BibRank: a language-based model for co-ranking entities in bibliographic networks

Laure Soulier, Lamjed Ben Jabeur, Lynda Tamine, Wahiba Bahsoun

BibRank: a Language-Based Model for Co-Ranking Entities in Bibliographic Networks

Laure Soulier
IRIT - Paul Sabatier University
118 route de Narbonne
31062 Toulouse, France
soulier@irit.fr

Lamjed Ben Jabeur
IRIT - Paul Sabatier University
118 route de Narbonne
31062 Toulouse, France
jabeur@irit.fr

Lynda Tamine
IRIT - Paul Sabatier University
118 route de Narbonne
31062 Toulouse, France
tamine@irit.fr

Wahiba Bahsoun
IRIT - Paul Sabatier University
118 route de Narbonne
31062 Toulouse, France
wbahsoun@irit.fr

ABSTRACT

Bibliographic documents are basically associated with many entities including authors, venues, affiliations, etc. While bibliographic search engines addressed mainly relevant document ranking according to a query topic, ranking other related relevant bibliographic entities is still challenging. Indeed, document relevance is the primary level that allows inferring the relevance of the other entities regardless of the query topic. In this paper, we propose a novel integrated ranking model, called BibRank, that aims at ranking both document and author entities in bibliographic networks. The underlying algorithm propagates entity scores through the network by means of citation and authorship links. Moreover, we propose to weight these relationships using content-based indicators that estimate the topical relatedness between entities. In particular, we estimate the common similarity between homogeneous entities by analyzing marginal citations. We also compare document and author language models in order to evaluate the level of author's knowledge on the document topic and the document representativeness of author's knowledge. Experiment results on the representative CiteSeerX dataset show that BibRank model outperforms baseline ranking models with a significant improvement.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Multi-entity Ranking, Bibliographic Network, Heterogeneous Information Network, Homogeneous Information Network

1. INTRODUCTION

Bibliographic indexers and scientific search engines, such as CiteSeerX[1] and Google Scholar[2] have largely contributed to develop sophisticated models for bibliographic access and scientific document ranking. For this purpose, several features such as citation links and document recency are used in order to rank bibliographic resources. Meanwhile, these approaches remain document-centered and do not consider other entities participating in scientific activities.

From a larger point of view, the bibliographic resources involve several entities comprising documents, authors, venues, research institutions, etc. In addition to the traditional keyword-based search that aims at retrieving documents discussing a specific topic, scientists express usually their need for accessing to relevant information related to other different bibliographic entities. For instance, they are interested in finding authors working on a specific topic, relevant venues to their research activity and specialized institutions in some research area. These information needs reveal a new information retrieval task that consists on ranking jointly different bibliographic entities considering a particular topic.

To tackle this problem, previous approaches in the area of literature access propose to either estimate the relevance of a bibliographic entity from related ones or jointly rank the multiple types of entities that interact with each other. In the first category of approaches, some work propose to estimate the importance of an author from his/her co-authorship links [18]. Likewise, the importance of documents can be estimated from their corresponding authors [11, 15]. The second category of approaches models bibliographic entities using heterogeneous networks and ranks entities according to their relationships to both similar and different types of entities [23, 27, 28, 30, 31]. This category differs from the first one by producing multiple ranking sets

[1]http://www.citeseer.ist.psu.edu/
[2]http://www.scholar.google.com/

as much as the number of entity types are embedded within the network.

In this paper, we are mainly interested in ranking both documents and authors considering a specific topic. We propose a novel ranking model that jointly ranks author and document entities according to their relationships. This model integrates content-based features to evaluate the topical relatedness between connected entities. In particular, a language model is used in order to estimate the author's knowledge on document topic and the document representativeness of author's knowledge.

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 introduces problem definition and used notations. We present in section 4 our BibRank algorithm for co-ranking bibliographic entities and we detail the proposed link-based and content-based indicators. Section 5 describes experiments carried out using CiteSeerX dataset and discusses the obtained results. In section 6, we conclude the paper and outline research directions for future work.

2. RELATED WORK

A bibliographic information retrieval system aims at retrieving and ranking bibliographic resources considering a particular topic. Previous work in the domain have investigated bibliometric indicators in order to evaluate the scientific quality of documents and therefore rank the search results by their importance [8, 10, 25]. These indicators help to evaluate document quality [25] as well as the prestige of corresponding authors [2, 6, 10, 13, 29] and journals [3, 8, 24]. In this context, the importance of a document is derived from related entities such as authors, journals, venues, etc. These entities mutually reinforce the quality of each other.

Having the meaningful relationships shown between the different bibliographic entities, some work propose to represent bibliographic resources using bibliographic networks. Such representation allows modeling the different relationships between entities. Two categories of bibliographic network models are proposed in the literature. The first category relies on representing bibliographic resources using a *homogenous network* including only one type of entities. The network model introduced in [16, 17, 20] represents documents and citation links between them. Work in [18] propose to model the bibliographic resources with a co-authorship network. The second category of models represents bibliographic resources using a *heterogeneous network* involving several entities. Zang et al [30] propose to include documents, authors and venues in the network. These entities are associated to both similar and different entities in the network.

Based on bibliographic network models introduced above and aiming at evaluating the scientific quality of bibliographic entities, related work have proposed alternative indicators for entity ranking. These indicators take into account the entity position in the network. In the case of homogeneous networks, link analysis algorithms such as PageRank [20], HITS [16] and Salsa [17] have been investigated for computing authority indicators for bibliographic entities [5, 7]. In the case of heterogeneous networks, previous approaches propose to either rank one type of entities by considering its relationships or jointly rank several types of entities. The first category of approaches are called *mono-type entity ranking approaches in heterogeneous bibliographic networks*. The second category is known as *multi-type entity ranking approaches in heterogeneous bibliographic networks*.

