On Ranking Relevant Entities in Heterogeneous Networks Using a Language-Based Model

Laure Soulier, Lamjed Ben Jabeur, Lynda Tamine, and Wahiba Bahsoun
IRIT, University of Toulouse, 118 route de Narbonne, 31062 Toulouse, France.
E-mail: {soulier, jabeur, tamine, wbahsoun}@irit.fr

A new challenge, accessing multiple relevant entities, arises from the availability of linked heterogeneous data. In this article, we address more specifically the problem of accessing relevant entities, such as publications and authors within a bibliographic network, given an information need. We propose a novel algorithm, called BibRank, that estimates a joint relevance of documents and authors within a bibliographic network. This model ranks each type of entity using a score propagation algorithm with respect to the query topic and the structure of the underlying bi-type information entity network. Evidence sources, namely content-based and network-based scores, are both used to estimate the topical similarity between connected entities. For this purpose, authorship relationships are analyzed through a language model-based score on the one hand and on the other hand, non topically related entities of the same type are detected through marginal citations. The article reports the results of experiments using the Bibrank algorithm for an information retrieval task. The CiteSeerX bibliographic data set forms the basis for the topical query automatic generation and evaluation. We show that a statistically significant improvement over closely related ranking models is achieved.

Introduction

Information networks include a large number of components, called entities, related to each other by relationships. They aim at sharing information and emphasize interdependencies between entities within the network. Authors, for instance (Zhou, Orshanskiy, Zha, & Giles, 2007; Zhang et al., 2008; Yan & Ding, 2010), mainly distinguish homogeneous networks from heterogeneous ones. The former are characterized by entities of the same type, connected to each other by one type of relationship whereas the latter include entities of multiple types and related to each other using several types of links. These kinds of networks are used in several application domains such as biology (Roy, Lane, & Werner-Washburne, 2008), transport (Emmerink, 1993), scientific collaboration (Coyle & Smyth, 2008), scholarly communication (Cabanac, 2012), and email and meeting management (Minkov & Cohen, 2006).

In this article, we address the problem of ranking entities in a heterogeneous information network within an information retrieval (IR) task. Our bi-type entity-based structure, namely a bibliographic network, contains heterogeneous entities including documents and authors, and their semantic relationships such as citation links and authorship links.

Previously, ranking algorithms such as PageRank (Page, Brin, Matwani, & Winograd, 1999), HITS (Kleinberg, 1999) and Salsa (Lempel & Moran, 2000) have been proposed for homogeneous document networks. In this context, authors (Page et al. 1999; Kleinberg, 1999; Lempel & Moran, 2000) have introduced hyper-link analysis between entities to emphasize authoritative entities. Heterogeneous networks have highlighted a new challenge, namely ranking jointly different types of entities, heterogeneous entities on the one hand, and on the other, their heterogeneous semantic relationships, which can be weighted differently.

A possible method to address this challenge consists in using bibliometric indicators for determining important entities (Ibáñez, Larranaga, & Bielza, 2011). These measures involve a network-based analysis between related entities. We can distinguish indicators based only on citation-links (Hirsch, 2005; Egghe, 2006; Zhang, 2009) and others, which, moreover, consider time features such as publication date (Garfield, 1955; Walker, Xie, Yan, & Maslov, 2006; Uddin, Houssain, Abbasi, & Rasmussen, 2012). For instance, Hirsch (2005) proposes an h-index indicator; this

Received January 24, 2012; revised May 23, 2012; accepted June 14, 2012

 DOI: 10.1002/asi.22762

measure computes, for an author, a value h in which for all its authored documents, h papers are cited at least h times and the other papers are cited less.

However, the main works dealing with entity ranking focus on graph structure analysis by using either network-based measures such as the PageRank algorithm or its variant to determine authoritative entities (Kleinberg, 1999; Kurland & Lee, 2005; Liu, Bollen, Nelson, & Van de Sompel, 2005; Zhao, 2006). More specifically, two lines of works are reported regarding the entity types being ranked in heterogeneous information networks. In (Kirsch et al., 2006; Kurland and Lee, 2006; Sayyadi and Getoor, 2009; Jabeur et al., 2010; Yang et al., 2010), one type of entity is ranked while in (Nie , 2006; Zhou et al., 2007; Zhang et al., 2008; and Yan and Ding, 2010), multi-type entities are jointly ranked.

In this article, we propose a novel approach for co-ranking documents and authors in a bi-type bibliographic network within an IR task. The core idea of the approach is that relevance should be measured using evidence issued from (1) topical intrinsic content of document subjects and author's scientific production, (2) structure of both homogeneous and heterogeneous citation and authoring subgraphs and (3) relevant inter-graph citations.

The rest of this article is structured as follows. In Related Work, we describe our survey on the area of entity ranking in homogeneous and heterogeneous information networks to clarify how entity ranking models occur in bibliographic networks within an IR task. Contribution and Comparison of Related Work explains how our contribution puts forward a particular stance towards related work presented in Related Work. In The Bibliographic Network section, we present definitions and preliminary notations about bibliographic networks. The BibRank Algorithm section details the BibRank algorithm and its qualitative and quantitative components. Experimental Evaluation describes the evaluation methodology and discusses the results of the experimental evaluation using CiteSeerX data set. Conclusion and Future Work provides concluding remarks and identifies future research directions.

Related Work

Literature access is a specific application domain of IR in which the main problem concerns the ranking of either publications or authors within a bibliographic network. Ranking entities within bibliographic networks is tackled generally by network-based approaches (Zhou, et al., 2007; Zhang et al., 2008; Jabeur, et al., 2010; Yan & Ding, 2010) that rank entities according to topical-based and network-based features.

The network-based approaches rank bibliographic entities in response to a query topic according to the basic assumption that important entities are related to other important ones. Similarly to our BibRank algorithm, all of these approaches use the mutual reinforcement principle between connected entities within a bibliographic network. Ranking algorithms were proposed for both homogeneous networks where ranking is proposed for homogeneous entities (Page, et al., 1999; Kleinberg, 1999; Kurland & Lee, 2005; Liu, et al., 2005; Zhao, 2006), and heterogeneous networks where ranking is computed for either mono-type (Kirsch et al., 2006; Kurland & Lee, 2006; Sayyadi & Getoor, 2009; Jabeur et al., 2010) or multi-type entities according to the network topology and the query topic (Nie et al., 2006; Zhou et al., 2007; Zhang et al., 2008; Tang, Jin, & Zhang, 2008; Yan & Ding, 2010; Yang et al., 2010). We introduce in what follows the two categories of models that rank entities within either a homogeneous network or a heterogeneous one.

Ranking Entities in Homogeneous Networks

In the case of bibliographic homogeneous networks, entities are of the same type, mainly documents, and are related to each other by citation links. The latter are exploited to detect important documents and rank them by the "surfer random walk" general model (Pearson, 1905). Prior, document ranking algorithms such as PageRank (Page et al., 1999) or HITS (Kleinberg, 1999) consider hyper-links in a web page collection to highlight important web pages connected by citation links. Some variants of PageRank algorithm have been proposed for ranking document entities (Kurland & Lee, 2005) and author entities (Liu et al., 2005; Zhao, 2006).

Concerning the ranking of document entities, Kurland and Lee (2005) propose to model and, therefore, weight relationships between independent documents to compute a PageRank-like algorithm applied on the corresponding connected graph. The weight of relationships between two documents reflects their textual similarity and is estimated with a smoothed Kullback-Leibler divergence measure (Kullback & Leibler, 1951) between the language models (Ponte & Croft, 1998) of the two respective documents. The ranking algorithm is divided into three steps: 1) generating the weighted document network applying the textual similarity measure, 2) computing the document centrality with a propagation algorithm, and 3) ranking documents by a multiplicative combination of the centrality measure and the topical relevance to the query topic. Experiments show that considering both textual relationships and centrality is effective for ranking independent documents.

Concerning the author ranking, the AuthorRank algorithm (Liu et al., 2005) enhanced the traditional PageRank algorithm by considering weighted coauthorship links rather than unweighted ones. Experimentation shows that AuthorRank and PageRank algorithms applied on the coauthorship network are highly correlated without significant impact on the model effectiveness. However, AuthorRank outperforms the rankings provided by bibliometric indicators such as degree or closeness (Hanneman & Riddle, 2005). In Zhao (2006), the authors propose an author cocitation analysis method (ACA) specifically feasible for ranking multiple

DOI: 10.1002/asi

author documents. The method exploits three main link types: author cocitation, inclusive all-author cocitation, and exclusive all-author cocitation. Experimental results have highlighted that considering all the paper authors in citation links improves the author's ranking. The difference between inclusive and exclusive all-author cocitation links depends on the purpose of the study. If the aim is to represent a research field considering the intellectual structure, exclusive all-author cocitation links feature is more appropriate, otherwise, inclusive all-author cocitation links analysis is recommended.

