_4\}$. The importance score of these four authors is shown in Table 1.

Table 1. Exemplar author ranking score.

Author Name	a_1	a_2	a_3	a_4
score	0.2	0.5	0.12	0.18

From the four methods mentioned above, we obtain the two papers’ ranking scores as shown in Table 2.

Table 2. AuthorRank for paper ranking.

	First Author	Last Author	Max Author	Average Author
$paper_1$	$a_1 = 0.2$	$a_2 = 0.5$	$a_2 = 0.5$	$(a_1 + a_2)/2 = 0.35$
$paper_2$	$a_1 = 0.2$	$a_4 = 0.18$	$a_1 = 0.2$	$(a_1 + a_3 + a_4)/3 = 0.167$

In this example, we found that the papers’ author ranking scores are different than their AuthorRank. We verify the effectiveness of publication ranking by AuthorRank in the following.

3.5. AuthorRank Combined with PRP

We have introduced two methods for publication ranking, PRP and AuthorRank. The former, PRP, depends on links in the graph to calculate the authority of each node. The latter, AuthorRank, is decided by author ranking. Ideally, we would recommend publications that have both a high AuthorRank and a high PageRank, meaning that they are really important for the topic. In contrast, papers with a low AuthorRank and a low PageRank have little importance for the topic.

In this part, we would like to use two methods to combine the results of PRP and AuthorRank to verify whether the performance of publication ranking can be improved. For these methods, we used AuthorRank to enhance the publication prior probability via publication prior probability smoothing.

Linear Combination:

$$I_{key_t}(p_i) = \alpha * I_{key_t}(p_i|PR) + (1 - \alpha) * I_{key_t}p_i|AR \tag{13}$$

where α is a parameter between 0 and 1, which controls the weight of $I_{key_t}(p_i|PR)$ and $I_{key_t}p_i|AR$. $I_{key_t}(p_i|PR)$ is the relative importance score for $paper_i$ on $topic_t$ calculated by the PRP algorithm, whereas $I_{key_t}p_i|AR$ is the importance ranking score for $paper_i$ on $topic_t$ generated by the AuthorRank algorithm.

The linear form assumes that the total score is a linear combination of the two scores. Each score's contribution to the total score is controlled by a parameter, α .

Cobb–Douglas Form:

$$I_{key_t}(p_i) = I_{key_t}(p_i|PR)^\alpha * I_{key_t}p_i|AR^{(1-\alpha)} \tag{14}$$

This form assumes that the total score is the product of the two scores, rendering it insensitive to small scores but sensitive to large scores.

As we already know, one limitation of the topic-based PageRank algorithm is that the importance scores of papers, $I_{key_m}(p_i|PR)$, are 0, if the paper is not related to the specific topic (key_t).

This may limit the performance of the harmonic form. Thus, if $I_{key_m}(p_i|PR)$ is 0, the final score of $I_{key_m}(p_i)$ is infinite. One smoothing method will improve the score of $I_{key_m}(p_i|PR)$, as follows:

$$I_{key_t}(p_i|PR)' = \sigma \cdot I_{key_t}(p_i|PR) + (1 - \sigma) \cdot P(z_{key_t}|Corpus) \tag{15}$$

where $I_{key_t}(p_i|PR)'$ is the score after smoothing, and $P(z_{key_t}|Corpus)$ is always larger than 0, the paper topic score is also always positive. The parameter, σ , controls the amount of smoothing. In this research, we used as a tentative value, $\sigma = 0.8$.

3.6. Evaluation Methods

In order to verify whether AuthorRank can replace PRP, that is, whether AuthorRank can improve on PRP's performance, and to determine which methods yield the best performance, we compared the results with several baseline approaches. The original, topic-insensitive PageRank algorithm was the first baseline. The other was based on PageRank with Priors (PRP). Two indicators were used in this paper to measure algorithm performance: mean average precision (MAP), and normalized discounted cumulative gain (nDCG) [30].