Focusing on mono-type entity ranking approaches in heterogeneous networks, previous work propose to rank bibliographic entities based on their topical relevance in response to the query and also according to their position in the network. In this context, bibliographic networks have been addressed from a social network point of view where documents are ranked by combining their topical relevance and the social importance of their corresponding authors [11, 15]. Liu et al. [18] propose to apply a revisited PageRank on a co-authorship network, where edges are weighted by co-authorship frequency, to evaluate the impact of an author in the network.

In comparison to mono-type entity ranking approaches in heterogeneous bibliographic networks, multi-type entity ranking approaches in the same networks jointly rank different types of entities based on the possible relationships between them. Arnetminer[3], Microsoft Academic Search[4] and Rexa[5] are typical examples of search engines that rank at least two types of entities, generally authors and documents. Related work for multi-type entity ranking models propose to either use link analysis approaches [27, 28, 31] or integrate topical indicators to represent entity-query similarity [23, 30].

Considering the link analysis for multi-type entity ranking approaches, Yan and Din [27] propose to jointly rank documents, authors and journals presuming that important authors publish important documents and important entities cite important ones. This model computes entity scores based on the relationships between them by propagating connected entity scores. Document scores are computed differently. First, an initial score is attributed to each document based on the score of other connected entities. Afterwards, a PageRank score is computed on the homogeneous network of documents. Yang et al. [28] propose to extend the Topical PageRank Model, previously introduced in [19], in order to highlight topical authorities in the bibliographic network. This algorithm propagates scores through the network by simulating the surfer behavior. Three actions are considered: stay on the same topic *"Follow-stay"*, move to a different topic *"Follow-jump"* or move to a random topic *"Jump-jump"*. In the same line, Zhou et al. [31] propose to rank jointly different types of entities by combining a PageRank score, expressing the authority of an entity in the homogeneous network, and a *BiWalk* score that emphasizes the authority of an entity through inter-graph relationships.

Beside the previous link analysis-based approaches, some work propose to integrate topical-based indicators in order to rank entities in a bibliographic network. Zang et al. [30] propose to recommend entities in the bibliographic network by combining two scores. The first, namely the topical score, is computed with a language model-based information retrieval framework [21]. The second score evaluates the authority of the entity in the heterogeneous network using a PageRank-like algorithm. *"Author-Topic-Conference"* (ACT) model, proposed by Tang et al. [23] and used in Arnetminer search engine, represents entities thanks to topic distributions inferred by the Latent Dirichlet Allocation

[3]http://www.arnetminer.org/
[4]http://www.academic.research.microsoft.com/
[5]http://www.rexa.info/

(LDA) model [26]. The ACT model ranks authors, documents and venues using a PageRank-like algorithm.

In this paper, we propose a novel ranking algorithm for heterogeneous bibliographic networks. In comparison to previous related work, our approach is different in at least three respects:

- First, we propose an integrated approach for ranking bibliographic entities unlike modular approaches that combine distinct relevance scores [11, 15, 30, 31]. The aim of the integrated approach is to consider, at the same time and within a unified process, both topical and link-based features.
- Second, our approach is based on content-based and link analysis features unlike previous work [27, 28, 31] that consider only link-based indicators. We propose in this work a ranking model that estimates, by means of a language model, the author's knowledge on the document topic and the document representativeness to author's knowledge. We also propose to consider marginal citations to discredit non topical relationships unlike work presented in [23] that models entities through topical distributions.
- Third, the core idea of our approach relies on the use of a language model-based approach to estimate the strength and/or weakness of the topical relationship between the bi-type entities jointly with the link-based relatedness. In contrast, [30] focuses on the link-based relatedness between entities and the topical feature is only addressed at the query level.

3. PROBLEM DEFINITION AND NOTATIONS

In this work, we propose to model bibliographic resources using a bi-type network that integrates documents and authors of scientific publications. In response to a user query Q, each type of entities is ranked according to a relevance score computed using a co-ranking function called *BibRank*. Below, we introduce some definitions and notations.

Definition 1. Document: A scientific document, noted d_i, is represented by a weighted vector of terms $\vec{d_i} = (w_{i1}, \ldots, w_{ip}, \ldots, w_{iT})$. w_{ip} denotes the weight of the p^{th} term occurring in document d_i. T is the number of indexed terms.

Definition 2. Author: A scientific author a_j may publish one or more documents. Let $\mathcal{D}(a_j)$ be the set of documents authored by $\vec{a_j}$. An author a_j is represented by a weighted vector including terms present in all his/her documents $\vec{a_j} = \sum_{d_i \in \mathcal{D}(a_j)} \vec{d_i}$.

Definition 3. Bi-type bibliographic network: Bibliographic resources are represented by a directed graph $G = (V, E)$ where nodes $V = A \cup D$ denote bibliographic entities with D and A corresponding to document set $D = \{d_1, ..., d_n\}$ and author set $A = \{a_1, ..., a_m\}$ respectively. The set of edges $E : V \times V$ represents bibliographic relationships between entities. A directed edge from $x \in V$ to $y \in V$ is represented by the pair $e = (x, y)$.

Definition 4. Document subgraph and author subgraph: Document and author subgraphs are directed graphs which include the subset of homogeneous nodes represented in the bi-type bibliographic network G. The document subgraph $G_D = (D, D \times D) \in G$ represents bibliographic documents and possible relationships between them. The author subgraph $G_A = (A, A \times A) \in G$ represents authors of bibliographic documents and all possible relationships between them.

Definition 5. Bibliographic relationships: In a bi-type bibliographic network and homogeneous related subgraphs, relationships are represented by edges that model semantic links between entities. Three types of relationships are identified:

Document citation associations: Two documents d_i and d_k are connected by a document citation association if document d_i *cites* document d_k. This relationship is modeled by an edge $(d_i, d_k) \in D \times D$.