Ranking Entities in Heterogeneous Networks

Heterogeneous bibliographic networks include multi-type entities. In this context, several works have proposed ranking models that rank one type of entity (Kurland & Lee, 2006; Kirsch et al., 2006; Sayyadi & Getoor, 2009; Jabeur et al., 2010) or rank jointly several types of entity (Nie et al., 2006; Zhou et al., 2007; Zhang et al., 2008; Tang et al., 2008; Yan & Ding, 2010; Yang et al., 2010).

Mono-type entity ranking approaches for heterogeneous bibliographic networks. According to these approaches, one kind of entity type is ranked considering its relationships with other graph entities of different types. Entity relevance is generally estimated as the related importance in the network. Regarding relevance estimation of entities, we mainly distinguish between modular models (Kurland & Lee, 2006; Kirsch et al., 2006; Jabeur et al., 2010) that combine topical and network-based features and integrated ones (Sayyadi & Getoor, 2009) that compute entity relevance as a whole using a spread activation process in the network.

Some of the modular approaches (Kirsch et al., 2006; Jabeur et al., 2010) consider the bibliographic network as a social one and therefore compute document relevance by combining the topical relevance and the social importance of their authors. Different network topologies are considered including citation network, co-authorship network, or both citation and authorship network. Results show that ranking entities in networks including citation links enhances the ranking effectiveness in comparison to co-authorship networks. Kurland and Lee (2006) use a bipartite network including documents and clusters of documents. Relationships between entities are weighted by a textual similarity measure computed by the Kullback-Leiber divergence measure (Kullback & Leibler, 1951). The algorithm is designed to re-rank documents with a mutual reinforcement algorithm between documents and clusters. This method is based on the assumption that central clusters should include a large percentage of relevant documents. For this purpose, the different clusters are ranked first by a centrality measure using a variant of HITS algorithm applied on the weighted graph. Then, each entity is ranked within each cluster according to the query topic. Finally the different rankings are merged ordering documents by combining their cluster centrality and topical relevance measure. Experimental evaluation shows that mutual reinforcement between documents and clusters are promising for both ranking documents and building clusters that include several relevant documents.

According to the integrated approach, Sayyadi and Getoor (2009) introduce a variant of the PageRank algorithm, called FutureRank. An entity score in a bi-type bibliographic network is computed using a personalized score propagation algorithm that uses evidence from current date and publication time. Experimental evaluation shows that considering citation links and time feature in ranking algorithms outperforms the traditional PageRank algorithm.

Multi-type entity ranking approaches for heterogeneous bibliographic networks. In these approaches, each type of entity is ranked according to its different semantic relationships with other entities. In several works (Zhou et al., 2007; Zhang et al., 2008; Tang et al., 2008; Yan & Ding, 2007; Yang et al., 2010), algorithms for score propagation in bibliographic networks are proposed. In the same way of *mono-type entities ranking approaches for heterogeneous networks*, we distinguish modular algorithms that combine different features (Zhou et al., 2007; Zhang et al., 2008) and integrated ones (Tang et al., 2008; Yan & Ding, 2007; Yang et al., 2010).

Among modular algorithms, Zhou et al. (2007) consider a bipartite graph including two homogeneous subgraphs of authors and documents. The entity relevance scores are computed using the mutual reinforcement principle based on the assumption that the more authoritative an author is, the more likely a document is perceived as relevant and reciprocally. Entity score results from the combination of a PageRank score in the homogeneous subgraph and a *biWalk* score that considers inter-graph relationships. This co-ranking algorithm is evaluated as effective for author ranking in comparison to the PageRank algorithm computed on the author subgraph. Document ranking effectiveness is not evaluated. Another algorithm (Zhang et al., 2008) aims at recommending heterogeneous entities in a weighted bibliographic network according to the social view. Weights are assigned to social relationships between authors, documents, resources, and tags are computed as network-based probabilities. A topical document entity relevance score according to the query is estimated and combined with its importance score regarding relationships between an entity and the other socially related ones.

According to the integrated approaches, variants of the PageRank algorithm are proposed introducing topical feature (Tang et al., 2008; Yang et al., 2010) or time feature (Yan & Ding, 2007). For instance, Arnetminer[1] is a "scholar" search engine that uses a PageRank-like algorithm including an "Author-Topic-Conference" (ACT) model. Authors propose three different ACT models and two ways of combining ACT scores within a random walk.

[1]http://arnetminer.org/

DOI: 10.1002/asi

The main idea is described as follows: the ACT model generates for each entity a topic distribution similar to the LDA algorithm (Blei et al., 2003). Documents and venue topics distribution are inferred from author topics. Then, ACT scores are considered as weights for score propagation using the mutual reinforcement between connected entities. Experimental results show that this approach outperforms a classical ranking model with BM25, and other closely related ranking models obtained by two competing academic search engines, namely, Libra, the Microsoft Academic Search engine[2] (+3,4% by MAP) and Rexa[3] (+15,6% by MAP). Yang et al. (2010) also propose an other model based on topic distribution. For this purpose, they apply the Topical PageRank (TPR) algorithm (Nie et al., 2006) in a multi-type citation network that entails authors, venues, authors, and papers nodes. TPR is a PageRank algorithm extension that considers three different surfing behaviors: "Follow-stay" when a surfer stays in the same topic, "Follow-jump" when a surfer changes the topic regarding previous entity topics and "Jump-jump" when the user accesses randomly a topic through a fixed entity. Experiments show that multi-type citation networks allow to improve entity rankings and TPR outperforms author ranking thanks to authority based measures. Yan and Ding (2007) propose a network-based analysis algorithm, called PRank, that ranks articles, authors and journals within a heterogeneous bibliographic network. This algorithm investigates two main properties of important bibliographic entities. First, important entities are cited by important ones (Page et al., 1999; Zhou et al., 2007; Yang et al., 2010). Second, recently published articles are more important because users are generally interested in the most recent work in their research area (Walker et al., 2006; Sayyadi & Getoor, 2009). Accordingly, PRank computes for each type of entity a relevance score that takes into account the importance of associated entities in the network as well as their freshness. For instance, document scores depend, first, on the score of their corresponding authors and journals, and second, on the document publication date. In the same way, author scores depend on the score of their corresponding articles. Experimental results show that time is not an effective feature for this model.

Contribution and Comparison with Related Work

In this article, we propose a bi-type entity ranking model for bibliographic networks. This model is used for an IR task that jointly ranks document and author entities for a particular topic. Two main features are used in our algorithm. The topical-based feature considers the topical relevance of an entity according to the query topic. The content-based feature estimates the topical similarity between connected entities. According to authorship links, the content-based scores allow us to evaluate an author's scientific production representativeness on document topic and the document representativeness of an author's scientific production. For document or author citation relationships, the content-based features allow us to detect marginal citation links, in other words, non-topically focused citation links. For this purpose, we have analyzed the topical similarity of the connected entities using a rank-based measure in order to estimate the joint similarity of the two entities regarding the query topic.

The model presented in this article is different from previous work in the same area in several respects. First of all, our model relies on an integrated approach of rankings unlike works relying on feature combination (Kurland & Lee, 2005; Kirsch et al., 2006; Kurland & Lee, 2006; Zhou et al., 2007; Zhang et al., 2008; Jabeur et al., 2010). Furthermore, our model provides joint bi-type entity rankings unlike previous works that provide mono-type entity rankings (Page et al., 1999; Kleinberg, 1999; Lempel & Moran, 2000; Kurland & Lee, 2005; Liu et al., 2005; Zhao, 2006; Kirsch et al., 2006; Kurland & Lee, 2006; Sayyadi & Getoor, 2009; Jabeur et al., 2010).

Regarding works that are most closely related to ours, we can highlight two main facets of differences. Regarding the sources of evidence used for ranking, our work integrates three distinct features based on the query topic, the graph structure and the topical similarity between connected entities, unlike previous works that exploit only the two first ones (Zhou et al., 2007; Zhang et al., 2008; Yan & Ding, 2010). From another perspective even if some authors (Tang et al., 2008; Yang et al., 2010) also consider a topical feature for entities, the topic is used however to represent entities themselves by means of topic distribution and not for investigating the strength of the relationships between entities. Considering the topical similarity between entities, we distinguish two main dissimilarities with the work of Kurland and Lee (2005; 2006). Besides the difference in the general approach and the objective, we have introduced a new metric that estimates topical relatedness between entities of the same type, called marginal citations. This measure is different from the measure proposed in Kurland and Lee (2005; 2006) in so far as we consider that citation links are marginal considering the query topic whereas Kurland and Lee use the textual similarity measure regardless of the query.