We clearly understand that it is difficult to obtain the "ground truth" for this experiment dataset, so we tried to use review or survey papers to find the most important publications for a specific scientific keyword. To achieve this goal, a list of review or survey papers along with their cited papers was collected. Collected review papers were screened so that they only focused on one topic (keyword). We assumed that if a publication was cited by a review paper, and if this review paper concentrated on a keyword, key_i , then this publication was important for key_i . Since the degree of importance of cited papers

may be different, we used the number of citations (by a review paper) to characterize the importance. Thus, if a review paper for keyword key_i cited $paper_1$ twice and $paper_2$ once, then, $Importance_{key_i}(paper_1) = 2$ and $Importance_{key_i}(paper_2) = 1$. If a paper was not cited by the target review paper, then the importance of this paper for the target topic was 0. If a paper was cited more times by the review paper, then we assume its maximum importance was equal to 4.

4. Experiments

4.1. Experimental Data

The experimental data for this paper were derived mainly from the ACM digital library. We used 41,370 publications and 223,810 citations, where full text and citations were extracted from XML files. In these publications, there are 63,323 authors. Among them, 49,101 authors (accounting for 77.54% of all the sampled publications) have only one publication. The selected dataset is a sub-graph in the database, which is reasonable for graph mining.

In this graph, we extracted 28,013 publications' text, including titles, abstracts, and full text. For the other 9879 publications, whose full texts were not available in our database, we used the title and abstract from a metadata repository to represent the content of the paper. For the remaining 3479 publications, only the title was available.

We then wrote a list of regular expression rules to extract all possible citations from the paper's full text. A text window surrounding the target citation, ($-n$ words, $+n$ words), was used to infer the citation topic distribution via LLDA. Intuitively, n should be a small number, as nearby words should provide more accurate citation information. However, n should not be too small to minimize randomness. In this experiment, we used an arbitrary parameter setting, where $n = 150$. In a total of 223,810 references, we successfully identified 94,051 references. The Table 3 shows possible citation formats in publications.

Table 3. Format of references in publications.

No.	References Format
1 [num1]
2 [num1,num2]
3 [num1-num2]
4 [num1, num2-num3]

For training the Labeled-LDA topic model, we first sampled 10,000 publications (with full text) and used author-provided keywords as topic labels. For instance, this paper has six author-provided keywords. Thus, our LLDA training would have assumed that this paper is a multinomial distribution over these six topics.

For the sampled publications, we first used tokenization to extract words from the title, abstract, and publication full text, and then employed Snowball stemming to extract the root of the target word. If a keyword appeared less than 10 times in the selected publications, we removed it from the training topic space. After that, we trained an LLDA model with 3910 topics (keywords) on 46,010 single words (bag of words), as p_{l,w_1} . These topics were used to infer the publication and citation topic distribution.

4.2. Experimental Result

As proposed in the section on methods, the PRP combined Labeled-LDA topic model with full-text citation analysis measured the relative importance of vertices (papers) in the publication networks. The vertices (41,370 publications) and edges (223,810 citations) represented a topic distribution on 3910 topics. For each topic, the graph may be totally different than for others.

In Figure 2, the first graph is the complete graph with 41,370 vertices and 223,810 edges, while the second one is a sub-graph with 580 vertices and 1148 edges based on the topic "Information Retrieval" (i.e., the publications used "Information Retrieval" as a keyword

or “Information Retrieval” was found in the paper abstract by using greedy matching). The last graph with 3356 vertices and 6671 edges is an extended graph of the “Information Retrieval” graph, which means each node on the graph is directly or conditionally (cited by a directly relevant paper) related to “information retrieval”. The biggest node in these two graphs is “R_291008: A language modeling approach to information retrieval”, which means that the citation count is the biggest one in our dataset.