Author citation associations: Two authors a_j and a_l are connected by an author citation association if one of the publications of author a_j *cites* at least one of the publications of author a_l. This relationship is modeled by and edge $(a_j, a_l) \in A \times A$.

Authorship associations: An author a_j and a document d_i are connected by an authorship association if author a_j *has authored* document d_i. This relationship is established in both directions and is modeled by a couple of edges $(d_i, a_j) \in D \times A$ and $(a_j, d_i) \in A \times D$.

Definition 6. Intra-graph and inter-graph relationships: A relationship (x, y) is called *intra-graph* association if both source node $x \in X$ and target node $y \in X$ are included in the same subgraph G_X where $X \in \{A, D\}$. Document and author citation associations are intra-graph relationships.
A relationship (u, v) is called *inter-graph* association if source node $u \in X$ and target node $v \in Y$ belong to two different subgraphs G_X and G_Y with $X \in \{A, D\}$, $Y \in \{A, D\}$ and $X \neq Y$. Authorship associations are inter-graph relationships.

Figure 1 presents an example of a bibliographic network and illustrates a graphical representation of the extracted bi-type bibliographic network.

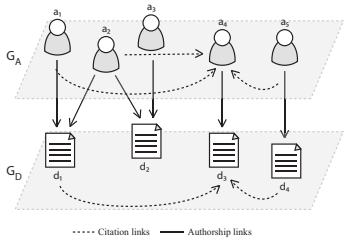


Figure 1: Bi-type bibliographic graph G

Definition 7. Entity Rank: A *Relevance Status Value* $RSV(d_i, Q)$ evaluates the document $d_i \in D$ relevance considering a query Q. This function measures the query-document similarity. Extending this notation, an entity-query similarity between a bibliographic entity $x \in A \cup D$ and a query Q is represented by a *Relevance Status Value*

$RSV(x, Q)$. This function ranks entities according to a relevance score. Let $rank(x)$ be the rank of entity x. The reciprocal rank is defined as follows:

$$r_x = \frac{1}{rank(x)} \tag{1}$$

Definition 8. BibRank function
Considering a bi-type bibliographic network $G = (V, E)$ and a query Q, BibRank function co-ranks author and document entities. A ranked list is produced for each type of entities. BibRank function is defined as follows:

$$\begin{gathered} BibRank : \{Q, G\} \longrightarrow \{R_A, R_D\} \\ \forall a_j \in A,\ 0 < R_A(a_j) < 1, \sum_{a_j \in A} R_A(a_j) = 1 \\ \forall d_i \in D,\ 0 < R_D(d_i) < 1, \sum_{d_i \in D} R_D(d_i) = 1 \end{gathered} \tag{2}$$

with R_D is the set of document scores. R_A is the set of author scores. $R_D(d_i)$ associates to each document d_i the corresponding score in R_D. Respectively, $R_A(a_j)$ associates to each author a_j the corresponding score in R_A.

4. BIBRANK ALGORITHM

In this section, we present our BibRank algorithm which computes a co-relevance score for both document and author entities of a bi-type bibliographic network. We notice that basic assumptions introduced in [28, 30, 31] are also maintained by our algorithm, particularly the assumption that important entities are cited by important ones. Indeed, BibRank algorithm introduces transition probabilities between the two homogeneous subgraphs and propagates relevance scores through connected entities. In addition to transition probabilities, content-based indicators are defined between connected entities to model the topical relatedness between them. Figure 2 illustrates transition probabilities and content-based indicators in the bi-type bibliographic network. These features are detailed in the next sub-sections.

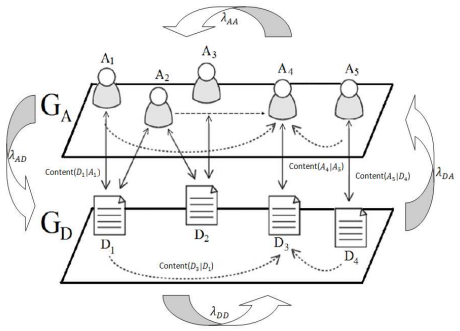


Figure 2: Transition probabilities and content-based indicator in a bi-type bibliographic network

4.1 Computing Transition Probabilities between Homogeneous Document and Author Subgraphs

Transition probabilities measure the possibility that a surfer moves from a subgraph to another one. Accessing a document node, the surfer has two choices: either stay on the same subgraph or move to the author subgraph. Therefore, we estimate the probability of moving from a subgraph to another one based on the current position of the surfer. We compute the transition probability from a subgraph G_X with $X \in \{A, D\}$ to a subgraph G_Y with $Y \in \{A, D\}$ as follows:

$$\begin{gathered} \lambda_{XY} = \frac{|\{\forall (x,y) \in X \times Y\}|}{|E|} \\ \lambda_{AD} + \lambda_{AA} = 1 \\ \lambda_{DA} + \lambda_{DD} = 1 \end{gathered} \tag{3}$$

where $|E|$ is the number of edges in the bibliographic network. The transition probabilities of staying on the same subgraph G_X with $X \in A \cup D$ is represented by λ_{XX}. Four types of transition probabilities are identified in the bi-type bibliographic network: λ_{AA}, λ_{DD}, λ_{DA} and λ_{AD}. The intra-graph transition probabilities are represented by λ_{AA} and λ_{DD}. The inter-graph transition probabilities are represented by λ_{AD} and λ_{DA}.