More specifically, the contributions of the paper are the following:

- A novel algorithm, called BibRank, for bi-entity ranking in a heterogeneous bibliographic network. The algorithm integrates topical and content-based features into a ranking model by providing insight on the global connexion between embedded homogeneous subnetworks. In the case of relationships between authors and documents, we introduce an author's scientific production representativeness to document topic and the document representativeness of an author's scientific production using a language model-based measure. In the case of relationships between entities of the same type, we

[2]http://academic.research.microsoft.com/

[3]http://rexa.info/

DOI: 10.1002/asi

Citing document	Cited document
d_1	d_3
d_3	d_2
d_4	d_3

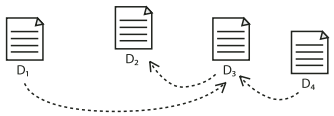


FIG. 1. Document citation network.

Author	Document
a_1	d_1
a_2	d_1
a_2	d_2
a_3	d_2
a_4	d_3
a_5	d_4

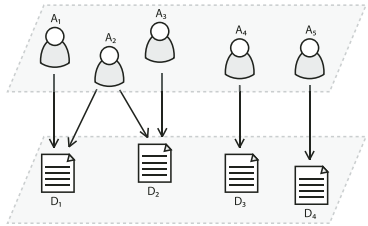


FIG. 2. Authorship network.

Citing author	Cited author
a_1	a_4
a_2	a_4
a_5	a_4

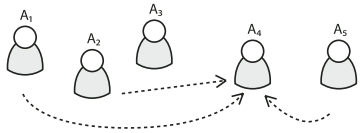


FIG. 3. Author citation network.

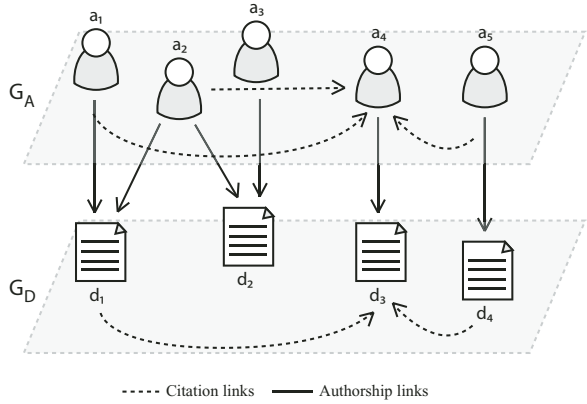


FIG. 4. Bi-type bibliographic graph *G*.

propose to discredit marginal citations measured by topical common interest between them considering the query topic.
- An intensive comparative evaluation with different state-of-the-art ranking models. We empirically show that the BibRank model outperforms significantly closely related ranking models.

The Bibliographic Network: Preliminaries and Notations

Bi-type bibliographic network. A bi-type bibliographic network is a graph of two types of entity: documents that represent information nodes and authors that represent individual nodes. These two types of entity are related by incoming and outgoing links. The bi-type bibliographic network is represented by a graph $G = \{V, E\}$ where $V = A \cup D$. $A = \{a_1, \ldots, a_{n_A}\}$ and $D = \{d_1, \ldots, d_{n_D}\}$ are entities which respectively correspond to a set of n_A authors and a set of n_D documents. $E \subseteq V \times V$ represents the set of edges of the graph that expresses relationships between entities. When entities are of the same type, relationships are called *intra-graph* whereas they are called *inter-graph* when entities are of different types. Edges represent semantic links as described below:

Document citation associations e_{DD}: the intra-graph link $e_{d_i d_{i'}}$ connects two scientific documents where document $d_i \in D$ *cites* at least once document $d_{i'} \in D$. Figure 1 shows document citation associations and their corresponding networks.

Authorship associations e_{DA}: the inter-graph link $e_{d_i a_j}$ (or $e_{a_j d_i}$) connects author $a_j \in A$ with his/her authored document $d_i \in D$. For example, $e_{d_1 a_2}$ means that the author a_2 (or document d_1) can be reached from document d_1 (respectively author a_2). Authorship links are represented by a bi-directional edge. In this way, Figure 2 shows authorship associations and their corresponding networks.

Author citation associations e_{AA}: the intra-graph link $e_{a_j a_{j'}}$ shows the connection from author $a_j \in A$ to author $a_{j'} \in A$, inferred from document citation links. Considering a citation link from a document d_1 to a document d_3 where the document d_1 is authored by two authors $\{a_1, a_2\}$ and the document d_3 is authored by one author a_4. The author citation links $e_{a_1 a_4}$ and $e_{a_2 a_4}$ can be deduced. We notice that self-citation links are not considered. Figure 3 presents author citation associations deduced from Figure 1 and Figure 2.

Author and document homogeneous subgraphs. As a bi-type bibliographic network is a heterogeneous network containing two types of entity, two homogeneous subgraphs can be deduced: one for authors $G_A = \{V_A, e_{AA}\}$ and the other for documents $G_D = \{V_D, e_{DD}\}$. V_A and V_D represent respectively author nodes and documents nodes and relationships e_{AA} and e_{DD} are intra-graph relationships described above. Subgraphs G_A and G_D are related by inter-graph relationships between authors and documents, previously called e_{DA}. Figure 4 shows the complete bi-type bibliographic network obtained thanks to document citation links (Figure 1), authorship links (Figure 2) and author citation links (Figure 3). This network is divided into two homogeneous subgraphs G_A and G_D.

DOI: 10.1002/asi

Author, document and collection language models. Ponte and Croft (1998) have defined language models for documents in order to estimate the topical similarity between a query Q and a document d_i by the probability $P(Q|M_{d_i})$ of the query Q regarding the language model of document d_i:

$$P(Q|M_{d_i}) = \prod_{t \in Q} P(t|M_{d_i}) \quad (1)$$

The language model M_{d_i} of document d_i analyzes the term distribution with a maximum likelihood method. The probability $P(t|M_{d_i})$ of term t considering the term distribution of document d_i is computed as follows:

$$P(t|M_{d_i}) = \frac{tf(t, d_i)}{dl_{d_i}} \quad (2)$$

where $tf(t, d_i)$ denotes the term frequency of t in document d_i and dl_{d_i} is the total number of terms in document d_i.

From this general model, we have inferred the author language model and the collection language model.

The author language model considers an author a_j as a textual entity that aggregates all his/her published documents. The probability $P(t|M_{a_j})$ of term t considering the term distribution in documents written by author a_j is computed as follow:

$$P(t|M_{a_j}) = \frac{tf(t, a_j)}{dl_{a_j}} \quad (3)$$

where $tf(t, a_j)$ denotes the term frequency of t for the document set written by author a_j and dl_{a_j} the total number of terms in documents published by author a_j.

The collection language model is similar to document or author model to some extent that it estimates the probability $P(t|M_c)$ of a term t considering the term distribution in the whole documents:

$$P(t|M_c) = \frac{tf(t, c)}{dl_c} \quad (4)$$

where $tf(t,c)$ denotes the term frequency of t for the whole document set in the collection c and dl_c the total number of terms in the whole document set.

BibRank Algorithm

General Description

In our study, a bibliographic network is used in an IR context for co-ranking two types of entity: authors and documents. The IR process is launched by a query $Q = \{w_{1q}, \ldots, w_{jq}, \ldots, w_{Kq}\}$ modeling an information need where w_{jq} is the weight of the j^{th} term of the query and K is the length of the query.

We introduce a BibRank function that ranks each type of entity included in the homogeneous subgraphs issued from the heterogeneous one $G = \{A \cup D, E\}$. Its underlying principle is illustrated in Figure 5. The BibRank function gives scores for each author a_j and document d_j entity where:

$$\begin{aligned} &BibRank : \{Q, G\} \rightarrow \{R_A, R_D\} \\ &\forall a_j \in A, 0 < R_A(a_j) < 1, \sum_{a_j \in A} R_A(a_j) = 1 \\ &\forall d_i \in D, 0 < R_D(d_i) < 1, \sum_{d_i \in D} R_D(d_i) = 1 \end{aligned} \quad (5)$$

R_A (a_j) represents the score of author a_j according to the query topic and the graph structure. In the same way, $R_D(d_j)$ represents the score of document d_i according to the query topic and the graph structure.

BibRank Score Computation

In this section, we detail the BibRank algorithm computations and the theoretical related justifications. BibRank is based on basic assumptions:

- *Assumption 1:* Important documents (respectively authors) are those cited by many other important documents (respectively authors) (Zhang et al., 2008; Yang et al., 2010).
- *Assumption 2:* Important documents are those authored by many important authors and reciprocally (Zhou et al., 2007; Yang et al., 2010).