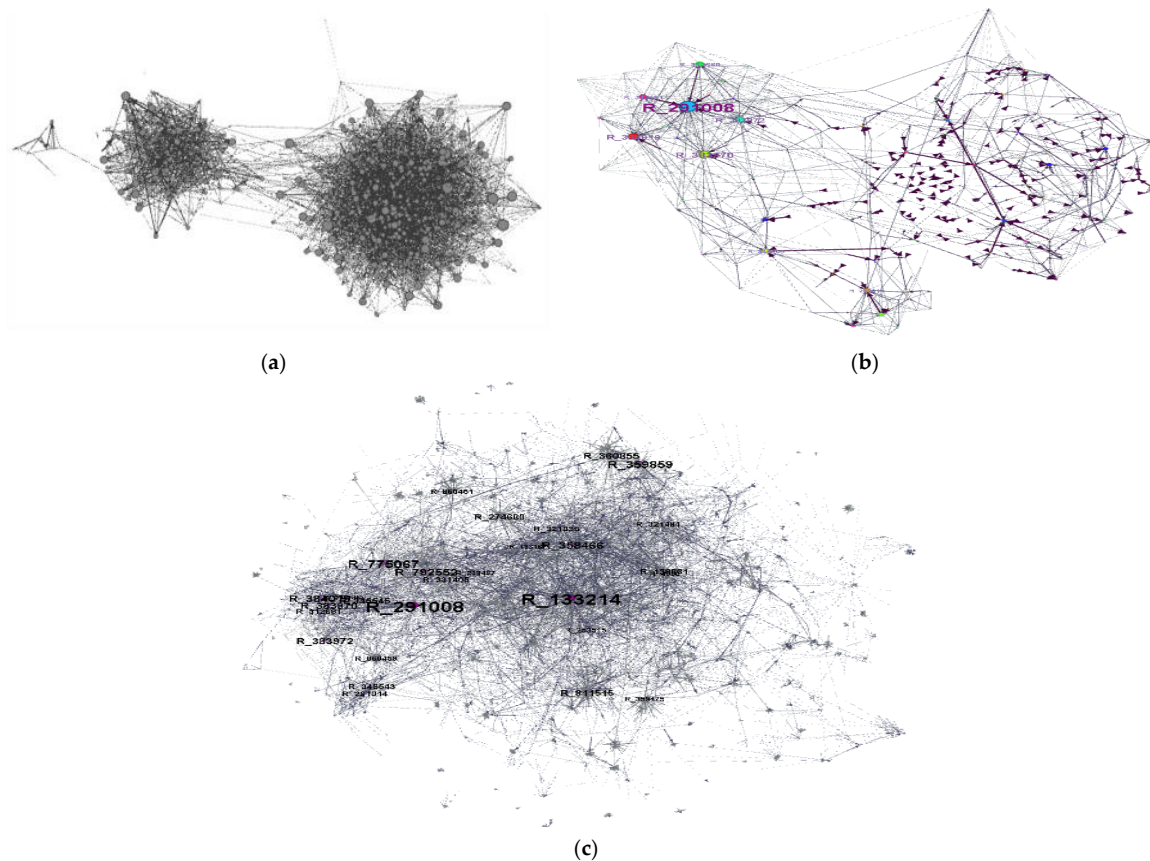


Figure 2. Publication graphs: (a) Visualization of complete graph (41,370 vertices and 223,810 edges). (b) Visualization of sub-graph on “Information Retrieval” (580 vertices and 1148 edges). (c) Visualization of extend graph on “Information Retrieval” (3356 vertices and 6671 edges).

In order to compare the different methods for publication ranking results, we took the topic “Information Retrieval” as our example and listed all the results by seven methods, as shown in Table 4 (top five results shown). PageRank is the original PageRank algorithm, where the damping factor is 0.85. PRP_1 is the PageRank with priors algorithm with only one prior: publication topic probability. PRP_1 neglects the other prior, the citation transition probability, thereby treating the scores of all the edges as equal. PRP_2 brings two kinds of prior knowledge into the citation graph by LLDA: publication topic probability and citation transition probability. First Author, Last Author, Max Author, and Average Author were proposed before as methods for calculating an AuthorRank.