4.2 Computing Content-Based Indicators

The content-based indicators measure the topical relatedness between two connected entities. Our bi-type bibliographic network is characterized by two types of relationships: inter-graph and intra-graph relationships. Accordingly, two content-based indicators are proposed. A first score is attributed to inter-graph relationships based on the topical relatedness between entities. This score is measured by a language model indicator that expresses, depending on the sense of the bi-directed edge, the document representativeness of author's knowledge and the author's knowledge on the document topic. Second, we attribute a content-based score for intra-graph relationships that estimates the significance of citation links and detects marginal citations. This score expresses the common interest of the two connected entities that they address regarding the query topic.

4.2.1 Document Representativeness of Author's Knowledge

For each author a_j, a content-based score is attributed to his/her authored document $d_i \in \mathcal{D}(a_j)$. This score determines the likelihood for document d_i to be representative for the author's knowledge. The content-based score $Content(d_i|a_j)$ from author a_j to document d_i is computed as follows:

$$Content(d_i|a_j) = \frac{P(a_j|M_{d_i})}{\max\limits_{\forall (a_l, d_k) \in A \times D} P(a_l|M_{d_k})} \tag{4}$$

$P(a_j|M_{d_i})$ is the probability of observing author a_j considering the language model M_{d_i} of document d_i. This probability is computed using the Jelineck-Mercer formula [12]:

$$P(a_j|M_{d_i}) = \prod_{t \in \vec{a_j}} [(1-\lambda)P(t|M_{a_j}) + \lambda P(t|M_{d_i})]^{n(t,\vec{a_j})} \tag{5}$$

$n(t, \vec{a_j})$ is the number of times that term t appears in $\vec{a_j}$. $P(t|M_{a_j})$ is the probability of observing term t having the language model M_{a_j} of author a_j. $P(t|M_{d_i})$ is the probability of observing term t having the language model M_{d_i} of document d_i.

For convenience, the probability $P(t|M_x)$ of a term t according to language model of entity $x \in A \cup D$ is computed as follows:

$$P(t|M_x) = \frac{tf(t,x)}{|x|} \tag{6}$$

where $tf(t,x)$ is the frequency of term t in x and $|x|$ is the number of terms included in x.

4.2.2 Author's Knowledge on the Document Topic

For a document d_i, we note the set of corresponding authors $\mathcal{A}(d_i)$. The likelihood of each author $a_j \in \mathcal{A}(d_i)$ considering the topic of document d_i is computed by a content-based score $Content(a_j|d_i)$:

$$Content(a_j|d_i) = \frac{P(d_i|M_{a_j})}{\max\limits_{\forall(a_l,d_k)\in A\times D} P(d_k|M_{a_l})} \quad (7)$$

$P(d_i|M_{a_j})$ is the probability of observing document d_i according to the language model M_{a_j} of author a_j. This probability is computed as follows:

$$P(d_i|M_{a_j}) = \prod_{t\in\vec{d_i}} [(1-\lambda)P(t|M_c) + \lambda P(t|M_{a_j})]^{n(t,\vec{d_j})} \quad (8)$$

where the collection c is represented by the weighted vector of terms included in all documents of the collection $c = \sum_{d_i\in D} \vec{d_i}$. Let $n(t,\vec{d_i})$ be the number of times that term t appears in $\vec{d_i}$. $P(t|M_c)$ is the probability of observing term t having to the language model M_c of collection c. This probability is computed as follows:

$$P(t|M_c) = \frac{tf(t,c)}{|c|} \quad (9)$$

where $tf(t,c)$ is the frequency of term t in the collection c and $|c|$ is the number of terms included in the collection c.

4.2.3 Marginal Citations

Marginal citations are detected by semantic links between two homogeneous entities. We assume that the significance of a citation link can be estimated by the relevance of each entity considering the query topic. The content-based indicator detects semantically related entities for intra-graph relationships. For this aim, we propose to discredit thus marginal citations and emphasize significant citation links by comparing the rank of each cited and citing entity in the result set. The closer the ranks of entities are, more significant the citation link is. The common similarity between two entities x, y of a homogeneous subgraph G_X with $X \in \{D, A\}$ is measured by the closeness of entity ranks:

$$Sim_{com}(x,y) = \frac{1}{|rank(x) - rank(y)|} \quad (10)$$

Similarly to inter-graph relationships, a content-based score $Content(x|y)$ of entity x from entity y is computed for intra-graph relationships:

$$Content(x|y) = \frac{Sim_{com}(x,y)}{\max\limits_{(v,w)\in X\times X} Sim_{com}(v,w)} \quad (11)$$

4.3 Detailed Algorithm

BibRank is a PageRank-like algorithm that ranks jointly document and author entities using both topical-based and link-based features. BibRank algorithm is processed in three main steps. First, it attributes for all the entities in the homogeneous subgraph an initial score based on the total number of similar nodes. Then, it propagates entity scores through the graph proportionally to the transition probabilities defined in formula 3. At each iteration, a new score is attributed to document and author entities as the sum of the predecessor node scores. These scores are weighted using a topical-based feature, modeled by the reciprocal rank defined in formula 1, and content-based indicators, as presented in formula 4, 7 and 11. Finally, a ranked list is produced for each type of entities.