The two homogeneous subgraphs of authors and documents are connected by transition probabilities. Moreover, the BibRank algorithm applies a score propagation process based on assumption 1 and assumption 2 in order to rank jointly each type of entity. Our algorithm integrates two indicators:

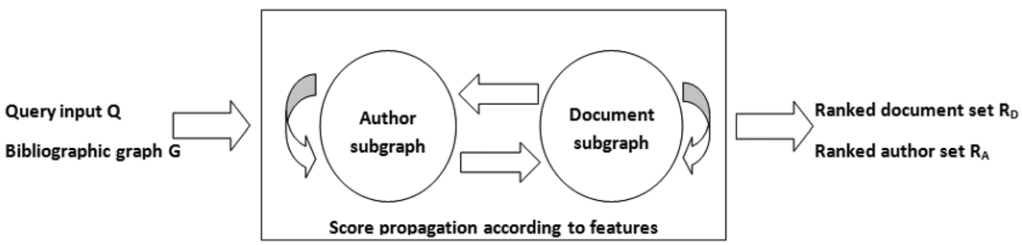


FIG. 5. BibRank: an integrated approach for co-ranking entities in heterogeneous information networks.

DOI: 10.1002/asi

the first one, *the topical-based indicator*, takes into account the similarity relevance to the query input whereas the second one, *the content-based indicator*, considers the topical similarity between authors and documents.

Computing transition probabilities between the homogeneous subgraphs. Transition probabilities enable to measure the moving actions from a subgraph to another one. Assuming the current node is a document, more the transition probability from document subgraph to author subgraph is high, more the likelihood to access an author node is high. For convenience, the transition probability of accessing a subgraph of type $Y \in \{A, D\}$ from a subgraph of type $X \in \{A, D\}$ is computed as follows:

$$\lambda_{XY} = \frac{C(G_X; G_Y)}{|E|}$$

$$\lambda_{AA} + \lambda_{AD} = 1$$

$$\lambda_{DD} + \lambda_{DA} = 1$$

where $C(G_X, G_Y)$ is the number of outgoing links from subgraph of type X to subgraph of type Y and $|E|$ is the number of edges in the bibliographic network.

Computing query-entity topical-based scores. Entity topical-based scores are estimated by computing the content similarity between an entity and the query input Q (Hiemstra, 1998). Assuming basically that top-ranked entities receive a higher score, their inverted rank is retained as an entity-query similarity indicator. For an entity $x_i \in A \cup D$, its reciprocal rank r_{x_i} is computed as follows:

$$r_{x_i} = \frac{1}{rank(x_i)} \quad (6)$$

where $rank(x_i)$ is obtained by ranking the query-entity similarity obtained by the language model (Hiemstra, 1998).

Computing entity-entity content-based scores. The content-based scores allow us to measure the topical relatedness between two connected entities in the graph. In our setting, both citation and authorship links, respectively intra-graph and inter-graph relationships, are considered. For this purpose, the content-based score $content(x_k|y_l)$ between two connected entities $x_k \in A \cup D$ and $y_l \in A \cup D$ is computed into two ways according to inter-graph and intra-graph relationships. We assume that the more one entity is similar to another related by an incoming link, the more the former receives the score of the latter.

For inter-graph relationships, a content-based score is computed using a language model. Two semantic interpretations between authors and documents are induced in order to model the directed link from an author to his documents and reciprocally from a document to its authors: (1) the document representativeness of author's scientific production and (2) the author's scientific production representativeness regarding the document topic.

For intra-graph relationships, marginal citations enable us to measure the common interest of two entities regarding the query input Q. For this purpose, we assume that two entities are topically related if their related ranks are closed considering the query. Therefore, the relationship between these entities is characterized by a semantically focused citation link as detailed below.

Document representativeness of author's scientific production. For a given author $a_j \in A$ which has authored a document set $\mathcal{D}(a_j)$, the score $content(d_k|a_j)$ for $d_k \in \mathcal{D}(a_j)$ determines the topical similarity between document d_k and its author a_j. We compute this score for each authorship link from authors to documents. Thus, the more a document is topically similar to its author's scientific production, the more the author contributes to document score. The score $content(d_i|a_j)$ is computed as follows:

$$content(d_i|a_j) = \frac{P(a_j|M_{d_i})}{\max_{\forall d_k \in D, a_l \in A; w(e_{a_l d_k})=1} P(a_l|M_{d_k})} \quad (7)$$

where $w(e_{a_l d_k}) = \begin{cases} 1 & if\ d_k \in \mathcal{D}(a_l) \\ 0 & otherwise \end{cases}$

$P(a_j|M_{d_i})$ is the probability of author a_j according to the language model of document d_i, described in section 4. It is computed by the Hiemstra formula (Hiemstra, 1998):

$$P(a_j|M_{d_i}) = \prod_{t \in a_j} \left[(1-\lambda) P(t|M_{a_j}) + \lambda P(t|M_{d_i}) \right]^{tf(t,a_j)} \quad (8)$$

Author's scientific production representativeness to the document topic. In the same way, for a current document $d_i \in D$, the representativeness of the scientific production of its authors $\mathcal{A}(d_i)$ according to the document topic and all authorships links is computed by the score $content(a_l|d_i)$:

$$content(a_j|d_i) = \frac{P(d_i|M_{a_j})}{\max_{\forall d_k \in D, a_l \in A; w(e_{a_l d_k})=1} P(d_k|M_{a_l})} \quad (9)$$

where $w(e_{a_l d_k}) = \begin{cases} 1 & if\ d_k \in \mathcal{D}(a_l) \\ 0 & otherwise \end{cases}$

$P(d_i|M_{a_j})$ is the probability of document d_i according to the language model of author a_j, detailed in section 4. It is computed as follows:

$$P(d_i|M_{a_j}) = \prod_{t \in d_i} \left[(1-\lambda) P(t|M_c) + \lambda P(t|M_{a_j}) \right]^{tf(t,d_i)} \quad (10)$$

Marginal citations. We analyze through marginal citations the likelihood of one entity being cited by another based on their topical relatedness to the query. More specifically,

DOI: 10.1002/asi

we assume that citation links are marginal if the two connected entities do not deal with the same topic. For this purpose, we investigate the detection of non-focused citation links called also marginal citations. Generally speaking, focused citation links express common topic interest between authors and/or documents. Analyzing the semantics of the citation link enables to gauge the reliability of the link itself. We assume that a citation link between two homogeneous entities is more reliable when entities are semantically related. Therefore, we propose to discredit nonsemantic citations leading to marginal ones. To achieve this objective, a common similarity indicator between two entities is computed using their corresponding ranks in a IR framework. Thus, we assume that two documents have a common interest if they are both relevant to the query topic. The common similarity indicator $sim_{com}(x_m, y_{m'})$ between two homogeneous entities $x_m \in X$ and $y_{m'} \in X$, where $X = \{A, D\}$, is computed as follows:

$$sim_{com}(x_m, y_{m'}) = \frac{1}{|r_{x_m} - r_{y_{m'}}|} \quad (11)$$

where r_{x_m} and $r_{y_{m'}}$ are the ranks obtained by entities x_m and $y_{m'}$ in an IR framework that ranks each entity according to their relevance scores according to the query topic.

We can deduce the content-based score of entity $x_m \in X$ from entity $x_p \in X$ relative to all the homogeneous entity citation links, written $content(x_m|x_p)$ as follows:

$$content(x_m|x_p) = \frac{sim_{com}(x_m, x_p)}{\max_{\forall x_k \in X, x_{k'} \in X; w(e_{x_k x_{k'}})=1} sim_{com}(x_k, x_{k'})} \quad (12)$$

where $w(e_{x_k x_{k'}}) = \begin{cases} 1 & if\ x_k\ cites\ entity\ x_{k'} \\ 0 & otherwise \end{cases}$

Detailed Algorithm

The BibRank algorithm enables us to rank each type of entity (document and author) using a score propagation process between connected entities. The BibRank algorithm steps within an IR task are launched by a query as detailed below:

(1) Initialize document and author scores with an equal probability estimated in the corresponding homogeneous subgraph.
(2) Compute, propagate and normalize relevance scores through the bibliographic network considering transition probabilities, topical-based and content-based scores.
(3) Rank each type of entity according to their score.

The detailed algorithm is presented in Algorithm 1.

Convergence Proof

The convergence of the BibRank algorithm is ensured considering the PageRank convergence (Haveliwala, 1999). Indeed, the BibRank algorithm is based on the PageRank algorithm structure. In BibRank, each entity score computation can be formulated with a matrix equation:

$$G_X = \frac{d}{|V|} e + (1-d)[\lambda_{XX} S_{XX} G_X + \lambda_{YX} S_{YX} G_Y] \quad (16)$$

G_X and G_Y denote respectively entity score vectors of each type of entity X and Y where $X \in \{A, D\}$ and $Y \in \{A, D\}$ with $X \neq Y$. e is the real vector of length $|X|$ corresponding to the number of terms included in X. Each vector element is equal to 1.