Table 4. Publication ranking results.

Ranking Results	PageRank	PRP_1	PRP_2	First Author	Last Author	Max Author	Average Author
Top1	R_359342	R_321930	R_951766	R_321035	R_363817	R_363817	R_321084
Top2	R_362685	R_321757	R_321930	R_321084	R_3197	R_3197	R_321035
Top3	R_805047	R_358466	R_321035	R_511295	R_321084	R_321035	R_511295
Top4	R_805694	R_321035	R_358466	R_122864	R_322053	R_321084	R_363817
Top5	R_805695	R_951766	R_321074	R_1095451	R_321035	R_322053	R_109545

In this table, we found that ranking result lists are totally different by different methods. For example, the paper “R_321035: On Relevance, Probabilistic Indexing and Information Retrieval” authored by M.E. Maron and published in the “Journal of the ACM” in 1960, was ranked high according to both PageRank and AuthorRank, but there also exist papers with high AuthorRank but low PageRank or, conversely, low AuthorRank and high PageRank. The former may indicate a new publication resource, one the importance of which has not yet been widely recognized. The latter might be an important paper by a young or relatively unknown author.

4.3. Experimental Evaluation

(1) If AuthorRank can replace PageRank for publication ranking

Based on our previous work [5], the PRP algorithm is better than other classical methods for publication ranking. In this paper, we tried to improve publication ranking by using AuthorRank to inform a topic-based PageRank algorithm. AuthorRank scores were computed by author ranking via different four methods: First Author, Last Author, Max Author, and Average Author. For evaluation, 104 topics were selected associated with review or survey papers. For each topic, we calculated each publication’s ranking score in the dataset by the methods: PageRank, PRP_1, PRP_2, First Author, Last Author, Max Author, and Average Author.

In Tables 5 and 6, we found that the best result was generated by PRP_2 both for MAP and nDCG, which verified that AuthorRank cannot replace the PageRank algorithm for publication ranking. We also found that topic-based PageRank (PRP_1) can significantly improve on the original PageRank algorithm, and that PRP_2 (which included the citation topic distribution) outperformed PRP_1 (which used the publication topic distribution only), especially for nDCG. Among AuthorRank results, we found the method of Average Author to be the best. The average contribution from all authors in an article is a more accurate guide to a paper’s AuthorRank than the other methods. However, the evaluation shows that AuthorRank cannot simply replace PageRank.

Table 5. Methods based on PageRank and AuthorRank (MAP).

	PageRank	PRP_1	PRP_2	First Author	Last Author	Max Author	Average Author
MAP@10	0.0168	0.2427	0.2238	0.1314	0.104	0.0953	0.1172
MAP@30	0.0192	0.1909	0.1933	0.1208	0.106	0.1031	0.1191
MAP@50	0.0186	0.1782	0.1804	0.1131	0.096	0.1021	0.1077
MAP@100	0.0182	0.1521	0.1619	0.0957	0.0681	0.08	0.0902
MAP@1000	0.011	0.106	0.1199	0.0519	0.0383	0.0407	0.0541
MAP@all	0.0037	0.091	0.0996	0.0382	0.0279	0.0288	0.0401

Table 6. Methods based on PageRank and AuthorRank (nDCG).

	PageRank	PRP_1	PRP_2	First Author	Last Author	Max Author	Average Author
nDCG@10	0.0093	0.1181	0.1229	0.0509	0.0427	0.0392	0.0593
nDCG@30	0.0076	0.1323	0.1451	0.063	0.0478	0.0478	0.0715
nDCG@50	0.0084	0.1482	0.1621	0.0773	0.0575	0.0595	0.0835
nDCG@100	0.0107	0.1767	0.1847	0.0963	0.0761	0.0768	0.1037
nDCG@1000	0.0392	0.2564	0.2738	0.1816	0.1635	0.1634	0.1893
nDCG@all	0.1904	0.3341	0.3445	0.2761	0.2619	0.2614	0.2816

(2) Whether AuthorRank can improve publication ranking results

To test whether the AuthorRank results can improve publication ranking results (as calculated by topic-based PageRank), we then used the combination methods evaluation results to compare with the PRP_2 results (best-performed method without Author Rank).