Algorithm 1: BibRank

Data: Q, $G = (V, E)$
Result: $BibRank : \{Q, G\} \longrightarrow \{R_D, R_A\}$
begin
  /* Step 1: Initialize */
  $\theta \leftarrow 0$
  $R_D^\theta(d_i) \leftarrow \frac{1}{|D|}$
  $R_A^\theta(a_j) \leftarrow \frac{1}{|A|}$
  /* Step 2: Propagate scores using transition probabilities, query input, graph structure and content-based scores */
  repeat
    /* Computing document scores */
    $R_D^{\theta+1}(d_i) \leftarrow \frac{d}{|V|} + (1-d)\times$
    $(\lambda_{AD} \sum\limits_{\substack{a_l;\\(a_l,d_i)}} \frac{R_A^\theta(a_l).w_{a_l}^{d_i}}{O(a_l)} + \lambda_{DD} \sum\limits_{\substack{d_k;\\(d_k,d_i)}} \frac{R_D^\theta(d_k).w_{d_k}^{d_i}}{O(d_k)})$
    $R_D^{\theta+1}(d_i) = Norm(R_D^{\theta+1}(d_i))$
    /* Computing author scores */
    $R_A^{\theta+1}(a_j) \leftarrow \frac{d}{|V|} + (1-d)\times$
    $(\lambda_{DA} \sum\limits_{\substack{d_k;\\(d_k,a_j)}} \frac{R_D^\theta(d_k).w_{d_k}^{a_j}}{O(d_k)} + \lambda_{AA} \sum\limits_{\substack{a_l;\\(a_l,a_j)}} \frac{R_A^\theta(a_l).w_{a_l}^{a_j}}{O(a_l)})$
    $R_A^{\theta+1}(a_j) = Norm(R_A^{\theta+1}(a_j))$
    $\theta \leftarrow \theta + 1$
  until *convergence*
  /* Step 3: Ranking result sets */
  $R_D \leftarrow Rank(R_D)$
  $R_A \leftarrow Rank(R_A)$
  Return $\{R_D, R_A\}$

where

- θ is the iteration number,
- $d \in [0,1]$ is a damping factor fixed to $d = 0.15$,
- $w_{a_l}^{d_i}$, $w_{d_k}^{d_i}$, $w_{a_l}^{a_j}$ and $w_{d_k}^{a_j}$ are the weighted factor for relationships (a_l, d_i), (d_k, d_i), (a_l, a_j) and (d_k, a_j) respectively. For convenience, the weighted factor w_x^y from entity $x \in A \cup D$ to entity $y \in A \cup D$ is computed as below:

$$w_x^y = r_x * Content(y|x) \quad (12)$$

- the normalization functions $Norm(R_D^{\theta+1}(d_i))$ and $Norm(R_A^{\theta+1}(a_j))$ normalize document and author scores as follows:

$$Norm(R_D^{\theta+1}(d_i)) = \frac{R_D^{\theta+1}(d_i)}{\sum_{k=1}^{|D|} R_D^{\theta+1}(d_k)} \quad (13)$$

$$Norm(R_A^{\theta+1}(a_j)) = \frac{R_A^{\theta+1}(a_j)}{\sum_{l=1}^{|A|} R_A^{\theta+1}(a_l)}$$

– the ranking functions $Rank(R_D)$ and $Rank(R_A)$ rank respectively document set and author set according to their BibRank score $R_D(d_i)$ and $R_A(a_j)$.

5. EXPERIMENTAL EVALUATION

We conduct a series of experiments in order to evaluate the effectiveness of BibRank ranking model within the task of accessing a bibliographic corpus. This section describes the experimental setup and discusses the obtained results.

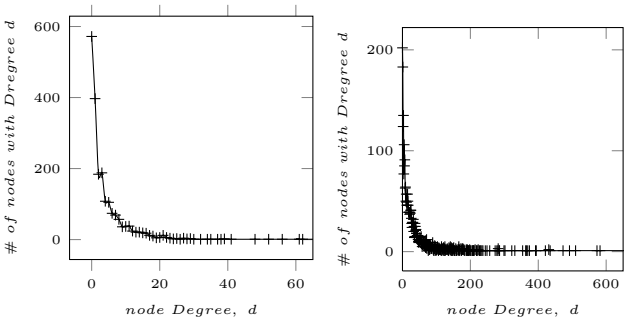
5.1 Experimental Framework

5.1.1 Data Collection Description

We use in this experimental evaluation the CiteSeerX[6] dataset downloaded in April 2011. This collection includes document title, abstract, the list of authors and outgoing citation links. We notice that the citation links between authors are extracted from document citation links by applying an exact matching on the name of authors. Table 1 and Figures 3, 4 and 5 show general statistics about the dataset and the extracted bibliographic network. Analyzing the distribution of nodes and relationships in the network, we note that the giant component includes about 73% of documents and authors entailed in the network. We also notice that author citation links represent 71.1% of the relationships in the network.

Documents	1 472 735
Authors	1 366 540
Citation links between documents	16 598 502
Citation links between authors	51 306 409
Document citation links per document	11.270
Author citation links per author	37.545
Authorship links	4 209 980
Documents per author	3.081
Authors per document	2.858

Table 1: CiteSeerX collection statistics



(a) Document citation network (b) Author citation network

Figure 3: Citation network density

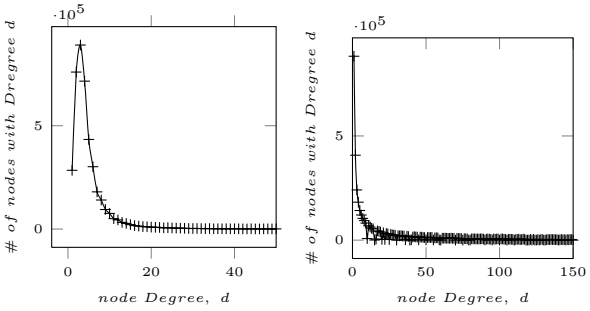
5.1.2 Topics and Assessments

We used the Latent Dirichlet Allocation (LDA) model [1] to extract first automatically a set of 200 topics. The LDA algorithm analyzes word-topic distributions and document-topic distributions. These probabilities allow determining the optimal number of extracted topics from the dataset, in our case a set of 200 topics. Starting from this set, we manually identified generic topics. Then, we extracted manually a subset of 35 topics. Each topic is represented by a list of terms that have been annotated manually to generate a query as shown in Table 2. For instance, three of the 35 topics are "web services", "intelligent agent system", and "markov chain model". For each topic, a subgraph including top documents and their authors is extracted.