Matrices $S_{XX} \in R^{|X|\times|X|}$ and $S_{YX} \in R^{|Y|\times|X|}$ are respectively transition matrices for citation and authorship relationships as detailed below:

$$S_{XX}(j,i) = \frac{w^{x_j}_{x_i}}{O(x_i)}\ with\ j,k \in \{1,\ldots,|X|\} \quad (17)$$

$$S_{YX}(j,k) = \frac{w^{x_j}_{y_k}}{O(x_k)}\ with\ k \in \{1,\ldots,|Y|\}$$

Experimental Evaluation

The main objective of the experimental evaluation is to measure the effectiveness of the BibRank ranking model. The IR tool used for this evaluation, namely indexing, ranking and retrieval effectiveness evaluation, is Terrier (Ounis et al., 2005). We detail in what follows the experimental data set used in the IR task setting, the baselines used for IR effectiveness comparison and then the results obtained.

Experimental Data Sets

It is well known that the evaluation of IR ranking effectiveness requires a document collection and a query test set. The latter comprises both information need descriptions and human relevance assessments. The next section provides the description of the data set used for our experiments.

Document collection. We used the CiteSeerX[4] collection including about 1,4 millions multi-disciplinary bibliographic documents. The data set was extracted on April 2011 using the XML interface of CiteSeerX website. This collection includes titles and abstracts of scientific publications, in addition to some metadata, such as authors and citation relationships. In order to extract the information network, an exact matching was applied on author names. Table 1 shows general statistics of the document data set and the bibliographic network. Analyzing the data set, an author has authored an average of three documents and cites 37 of

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

DOI: 10.1002/asi

ALGORITHM 1. Multi-entity ranking algorithm in a bibliographic network.

Input: Q, $G = <V,E>$ with $V = A \cup D$ and $E = e_{AA} \cup e_{AD} \cup e_{DD}$

Output: Bibrank: $\{Q, G\} \rightarrow \{R_D, R_A\}$

$\Theta \leftarrow 0$;

$$R_D(d_j)^{\Theta} \leftarrow \frac{1}{|D|};$$

$$R_A(a_j)^{\Theta} \leftarrow \frac{1}{|A|};$$

repeat Computing scores propagation algorithm considering transition probabilities, the query input, graph structure and content-based scores

$$R_D(d_j)^{\Theta+1} = \frac{df}{N} + (1-df)\left(\lambda_{AD} \sum_{w(e_{a_l d_i})=1} \frac{R_A(a_l)^{\Theta} \cdot w_{a_l}^{d_i}}{O(a_l)} + \lambda_{DD} \sum_{w(e_{d_k d_i})=1} \frac{R_D(d_k)^{\Theta} \cdot w_{d_k}^{d_i}}{O(d_k)}\right);$$

$R_D(d_i)^{\Theta+1} \leftarrow \mathcal{N}(R_D(d_i)^{\Theta+1})$;

$$R_A(a_j)^{\Theta+1} \leftarrow \frac{df}{N} + (1-df)\left(\lambda_{AA} \sum_{w(e_{a_l a_j})=1} \frac{R_A(a_l)^{\Theta} \cdot w_{a_l}^{a_j}}{O(a_l)} + \lambda_{DA} \sum_{w(e_{d_k a_j})=1} \frac{R_D(d_k)^{\Theta} \cdot w_{d_k}^{a_j}}{O(d_k)}\right);$$

$R_A(a_j)^{\Theta+1} \leftarrow \mathcal{N}(R_A(a_j)^{\Theta+1})$;

$\Theta \leftarrow \Theta + 1$;

until convergence

$R_D \leftarrow Rank(R_D)$;

$R_A \leftarrow Rank(R_A)$;

Return $\{R_D, R_A\}$;

where

- $df \in [0,1]$ is the damping factor. BibRank is a PageRank-like algorithm, we use also the default value of $df = 0.15$.
- $R_D(d_i)^{\Theta+1}$ and $R_A(a_j)^{\Theta+1}$ are respectively the score of document d_i and author a_j at iteration $\Theta + 1$.
- $w_{a_l}^{d_i}$, $w_{d_k}^{d_i}$, $w_{a_l}^{a_j}$ and $w_{d_k}^{a_j}$ are respectively the weighted factor for relationships $e_{a_l d_i}$, $e_{d_k d_i}$, $e_{a_l a_j}$ and $e_{d_k a_j}$. For convenience, the weighted factor $w_{x_i}^{y_j}$ from entity $x_i \in A \cup D$ to entity $y_j \in A \cup D$ is computed as below:

$$w_{x_i}^{y_j} = (r_{x_i}) * (content(y_j | x_i)) \tag{13}$$

- the function $w(e_{x_i y_j})$ denotes the presence or the absence of the relationship from entity x_i to entity y_j with $x_i \in A \cup D$ and $y_j \in A \cup D$.

$$w(e_{x_i y_j}) = \begin{cases} 1 & \text{if the relationship } e_{x_i y_j} \text{ exists} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

- the functions $\mathcal{N}(R_A(a_j))^{\Theta+1}$ and $\mathcal{N}(R_D(d_i))^{\Theta+1}$ normalize entity scores as follows:

$$\begin{aligned} \mathcal{N}\left(R_A(a_j)^{\Theta+1}\right) &= \frac{R_A(a_j)^{\Theta+1}}{\sum_l R_A(a_l)^{\Theta+1}} \\ \mathcal{N}\left(R_D(d_i)^{\Theta+1}\right) &= \frac{R_D(d_i)^{\Theta+1}}{\sum_k R_D(d_k)^{\Theta+1}} \end{aligned} \tag{15}$$

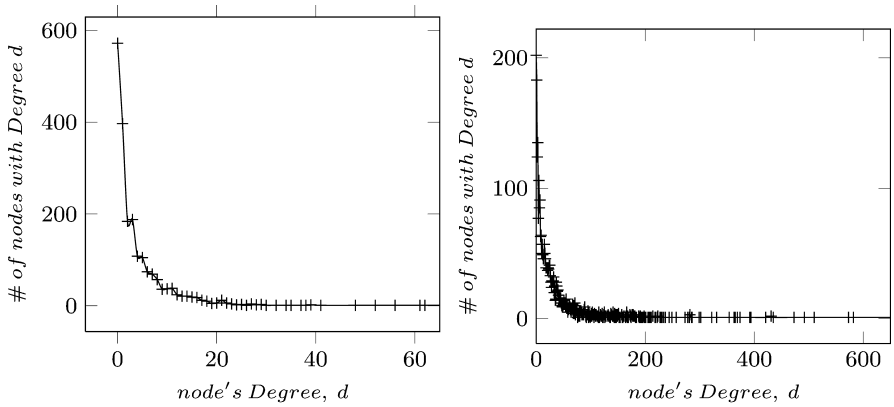
- the ranking function $Rank(R_D)$ ranks the document set according to the values $R_D(d_i)$ for $d_i \in D$. Similarly, the function $Rank(R_A)$ ranks the author set according to the values $R_A(a_j)$ for $a_j \in A$.

his/her colleagues. The number of document authors is about 3 and each document cites about 11 authors.

The density distribution of nodes in each homogeneous subgraph is estimated by the number of in-coming links and is illustrated in Figure 6. We notice that this distribution follows an exponential function.

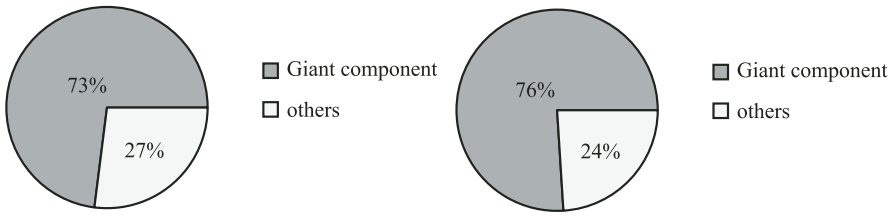
We study the portion of the giant component in the CiteSeerX data set for the different networks based on the citation relationships: document citation links and authorship links. Figure 7 shows the distribution of the giant component in each subnetwork. We can observe that in both subnetworks, the giant component includes more than 73% of the

DOI: 10.1002/asi



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

FIG. 6. Citation network density.



(a) Document citation links (b) Authorship links

FIG. 7. Giant component analysis.

TABLE 1. CiteSeerX collection statistics.

Documents	1,472,735
Authors	1,366,540
Citation links between documents	16,598,502
Citation links between authors	51,306,409
Author citation links per author	37,545
Document citation links per document	11,270
Authorship links	4,209,980
Documents per author	3,081
Authors per document	2,858

network nodes. Thus, we can conclude that entities in the bibliographic network are well connected by citation links viewed as interactions.

Topics. As topics for the CiteSeerX collection are as yet unavailable, we carried out an automatic process for topic generation. For this purpose, we have chosen to use the Latent Dirichlet allocation (LDA) model (Blei et al., 2003) to extract a set of 35 queries from the document titles. This model enables to characterize the data set with topic-distributions. This algorithm computes word-topic distribution $phi_{w|t}$ and document-topic distribution $theta_{d|l}$ that respectively analyze the probability of a word *w* under a topic *t* and the probability of a document *d* under a topic *t*. The LDA algorithm considers some parameters: the number of topics *K*, two free parameters α and β, and the number of iterations *iter*. Regarding parameters α, β and *iter*, we have considered the default values, respectively $\frac{50}{K}$, 0.1 and 2000. The optimal number of topics *K* is contingent to a maximum log-likelihood. This maximum is reached when the probability of the words under the extracted topics is maximal. The log-likelihood $l(nTopic|w,t)$ is estimated as follows:

$$l(nTopic|w,t) = \sum_{w \in W} \log\left(\sum_{t \in K} phi_{w|t}\right) \quad (18)$$

where *W* is the set of words extracted from the data set and *T* is the topic set extracted by LDA algorithm.