Linear combination:

For each AuthorRank method (First Author, Last Author, Max Author, or Average Author), the parameter α , which controls the relative contributions of a PRP and AuthorRank to a publication’s topical importance score, was trained from 0 to 1 with a step of 0.1.

Tables 7 and 8 display results by the linear combination method for AuthorRank in informing publication ranking. The best result in the training results for each method is shown in the tables.

Table 7. Linear Combination results (MAP).

	PRP_2	First Author	Last Author	Max Author	Average Author
		$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.9$
MAP@10	0.2238	0.2271	0.2265	0.2266	0.2286
MAP@30	0.1933	0.1948	0.1943	0.1941	0.1935
MAP@50	0.1804	0.1812	0.1814	0.1805	0.1818
MAP@100	0.1619	0.1623	0.1626	0.1626	0.1627
MAP@1000	0.1159	0.1172	0.1088	0.1158	0.1182
MAP@all	0.0996	0.1002	0.1002	0.1005	0.1007

Table 8. Linear Combination results (nDCG).

	PRP_2	First Author	Last Author	Max Author	Average Author
		$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.9$
nDCG@10	0.1229	0.1254	0.1248	0.1243	0.1248
nDCG@30	0.1451	0.1471	0.1456	0.1454	0.1463
nDCG@50	0.1621	0.1639	0.1638	0.1646	0.165
nDCG@100	0.1847	0.1868	0.1872	0.1865	0.1867
nDCG@1000	0.2738	0.2743	0.2739	0.2778	0.2776
nDCG@all	0.3445	0.3456	0.3456	0.3467	0.3469

As shown in the above tables, when the parameter $\alpha = 0.9$ in each method, AuthorRank combined with PageRank received the best result, meaning that AuthorRank can improve ranking results, but cannot replace the traditional link analysis ranking algorithm (PageRank with Priors). For MAP@n, the First Author method was better than the others when $n \leq 30$, but for $n \geq 50$, the Average Author method was better than all of the others.

nDCG@n is a more important indicator in this research, for it tells the degree of (publication topic) importance. If an nDCG score is large, the target algorithm can prioritize the most important on the ranking list. In the tables, it is clear that the First Author method is better than the others when $n < 30$, but for $n > 1000$, the Average Author method is the best one.

We also used significance testing to compare each method with the baseline PRP_2, and $t < 0.001$.

Cobb–Douglas Form:

This combination method is insensitive to small scores and sensitive to large scores. The parameter, α , was trained from 0 to 1 with a step of 0.1. The best results for each method are shown in the following tables.

For MAP@n in Table 9, the Max Author method was better than the others when $n \leq 10$, and Average Author was the best when $30 < n < 50$ and $n > 1000$, while for $100 < n < 100$, the First Author method was better than the others.

Table 9. Cobb–Douglas Combination results (MAP).

	PRP_2	First Author	Last Author	Max Author	Average Author
		$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.8$	$\alpha = 0.9$
MAP@10	0.2238	0.2309	0.2248	0.2314	0.2306
MAP@30	0.1933	0.1953	0.1949	0.1966	0.1975
MAP@50	0.1804	0.182	0.1809	0.1828	0.1845
MAP@100	0.1619	0.1642	0.1637	0.163	0.1638
MAP@1000	0.1199	0.1213	0.1199	0.1204	0.1209
MAP@all	0.0996	0.101	0.1009	0.1006	0.1013

It is clear in Table 10 that Average Author is always better than all the other baseline methods, especially for the nDCG indicator.