Considering the query "intelligent agent system", Figure 6 illustrates, for an entity, its number of incoming citation links regarding its rank obtained with the BibRank model. It appears the most cited entities are highly ranked entities, that highlights the impact of graph structure in our ranking algorithm. Taking into account the fact that a subgraph is extracted from a query topic on one side, and the impact of graph structure on our ranking algorithm on the other side, we can confirm through observed values that the final relevance score is a combination of a *topical relevance assessment* and an *authority relevance one*.

The topical relevance assessment expresses the similarity of the document content to address the query topic. It was performed using a pool-based process. First of all, BibRank ranking and baseline rankings have been computed. We have merged the top 20 documents obtained in each ranking model. We then asked 25 colleagues to assign to documents a binary topical relevance score considering the query topic. Among assessors, we count 9 assistant professors, 13 Phd students, 2 Master students and 1 engineer. Each query is assessed by two different judges. The average agreement between assessors, measured by the Kappa indicator [4], is a

[6]http://citeseer.ist.psu.edu/



(a) Authors per document (b) Documents per authors

Figure 4: Distribution of authors and documents in the authorship network

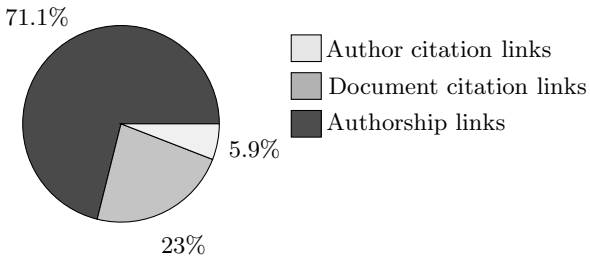
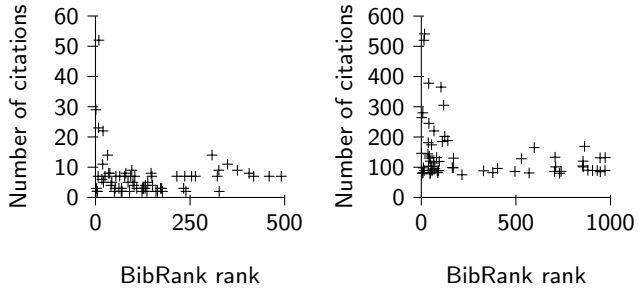


Figure 5: Distribution of links in the network

words	word-topic distributions	query
agent	0.244	
intelligent	0.243	
system	0.083	
artificial	0.067	
autonomy	0.039	intelligent agent
negotiation	0.032	system
cooperation	0.029	
behavior	0.027	
automation	0.023	
coordination	0.022	

Table 2: Topic extraction



(a) Document entities (b) Author entities

Figure 6: Number of incoming citation links according to the BibRank ranking for the query "intelligent agent system"

moderate agreement of $k = 0.57$. Finally, a topical relevance score is automatically computed for each author as the topical assessment score average of his/her documents in the collection.

The authority relevance assessment expresses how much the author expertise covers the query topic. It was inferred from a PageRank score classification computed in each homogeneous subgraph. Entity authority relevance is equal to 1 if the PageRank score is higher than the average PageRank score or equals to 0 otherwise.

5.1.3 Baselines

We compare the retrieval effectiveness of our model to three baselines:

- **the BM25 model** [22] denotes a classical IR matching model that computes the query-entity similarity by a probabilistic model.
- **the Language Model (LM)** [9] denotes a classical IR matching model that computes the query-entity similarity by a smoothed language model, namely the Hiemstra model. The language model is used in our model to compute the content-based score.
- **the PRank model** [27] denotes a matching model for ranking entities in a bibliographic network. This model has been adapted considering author and document nodes rather than author, document and journal nodes.

5.2 Results and Discussion

We have analyzed the retrieval effectiveness of the BibRank algorithm for ranking both document and author entities. To evaluate the retrieval effectiveness for each entity type ranking and considering the two-levels assessment, we use the Normalized Discount Cumulative Gain at 20^{th} document, denoted NDCG@20 as the effectiveness metric that normalizes the n-top retrieved results by the ideal ranking. Improvement significance was computed using student t-test.

We underline that document and author rankings are jointly produced by the BibRank model as outlined in algorithm 1 but for convenience, they are presented below and discussed separately.

5.2.1 Document Ranking Effectiveness

We compare in Table 3 the effectiveness of BibRank algorithm with state-of-the-art and traditional ranking models. *"% change"* column denotes the percentage of BibRank improvement regarding the baselines. We notice significant improvements between 7.03% for PRank model and 113.13% for the Hiemstra language model. The difference of NDCG@20 values for BibRank and PRank rankings is less important comparing to other baseline rankings because PRank is the closest one since it aims at ranking jointly heterogeneous entities, as BibRank algorithm, whereas BM25 and LM are devoted to rank only one type of entities. We also conclude that considering only a topical feature, as in baselines BM25 and LM, or only graph structure, as in PRank algorithm, is not sufficient to estimate the relevance of documents. Compared to these three baselines, our algorithm integrates jointly in addition to these features a content-based score that analyzes the topical relatedness between linked entities.

Model	NDCG@20	% change	
BM25	0.429	**59.77%**	***
LM	0.322	**113.13%**	***
PRank	0.641	**7.03%**	*
BibRank	**0.686**		

Table 3: Document ranking retrieval effectiveness. % change: BibRank improvement. Student test significance *: $0.01 < t \leqslant 0.05$; **: $0.001 < t \leqslant 0.01$; ***: $t \leqslant 0.001$

Figure 7 illustrates the NDCG values at different levels of rank. We notice that the BibRank curve is always over the baseline ones and more particularly when NDCG is computed before rank 5. This means that our algorithm returns in the top 5 results the most relevant documents.