Figure 8 highlights the evolution of the log-likelihood according to the number of topics used in the test data set. Even if the curve has not a maximum, we can notice a logarithmic function. So we have chosen to consider 200 topics to counterbalance likelihood gain by execution time gain. Among these 200 topics, some are general and do not really characterize the collection. We extract manually 35 topics among the specific ones and, for each topic,

DOI: 10.1002/asi

TABLE 2. Instances of test topics.

Query	Description	Keywords
Linear algebra and mathematics	General documents about mathematics and linear algebra are waited: matrix and its characteristics. Mathematics applications in technical domains are less relevant than theories	matrix, polynomial, factor, orthogonal, symmetry
Markov chain model	General documents about markov chain model and its derivative models are waited	markov chain model, hidden markov chain, monte carlo
Web services	Documents that deal with web services and internet (architecture and management for example)	internet, web service, architecture
Mobil agent	Documents that deal with mobile agent and speak about autonomy and robot. Agent can be geographically mobile or adapt their behaviour to the situation according to other processes	mobile agent, device, smart, environment, platform
Object identification in pattern recognition	Documents that deal with object identification thanks to pattern recognition. How can be the subject represented, what kind of orientation pattern recognition it can have?	object identification, pattern recognition, classification, representation

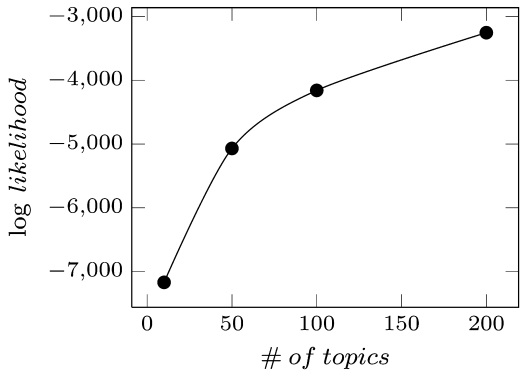


FIG. 8. Maximizing likelihood of query tests.

we generate a query that includes the top representative terms.

Table 2 illustrates some test topics, their description and keywords. For each query, a subgraph is extracted including the most relevant documents and their corresponding authors.

Relevance assessments. As the relevance assessments of both documents and authors are unavailable, we have undertaken a human relevance study described in the following. Considering the use of both topical-based and network-based features in the BibRank algorithm, we combine two binary metrics related to these two binary features to estimate the relevance of each type of entity. The topical one A_{topic} (e_i) is performed through a pooling-based process. The authority-based one $A_{authority}$ (e_i) is attributed automatically to each entity revealing its authority in its homogeneous network. The 3-levels final relevance score R_{Final} $(e_i) \in [0; 2]$ for an entity e_i is estimated as follows:

$$R_{Final}(e_i) = A_{Topic}(e_i) + A_{Authority}(e_i) \tag{19}$$

We detail in what follows how these two intermediary indicators are estimated for documents and authors.

Relevance judgements for documents. The *topical-based indicator* A_{topic} (d_i) is obtained by a pool-based process, close to TREC pooling (Voorhees & Harman, 1998). First of all, for each test query, the ranking of both authors and documents has been computed separately using the baselines introduced above and the BibRank algorithm. The lists of the 20 top document results have been merged. We asked 25 colleagues to assess the relevance of documents in the merged list considering the query topics. Nine of the assessors are assistant professors, 13 are Phd students, two are Master students and one is an engineer. All of them have experience in using search engines. Each topic is evaluated by two different assessors. A score A_{topic} (d_i) between 0 and 1 is assigned to each document $d_i \in D$ to express its relevance to the query: 0 for not relevant and 1 for relevant.

We have analyzed the degree of agreement between assessors for each test topic with the Kappa measure κ (Cohen, 1960). This indicator takes into account the proportion of agreement between assessors $\overline{P}$ and the proportion of expected agreement between assessors by chance $\overline{P}_e$. The Kappa measure κ is equal to 1 if assessors always agree, 0 if they agree only by chance. κ is negative if the agreement between assessors is worse than random. The Kappa measure is computed as follows:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{20}$$

$$with\ \overline{P} = \frac{1}{n}\sum_{i=1}^{r} n_{ii}\ and\ \overline{P}_e = \frac{1}{n^2}\sum_{i=1}^{r} n_{i1}n_{i2}$$

Let n be the total number of assessments supplied by the whole assessors, r be the number of assessment categories (in our case 2 categories: 0 and 1). n_{ii} is the number of agreements between the two assessors for agreement i with $i \in r$, n_{i1} is the total of assessments i given by assessor 1 and n_{i2} is the total of assessments i given by assessor 2.

Figure 9 shows the distribution of the Kappa measure according to the query test set. We notice that the agreement measure ranges from 0.37 to 0.86. The average agreement

DOI: 10.1002/asi

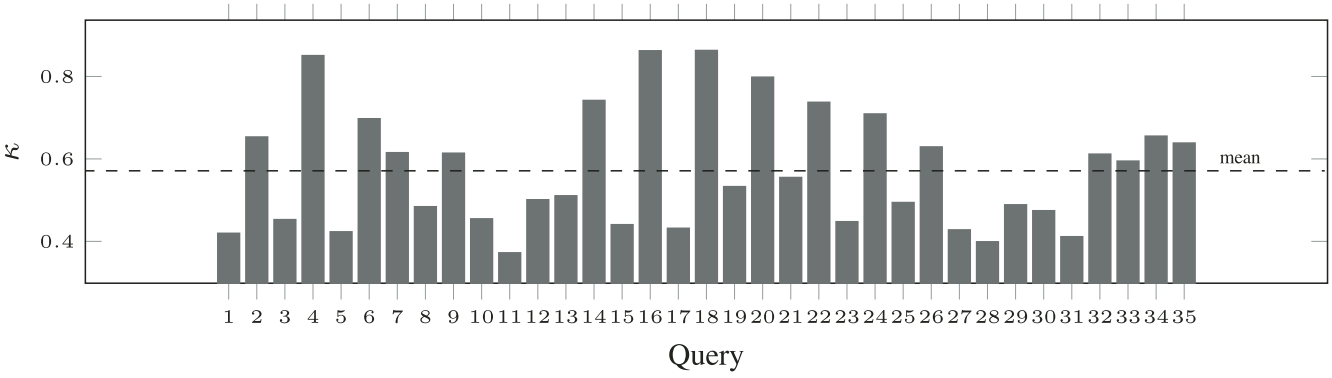


FIG. 9. Distribution of the Kappa measure κ per query. <0 poor agreement, 0.0–0.2 slight agreement, 0.21–0.4 fair agreement, 0.41–0.6 moderate agreement, 0.61–0.8 substantial agreement, 0.81–1 perfect agreement.

measure between assessors is 57.1%, which corresponds to moderate agreement.

The *authority-based indicator* $A_{Authority}(d_i)$ is estimated using a PageRank score classification computed in the document subgraph. Accordingly, a score equal to 1 is assigned to the authoritative documents above the mean PageRank score and 0 to the remaining ones.

Relevance judgements for authors. The *topical-based indicator* $A_{Topic}(a_j)$ is inferred from the document topical assessments. For this purpose, we have built a document pool that merges each document published by the top 20 authors in each ranking list. We have also added to the assessment pool described previously documents authored by the top 20 authors of each author ranking list not already included in the merged list. The whole document set has been assessed in the same way. A topical relevance score is automatically computed for each author as the assessment score average of his/her documents in the collection. The assessment $A_{Topic}(a_j)$ of an author $a_j \in A$ regarding his/her documents $\mathcal{D}(a_j)$ is computed as follows:

$$A_{Topic}(a_j) = \left\lceil \frac{\sum_{d_i \in \mathcal{D}(a_j)} rel(d_i)}{|\mathcal{D}(a_j)|} \right\rceil \quad (21)$$

where $\lceil (x \rceil$ is the ceil function and $|\mathcal{D}(a_j)|$ is the number of documents published by author a_j.

The *authority-based indicator* $A_{Authority}(a_j)$ is estimated using a PageRank score classification computed in the author subgraph. Accordingly, score equal to 1 is assigned to the authoritative authors above the mean PageRank score and 0 to the remaining ones.