Table 10. Cobb–Douglas Combination results (nDCG).

	PRP_2	First Author	Last Author	Max Author	Average Author
		$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.8$	$\alpha = 0.9$
nDCG@10	0.1229	0.1261	0.1257	0.1289	0.1319
nDCG@30	0.1451	0.149	0.148	0.1489	0.1499
nDCG@50	0.1621	0.1653	0.1641	0.1652	0.1666
nDCG@100	0.1847	0.1871	0.1857	0.1881	0.1902
nDCG@1000	0.2738	0.2764	0.2765	0.28	0.2813
nDCG@all	0.3445	0.3465	0.346	0.3472	0.3486

5. Conclusions

In this paper, we aim to test whether and how topical AuthorRank can replace or enhance classical PageRank for publication ranking. From the results of our experiment, we can conclude that:

(1) AuthorRank cannot replace PageRank for publication ranking. This conclusion is supported by the results from Tables 5 and 6. We also found that PRP with two priors, publication topic probability and citation transition probability, outperforms significantly the original topic-insensitive algorithm of PageRank, and is better than PRP with only the publication topic probability.

(2) AuthorRank can improve publication ranking results, it also proves that the article written by influential authors often deserves a higher ranking in information retrieval. When we combine the results of PRP and AuthorRank by linear combination method and Cobb–Douglas combination, the results calculated by MAP and nDCG are better than PRP without AuthorRank.

(3) By comparing the linear combination method with the Cobb–Douglas combination method, we found that calculating AuthorRank results by Average Author is the best method for improving publication ranking. This conclusion is supported by Tables 7–10. Although the Cobb–Douglas combination method for AuthorRank and PRP is better than the linear combination method, this advantage is not significant. We also found that AuthorRank is effective for assessing the importance of publications where content or citation metadata is missing or partially missing. When we do not have publication content information, we cannot use topic modeling to infer the topic distribution, but AuthorRank can still help us to estimate the prior probability of these papers.