5.2.2 Author Ranking Effectiveness

Considering the ranking of author entities, our model overpasses significantly the different baselines, as shown in Table 4. This confirms the benefit of integrating bibliographic topical and content features in addition to the graph structure for ranking entities in a bibliographic network. However, we notice that NDCG@20 values for author ranking are less important than for document ranking. It can be explained by the way of modeling topical relevance assessments. As said previously, author topical relevance have been inferred from document topical one.

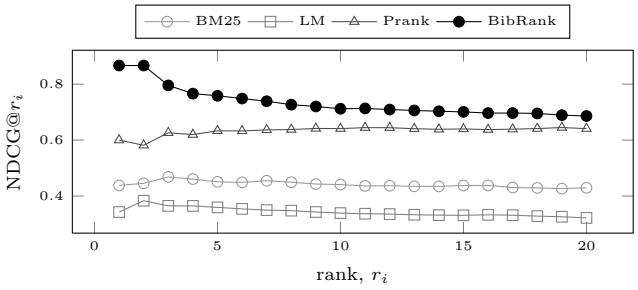


Figure 7: NDCG at different document entity rankings

Model	NDCG@20	% change	
BM25	0.376	**38.26%**	***
LM	0.428	**21.47%**	**
PRank	0.455	**14.29%**	*
BibRank	**0.520**		

Table 4: Author ranking retrieval effectiveness. % change: BibRank improvement. Student test significance *: $0.01 < t \leqslant 0.05$; **: $0.001 < t \leqslant 0.01$; ***: $t \leqslant 0.001$

As for document ranking, Figure 8 introduces the NDCG values computed at different levels of rank. BibRank curve rises above the baseline ones. Compared to document rankings where the BibRank curve remains higher than the baseline ones but declines, this latter grows for author ranking, probably also due to the way of inferring topical relevance assessments for authors.

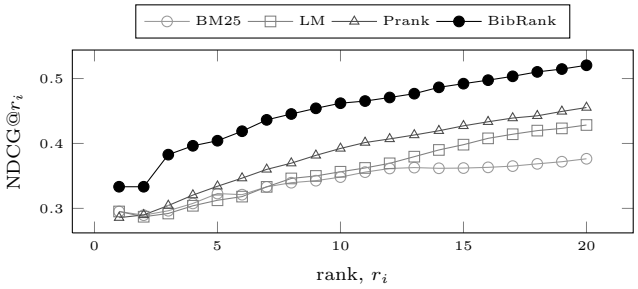


Figure 8: NDCG at different author entity rankings

5.2.3 Ranking Correlation Analysis

Ranking correlation analysis is performed through the Kendall's tau (τ) rank correlation coefficient [14] that analyzes the agreement between two ranking models considering concordant and discordant pairs. The more similar (respectively reversed) the rankings are, the more the correlation coefficient τ approaches 1 (respectively -1). If ranking models are independent τ is null.

Table 5 shows entity ranking coefficient correlation τ. For both document and author rankings, a couple of ranking models is particularly correlated: BibRank and PRank models. They both integrate graph structure and have close rankings. Nevertheless, we have noticed in Table 3 that PRank and BibRank ranking effectivenesses are significantly different.

For document ranking, BM25 and Hiemstra models are weakly correlated. Indeed, they are both classical matching models that rank entities according to a query. Other couples of ranking models are not particularly correlated. In other words, the different rankings between these models considered by pairs are not similar.
For author ranking, we notice that some couples of ranking models are weakly correlated by a negative value, that means that one model ranks authors in a decreasing order regarding the other model.

For instance, Figure 9 shows an instance of rank correlation coefficient for a particular query "intelligent agent system". Both axes represent entity rank for a given ranking model (BM25, LM, PRank or BibRank). *LM-BM25*, for instance, compares BM25 ranking to LM ranking.
For document ranking, most of the scatter plots have points distributed in the whole graph, it explains therefore the weak coefficients reported in Table 5. The graph that opposes BibRank and PRank rankings illustrates a diagonal line that emphasizes the Kendall rank correlation value of 0.594 reported in Table 5.
For author ranking, scatter plots have a different point repartition with the diagonal line only in graph opposing BibRank and PRank which represents indeed the most correlated couple of ranking models. In the other graphs, points are not correlated and present moreover a vertical trend, related to a negative value for Kendall rank correlation coefficient in Table 5. We also notice that some points are concentrated along the abscissa axis as in Figures (b)-(2) to (b)-(5). This expresses the fact that the authors returned by the ranking model represented within the abscissa axis are not returned by the ranking model represented within the ordinate axis. This confirms the negative correlation between the ranking models. This fact is emphasized particularly, as shown in Figures (b)-(2) to (b)-(5), in the case of BM25 and LM ranking models that return few results regarding only the query.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a bi-type entity ranking algorithm for bibliographic networks, called BibRank. This algorithm propagates entity scores by considering both link-based and content-based features in the network. For this purpose, we estimate the significance of citation links by measuring the common similarity between homogeneous entities regarding a topic and we also evaluate, using a language model, the document representativeness to author's knowledge and the author's knowledge on document topic.

We conducted a series of experiments on CiteSeerX dataset that shows a significant improvement in comparison to other models for both document and author rankings. We also conclude that matching models based only on the query topic are not sufficient for ranking entities in heterogeneous subgraphs. Analyzing ranking correlations, we note that PRank and BibRank algorithms present correlated entity rankings. Meanwhile, this correlation agreement is counterbalanced by a significant improvement for BibRank algorithm.