Evaluation Measures and Baselines

For effectiveness measurement purposes, we used the Normalized Discount Cumulative Gain (NDCG) measure (Järvelin & Kekalainen, 2002) that considers relevant documents' position for the n-top results compared with the perfect ranking that we should obtain.

The BibRank ranking model is compared to the following state-of-the-art ranking ones:

- *BM25 textual similarity-based model* denotes the well known probabilistic IR model (Robertson & Walker, 1994). The BM25 relevance $RSV(e,Q)$ between the query Q and an entity e, either document d_i or author a_j is computed as follows:

$$RSV(e,Q) = \sum_{t \in Q \cup_e} \frac{(k_3+1)c(t,Q)}{k_3 + c(t,Q)} \frac{(k_1+1)c(t,Q)}{k_1\left(1-b+b\frac{|e|}{avdl}\right)+c(t,Q)} \log\left(\frac{N+1}{ef(t)+0.5}\right)$$

  with k_1, k_3 and b are free parameters respectively fixed to the default values 1.2, 8 and 0.75. The occurrence number of term t in the query Q is estimated by $c(t,Q)$. The number of terms included in entity e is noted $|e|$ and $avdl$ represents the average number of terms included in entities of the same type. N denotes the total number of entities and $ef(t)$ the number of entities including t.
- *Hiemstra textual similarity-based model:* denotes the traditional language based IR model (Hiemstra, 1998). Our motivation behind comparing to Hiemstra model is that this latter is the basis of the relevance scoring in BibRank. The relevance score $RSV(e,Q)$ between an entity e and a query Q computed with the Hiemstra model is estimated as follows:

$$RSV(e,Q) \simeq \prod_{t \in Q} \lambda P(t_i|M_e) + (1-\lambda) P(t_i|M_c) \quad (22)$$

  where $P(t_i|M_e)$ and $P(t_i|M_c)$ are relevance scores of term t_i according respectively to the entity language model, namely document or author language model defined in section 4, and the collection language model.
- *Structure-based ranking model (PRank):* denotes a retrieval model that ranks heterogeneous entities in a bibliographic network using a score propagation principle proposed in Yan and Ding (2010). The PRank algorithm is designed for ranking authors, documents and journals. We have

DOI: 10.1002/asi

ALGORITHM 2. Prank.

Input: $G = <V,E>$ with $V = A \cup D$ and $E = e_{AA} \cup e_{AD} \cup e_{DD}$
Output: PRank:
$\{G\} \rightarrow \{R_D, R_A\}$
$R_D(d_i) = 1$;
$R_D = PageRank(R_D, e_{DD})$;
repeat

$$R_A(a_j) \leftarrow \sum_{w(e_{d_k d_i})=1} R_D(d_k);$$

$$R_D(d_i) \leftarrow \sum_{w(e_{d_k d_i})=1} \frac{R_A(a_j)}{\sum_{a_l \in A} R_A(a_l)};$$

$R_D = PageRank(R_D, e_{DD})$;
until convergence
$R_D \leftarrow Rank(R_D)$;
$R_A \leftarrow Rank(R_A)$;
Return $\{R_D, R_A\}$;
where

- the function PageRank(R_D, e_{DD}) computes the PageRank algorithm through the document citation network considering the initial score R_D,
- the ranking function $Rank(R_D)$ ranks the document set according to the values $R_D(R_i)$ for $d_i \in D$. Similarly, the function $Rank(R_A)$ ranks the author set according to the values $R_A(a_j)$ for $a_j \in A$.

implemented this algorithm and used it for ranking only authors and documents. As experiments in Yan and Ding (2010) show that time feature has no impact on the retrieval effectiveness of PRank algorithm, we voluntarily do not consider this feature. By this way, document ranking depends only on document citation links and authorship links whereas author ranking depends only on authorship links. More specifically, we have implemented the algorithm detailed in Table 4.

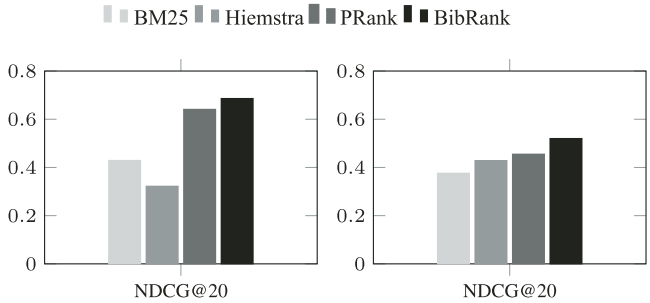
Retrieval Effectiveness Evaluation

The experiments focus here on comparative evaluation of BibRank effectiveness with the baselines described in section 6.2. Table 3 and Figure 10 show the results obtained using the baseline models (BM25, Hiemstra, and PRank) and the BibRank model, for both authors and documents; the improvements achieved using the BibRank model (% change) are computed and significance tested using the student t-test.

We can notice a significant improvement with the BibRank algorithm for both author and document rankings regarding the three baselines. Improving textual similarity-based retrieval models, such as BM25 and Hiemstra, proves that integrating both the graph structure and the topical relevance in the joint relevance scoring model is effective. Compared to the PRank algorithm, BibRank includes a content-based score that estimates the topical relatedness between connected entities. We notice that including this feature increases the NDCG metric around 7% for document ranking and around 14% for author ranking. We notice

TABLE 3. Retrieval effectiveness with the NDCG@20 measure and significance. % change: BibRank improvement. Student test significance *: $0.01 < t \leq 0.05$; **: $0.001 < t \leq 0.01$; ***: $t \leq 0.001$.

Model	NDCG@20	% Change
BM25	0.429	**+59.77**%***
Hiemstra	0.322	**+113.13**%***
PRank	0.641	**+7.03**%*
BibRank	**0.686**	
Model	NDCG@20	% change
BM25	0.376	**+38.26**%***
Hiemstra	0.428	**+21.47**%**
PRank	0.455	**+14.29**%*
BibRank	**0.520**	



(a) Document Ranking (b) Author Ranking

FIG. 10. Comparing retrieval effectiveness.

however that improvement compared to the PRank is less important compared to the BM25 and Hiemstra models making so in advance the impact of the graph structure on entity ranking. Nevertheless, features considered in BibRank algorithm enable to increase ranking including a content analysis.

However, we notice that the BibRank improvement for author ranking is weaker than for document ranking. A possible explanation is that document topical relevance was computed by human judges whereas author topical relevance was inferred from document topical relevance.

Figures 11 and 12 illustrate, respectively, the NDCG curves for document and author rankings between 1 and 20. We can see that the BibRank curve rises above the baseline ones, particularly for ranks prior to rank 5. That means BibRank ranks the most relevant documents in the top.

Moreover, Figures 11 and 12 highlight that the NDCG value of the BibRank ranking declines for document ranking and increases for author ranking. These trends imply that a good ranking is more likely for documents, maybe because of the way of modeling authors by aggregating authored documents.

We have listed in Table 4 and Table 5, respectively, the top five documents and top five authors obtained by BibRank algorithm for the topic "object identification in pattern recognition." For each document (respectively

DOI: 10.1002/asi

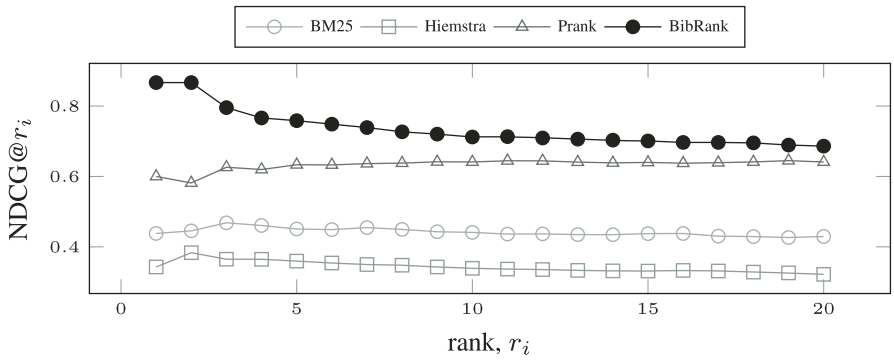


FIG. 11. NDCG at different document ranking.

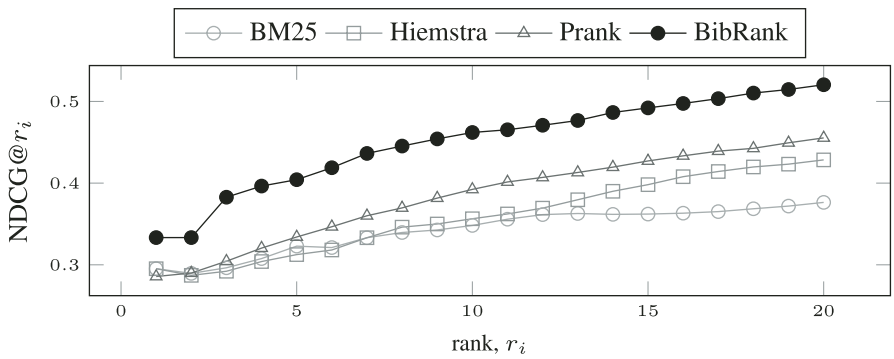


FIG. 12. NDCG at different author ranking.