Author Contributions: Conceptualization: X.L.; methodology: X.L.; validation: J.Z.; formal analysis: J.Z.; data curation: J.Z.; writing—original draft preparation: J.Z.; writing—review and editing: X.L.; visualization: J.Z.; supervision: X.L.; project administration: X.L.; funding acquisition: J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Funds for the Central Universities, grant number 3132022289, the Liaoning Revitalization Talents Program, grant number XLYC1907084, the China Postdoctoral Science Foundation, grant number 2016M591421.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Garfield, E. Citation indexes for science: A new dimension in documentation through association of ideas. *Science* **1955**, *122*, 108–111. [CrossRef]
- Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Stanford InfoLab: Stanford, CA, USA, 1999.
- Haveliwala, T.H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.* **2003**, *15*, 784–796. [CrossRef]
- Huang, X.; Chen, C.A.; Peng, C.; Wu, X.; Fu, L.; Wang, X. Topic-sensitive influential paper discovery in citation network. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, Australia, 15–18 May 2018; Springer: Cham, Switzerland, 2018; pp. 16–28.
- Liu, X.; Zhang, J.; Guo, C. Full-text citation analysis: Enhancing bibliometric and scientific publication ranking. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, HI, USA, 29 October–2 November 2012; pp. 1975–1979.
- Ramage, D.; Hall, D.; Nallapati, R.; Manning, C.D. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 248–256.
- Ahmedi, L.; Halilaj, L.; Sejdiu, G.; Bajraktari, L. Ranking Authors on the Web: A Semantic AuthorRank. In *Social Networks: Analysis and Case Studies*; Springer: Vienna, Austria, 2014; pp. 19–40.
- Amjad, T.; Daud, A.; Che, D.; Akram, A. MulCE: Mutual influence and citation exclusivity author rank. *Inf. Process. Manag.* **2016**, *52*, 374–386. [CrossRef]
- Amodio, P.; Brugnano, L.; Scarselli, F. Implementation of the PaperRank and AuthorRank indices in the Scopus database. *J. Informetr.* **2021**, *15*, 101206. [CrossRef]
- Gross, P.L.K.; Gross, E.M. College libraries and chemical education. *Science* **1927**, *66*, 385–389. [CrossRef] [PubMed]
- Garfield, E. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science* **1972**, *178*, 471–479. [CrossRef]
- Stern, D.I. High-ranked social science journal articles can be identified from early citation information. *PLoS ONE* **2014**, *9*, e112520. [CrossRef]
- An, J.; Kim, N.; Kan, M.-Y.; Chandrasekaran, M.K.; Song, M. Exploring characteristics of highly cited authors according to citation location and content. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 1975–1988. [CrossRef]
- Liu, S.; Chen, C.; Ding, K.; Wang, B.; Xu, K.; Lin, Y. Literature retrieval based on citation context. *Scientometrics* **2014**, *101*, 1293–1307. [CrossRef]
- Yan, E. Topic-based Pagerank: Toward a topic-level scientific evaluation. *Scientometrics* **2014**, *100*, 407–437. [CrossRef]
- Nykl, M.; Ježek, K.; Fiala, D.; Dostal, M. PageRank variants in the evaluation of citation networks. *J. Informetr.* **2014**, *8*, 683–692. [CrossRef]
- Dangur, I.; Bekkerman, R.; Minkov, E. Identification of topical subpopulations on social media. *Inf. Sci.* **2020**, *528*, 92–112. [CrossRef]
- Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
- Tao, M.; Yang, X.; Gu, G.; Li, B. Paper Recommend Based on LDA and PageRank. In Proceedings of the International Conference on Artificial Intelligence and Security, Hohhot, China, 17–20 July 2020; Springer: Singapore, 2020; pp. 571–584.
- Gollapalli, S.D.; Li, X. Using PageRank for characterizing topic quality in LDA. In Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, Tianjin, China, 14–17 September 2018; pp. 115–122.
- Kohn, A.J. Author Rank. 2012. Available online: <http://www.blindfiveyearold.com/author-rank> (accessed on 15 July 2017).
- Onodera, N.; Yoshikane, F. Factors affecting citation rates of research articles. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 739–764. [CrossRef]
- Waheed, W.; Imran, M.; Raza, B.; Malik, A.K.; Khattak, H.A. A hybrid approach toward research paper recommendation using centrality measures and author ranking. *IEEE Access* **2019**, *7*, 33145–33158. [CrossRef]
- Silva, J.; Aparício, D.; Silva, F. Otarios: Optimizing author ranking with insiders/outside subnetworks. In Proceedings of the International Conference on Complex Networks and their Applications, Cambridge, UK, 11–13 December 2018; Springer: Cham, Switzerland, 2018; pp. 143–154.
- Balog, K.; Fang, Y.; De Rijke, M.; Serdyukov, P.; Si, L. Expertise retrieval. *Found. Trends Inf. Retr.* **2012**, *6*, 127–256. [CrossRef]

26. Gonçalves, R.; Dorneles, C.F. Automated expertise retrieval: A taxonomy-based survey and open issues. *ACM Comput. Surv.* **2019**, *52*, 1–30. [[CrossRef](#)]
27. Zhang, J.; Guo, C.; Liu, X. Topic Based Author Ranking with Full-Text Citation Analysis. In Proceedings of the Asia Information Retrieval Symposium, Tianjin, China, 17–19 December 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 477–485.
28. Strumia, A.; Torre, R. Biblioranking fundamental physics. *J. Informetr.* **2019**, *13*, 515–539. [[CrossRef](#)]
29. Zhang, J.; Liu, X. Full-text and topic based authorrank and enhanced publication ranking. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital libraries, Indianapolis, IN, USA, 22–26 July 2013; pp. 393–394.
30. Järvelin, K.; Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* **2002**, *20*, 422–446. [[CrossRef](#)]