For future research work, we plan to extend our model to large-scale bibliographic networks including more entity types and relationships, such as co-authorship, the author's affiliation to an institution. We also plan to integrate temporal aspects of citation links in order to measure the author-

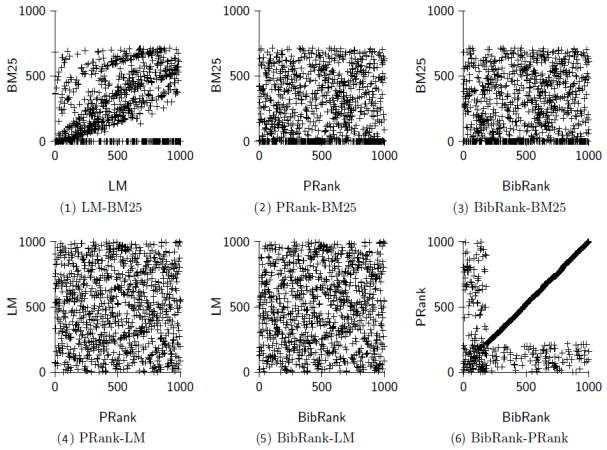
	BM25	LM	PRank	BibRank
BM25	1	**0.160**	0.002	0.015
LM		1	0.022	0.039
PRank			1	**0.594**
BibRank				1

(a) Document rankings

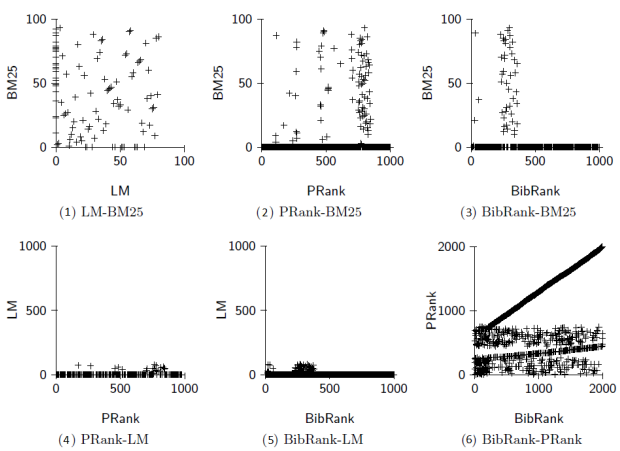
	BM25	LM	PRank	BibRank
BM25	1	-0.187	-0.082	-0.160
LM		1	-0.072	-0.151
PRank			1	**0.600**
BibRank				1

(b) Author rankings

Table 5: Rank correlation coefficient for document and author rankings



(a) Document rankings



(b) Author rankings

Figure 9: Comparison of document and author rankings for query "intelligent agent system"

ity acquired by a document regarding its publication date. In this mind, the importance of documents is not therefore measured by the number of in-coming links but by the average number of citation links per year. We expect that this revised approach would favour efficient entities rather than basically most cited ones.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 19–26, New York, NY, USA, 2006. ACM.

[2] S. Alonso, F. J. Cabrerizo, E. H. Viedma, and F. Herrera. hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*, 82:391–400, 2010.

[3] C. T. Bergstrom, J. D. West, and M. A. Wiseman. The eigenfactor metrics. *Journal of Neuroscience*, 28(45):11433–11434, 2008.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[5] A. Dahlan and B. Sitohang. Combining pagerank and citation analysis to measure information credibility in internet. In *Proceedings of iiWAS*, pages 375–382, 2007.

[6] L. Egghe. An improvement of the h-index: the g-index. *ISSI Newsletter*, 2:8–9, 2006.

[7] M. Frické and D. Fallis. Indicators of accuracy for answers to ready reference questions on the internet. *J. Am. Soc. Inf. Sci. Technol.*, 55:238–245, 2004.

[8] E. Garfield. The history and meaning of the journal impact factor. *JAMA: The Journal of the American Medical Association*, 1, 2006.

[9] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, pages 569–584, London, UK, 1998. Springer-Verlag.

[10] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102:16569–16572, 2005.

[11] L. B. Jabeur, L. Tamine, and M. Boughanem. A social model for literature access: towards a weighted social network of authors. In *Proceedings of Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 32–39, 2010.

[12] F. Jelinek and R. L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands: North-Holland, 1980.

[13] B. Jin, L. Liang, R. Rousseau, and L. Egghe. The r- and ar-indices: Complementing the h-index. *Chinese Science Bulletin*, 52:855–863, 2007.

[14] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[15] S. M. Kirsch, A. Prof, D. Armin, B. Cremers, and S. I.

A. D. Rheinischen. Social information retrieval, 2005.

[16] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, 1999.

[17] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Comput. Netw.*, 33:387–401, 2000.

[18] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Inf. Process. Manage.*, 41:1462–1480, 2005.

[19] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 91–98, New York, NY, USA, 2006. ACM.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[21] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281. ACM, 1998.

[22] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[23] J. Tang, R. Jin, and J. Zhang. A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 1055–1060, Washington, DC, USA, 2008. IEEE Computer Society.

[24] S. Uddin, L. Hossain, A. A., and K. Rasmussen. Trend and efficiency analysis of co-authorship network. *Scientometrics*, In press, 2011.

[25] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a simple model of network traffic. *Society*, pages 1–5, 2006.

[26] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

[27] E. Yan and Y. Ding. Measuring scholarly impact in heterogeneous networks. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 1–7, Silver Springs, MD, USA, 2010. American Society for Information Science.

[28] Z. Yang, L. Hong, and B. D. Davison. Topic-driven multi-type citation network analysis. In *Proceedings of Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 24–31, 2010.

[29] C.-T. Zhang. The e-index, complementing the h-index for excess citations. *PLoS ONE*, 4, 2009.

[30] J. Zhang, J. Tang, B. Liang, Z. Yang, S. Wang, J. Zuo, and J. Li. Recommendation over a heterogeneous social network. In *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management*, WAIM '08, pages 309–316, Washington, DC, USA, 2008. IEEE Computer Society.

[31] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles. Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 739–744, Washington, DC, USA, 2007. IEEE Computer Society.