TABLE 4. Ranks of the top 5 documents in BibRank and baseline rankings for "object identification in pattern recognition" query.

	Rank			
Title	BM25	Hiemstra	PRank	BibRank
Probabilistic object recognition and localization	304	372	47	1
Discriminant analysis for recognition of human face images	584	93	109	2
Gait-based human identification from a monocular video sequence	568	780	90	3
3D model enhanced face recognition	62	200	641	4
Is combining useful for dissimilarity representations?	434	557	120	5

TABLE 5. Ranks of the top 5 authors in BibRank and baseline rankings for "object identification in pattern recognition" query.

	Rank			
Title	BM25	Hiemstra	PRank	BibRank
Alex Pentland	—	—	454	1
Rama Chellappa	—	—	46	2
Aravind Sundaresan	—	—	504	3
Robert P.W. Duin	—	—	107	4
Kamran Etemad	—	—	496	5

author), we have listed its title (respectively author name) and ranks obtained in the three chosen baselines. For authors, the BM25 and Hiemstra models do not compute a rank. These authors are in fact well cited, more than 1,000 in-coming citation links, or are the authors of at least one of the top five documents.

Conclusion and Future Work

In this article, we proposed a bi-type entity ranking algorithm that aims to rank jointly documents and authors in a bibliographic network regarding a topical query. More specifically, the BibRank ranking model relies mainly on evidence sources issued from both content-based and network-based features. These features allow us to have a sense of the appropriateness of author's scientific production and document topic regarding the general description of the subject research held by the query. According to PageRank general form, partial scores are aggregated and propagated through the heterogeneous network. Experiments undertaken on the CiteSeerX collection demonstrate the effectiveness of the BibRank algorithm in comparison to state of the art ranking models. Improvements achieved using automatic generated queries based on LDA algorithm are estimated between 7% and 113% for document ranking and between 14% and 38% for author

DOI: 10.1002/asi

ranking. The experimental results are thus promising. However, they should be considered with care. Indeed, the main limit of the empirical evaluation design model consists of the availability of the citation data that need to be extracted from textual content and matched with preexisting documents in the database. Therefore, the quality of the citation network would depend on the quality of the extraction tool and the size of the database.

For short-term future work, we plan to extend this model to larger bibliographic networks, including more types of entities, such as conference proceedings and attendees and consequently more semantic relationships between entities. This extension may conduct to integrate more specific social relevance features such as social distance between entities. Moreover, we would like also to apply the BibRank model in other application domains in addition to literature access typically from social applications on the web and collaborative communities.

For long-term future work, we plan to investigate another task in literature access area, namely, identifying potential collaborators and locating innovative authors and group works. We believe that additional social network analysis methods and algorithms should be considered in order to model the semantic and the strength of the social relations between the heterogeneous entities.

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Cabanac, G. (2012). Shaping the landscape of research in information systems from the perspective of editorial boards: A scientometric study of 77 leading journals. Journal of the American Society for Information Science and Technology, 63(5), 977–996.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1(20), 213–220.

Coyle, M., & Smyth, B. (2008). (web search) shared: Social aspects of a collaborative, community-based search network. In Proceedings of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH '08 (pp. 103–112). Berlin: Springer-Verlag.

Egghe, L. (2006). An improvement of the h-index: The g-index. ISSI Newsletter, 2, 8–9.

Emmerink, R. (1993). Effects of information in road transport networks with recurrent congestion. Technical report.

Garfield, E. (1955). Citation indexes for science. A new dimension in documentation through association of ideas. Science, 122, 1123–1127.

Hanneman, R.A., & Riddle, M. (2005). Introduction to Social Network Methods. Riverside, CA: University of California, Riverside.

Haveliwala, T. (1999). Efficient computation of pagerank. Technical Report 1999-31, StanfordInfoLab.

Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. In Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, ECDL '98 (pp. 569–584). London: Springer-Verlag.

Hirsch, J.E. (2005). 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.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2011). Using bayesian networks to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals. Scientometrics, 89(2), 523–551.

Jabeur, L.B., Tamine, L., & Boughanem, M. (2010). A social model for literature access: towards a weighted social network of authors. In Proceedings of the 9th International Conference Recherche d'Information Assistée par Ordinateur: Adaptivity, Personalization and Fusion of Heterogeneous Information, RIAO '10 (pp. 32–39). Paris, France: Le centre de hautes études internationales d'informatique documentaire.

Järvelin, K., & Kekalainen, J. (2002). Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems, 20(4), 422–446.

Kirsch, S.M., Gnasa, M., & Cremers, A.B. (2006). Beyond the web: Retrieval in social information spaces. In Proceedings of the 28th European conference on Advances in Information Retrieval Research, ECIR'06 (pp. 84–95).

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

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22, 79–86.

Kurland, O., & Lee, L. (2005). Pagerank without hyperlinks: structural re-ranking using links induced by language models. In ACM SIGIR Special Interest Group on Information Retrieval, SIGIR'05 (pp. 306–313).

Kurland, O., & Lee, L. (2006). Respect my authority!: Hits without hyperlinks, utilizing cluster-based language models. In ACM SIGIR Special Interest Group on Information Retrieval, SIGIR'06 (pp. 83–90).

Lempel, R., & Moran, S. (2000). The stochastic approach for link-structure analysis (salsa) and the tkc effect. Computer Networks, 33, 387–401.

Liu, X., Bollen, J., Nelson, M.L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. Information Processing & Management, 41, 1462–1480.

Minkov, E., & Cohen, W. W. (2006). An email and meeting assistant using graph walks. In Proceedings of the 3rd Conference on Email and Anti-Spam, CEAS'06.

Nie, L., Davison, B.D., & Qi, X. (2006). Topical link analysis for web search. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06 (pp. 91–98). New York: ACM.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, B., & Johnson, D. (2005). Terrier information retrieval platform. In Proceedings of the 27th European Conference on IR Research, ECIR '05 (pp. 517–519). Springer.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Pearson, K. (1905). The problem of the random walk. Nature, 72(1867), 342–342.

Ponte, J.M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98 (pp. 275–281). New York: ACM.

Robertson, S.E., & Walker, S. (1994). 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 (pp. 232–241. New York: Springer-Verlag.

Roy, S., Lane, T., & Werner-Washburne, M. (2008). Integrative construction and analysis of condition-specific biological networks. In Proceedings of the 23rd National Conference on Artificial Intelligence, Volume 3 (pp. 1867–1868). Palo Alto, CA: AAAI Press.

Sayyadi, H., & Getoor, L. (2009). Futurerank: Ranking scientific articles by predicting their future pagerank. In Proceedings of the 9th SIAM International Conference on Data Mining, SIAM'09 (pp. 533–544).

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In Proceedings of the 2008 8th IEEE International Conference on Data Mining (pp. 1055–1060). Washington, DC: IEEE Computer Society.

Uddin, S., Hossain, L., Abbasi, A., & Rasmussen, K. (2012). Trend and efficiency analysis of co-authorship network. Scientometrics, 90(2), 687–699.

Voorhees, E.M., & Harman, D. (1998). The text retrieval conferences (trecs). In Proceedings of a Workshop on TIPSTER '98, (pp. 241–273). Stroudsburg, PA: Association for Computational Linguistics.

Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2006). Ranking scientific publications using a simple model of network traffic. Society, 1–5.

Yan, E., & Ding, Y. (2010). 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 (pp. 1–7). Silver Springs, MD: American Society for Information Science and Technology.

Yang, Z., Hong, L., & Davison, B.D. (2010). Topic-driven multi-type citation network analysis. In Proceedings of the 9th International Conference Recherche d'Information Assistée par Ordinateur: Adaptivity, Personalization and Fusion of Heterogeneous Information, RIAO '10 (pp. 24–31). Paris, France: Le centre de hautes études internationales d'informatique documentaire.

Zhang, C.-T. (2009). The e-index, complementing the h-index for excess citations. PLoS ONE, 4.

Zhang, J., Tang, J., Liang, B., Yang, Z., Wang, S., Zuo, J., & Li, J. (2008). Recommendation over a heterogeneous social network. In Proceedings of the 2008 The 9th International Conference on Web-Age Information Management, WAIM '08 (pp. 309–316). Washington, DC: IEEE Computer Society.

Zhao, D. (2006). Towards all-author co-citation analysis. Information Processing & Management, 42, 1578–1591.

Zhou, D., Orshanskiy, S.A., Zha, H., & Giles, C.L. (2007). Co-ranking authors and documents in a heterogeneous network. In Proceedings of the 7th IEEE International Conference on Data Mining (pp. 739–744). Washington, DC: IEEE Computer Society.

DOI: 10.1002/asi