Topic-based Heterogeneous Rank

Tehmina Amjad[a,b], Ying Ding[a,d], Ali Daud[b], Jian Xu[c], Vincent Malic[a]

[a] School of Library and Information Science, Indiana University, Bloomington, Indiana, USA
[b] Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan
[c] School of Information Management, Sun Yat-sen University, Guangzhou, China
[d] Library, Tongji University, Shanghai, China

Abstract

Topic-based ranking of authors, papers and journals can serve as a vital tool for identifying authorities of a given topic within a particular domain. Existing methods that measure topic-based scholarly output are limited to homogeneous networks. This study proposes a new informative metric called Topic-based Heterogeneous Rank (TH Rank) which measures the impact of a scholarly entity with respect to a given topic in a heterogeneous scholarly network containing authors, papers and journals. TH Rank calculates topic-dependent ranks for authors by considering the combined impact of the multiple factors which contribute to an author's level of prestige. Information retrieval serves as the test field and articles about information retrieval published between 1956 and 2014 were extracted from Web of Science (WoS). Initial results show that TH Rank can effectively identify the most prestigious authors, papers and journals related to a specific topic.

Keywords: Topic-based rank, Topic sensitive ranking, Heterogeneous networks, Topic modeling

1. Introduction

Citation analysis has often been used as a formal metric for the ranking of academic entities (i.e., author, paper or journal). This type of analysis aggregates the citations received by an academic entity and uses the resulting aggregate as a metric for that entity's impact. These methods cannot discriminate citations received from important journals from ones received from less important journals. All citations are given equal weight and hence they are treated as equally important, even though all citations arguably do not transmit same amount of prestige. Pinsky and Narin (1976) argue that more weight should be given to citations from prestigious journals than from peripheral ones.

The PageRank algorithm (Page et al. 1999) pioneered the concept of weighted nodes and introduced ranking of nodes based on link structure. PageRank and its variants have been used for measuring scholarly impact of authors (Liu et al. 2005; Ding 2011a; Yan & Ding 2009; Yan & Ding 2011; Fiala et al. 2008), journals (Bollen et al. 2006), and publications (Yan & Ding 2010a) by separating prestigious units from peripheral ones. These methods treat scholarly networks as homogeneous networks where all nodes belong to one type (i.e., author, journal or paper). Ding et al. (2014) provides a handbook for analyzing scholarly communication and the informetrics used for the evaluation of scholarly impact. However, in scholarly networks entities like authors, papers and journals/conferences are not independent, and their influence must be incorporated while

ranking them. Authors gain rank via their papers and papers are influenced by the venue they are published in and vice versa.

Academic entities such as papers, authors, and journals can be ranked simultaneously in heterogeneous networks by considering the impact of each entity type on the other (Su & Han, 2013). Yan et al. (2011) proposed the P-Rank algorithm, which calculates the impact of authors in a heterogeneous network. They did not consider, however, the topical information of a node in this network while ranking it. This can lead to undesirable results because papers that are important in one field may not be important in another field; similarly, authors who are experts in one field may not be experts in another field.

We anticipate that ranking of academic entities within the context of topics will grow in importance as trans-disciplinary collaboration continues to become a norm in scholarly practice. Trans-disciplinary collaboration is particularly essential when it comes to facing the challenges of analyzing big data. Within the field of information retrieval, for example, there are numerous diverse subfields such as query processing, medical information retrieval, database processing, and multimedia retrieval. A single scholar cannot be an expert in all of these subfields. Some topic-based variants of PageRank have been proposed (Haveliwala 2002; Ding 2011b) but these variants do not consider the heterogeneity of a scholarly network. Although Yan et al. (2011) did consider the heterogeneity of network, their method is topic insensitive.

This paper addresses the topic-based prestige of academic entities in a heterogeneous network and considers the mutual influence of authors, papers and venues such as journals and conferences. We propose a Topic-based Heterogeneous Rank (TH Rank) algorithm to measure scholarly impact of an academic entity with respect to its topical context in a heterogeneous network. TH Rank was applied on papers published from 1956 to 2014 in the field of information retrieval, which produced a ranking of the collected authors, papers and journals. The structure of this paper is as follows: Section 2 reviews related literature; Section 3 gives details of the data collection process and the proposed TH Rank method; Section 4 discusses the results; and Section 5 provides our conclusions and suggestions for future work.

2. Related Work

In this section we survey existing types of ranking methods and summarize their general features. This summarization is depicted in Table 1.

2.1 Ranking Methods for different networks and weight measures

Growth of interaction between scholars over the past decades has resulted in increasingly complex academic networks. This, in turn, has prompted a wide variety of methods for analysis of these networks. These studies have applied the PageRank algorithm or extended versions of it to analyze co-authorship networks (Liu et al. 2005; Yan & Ding 2009; Liu et al. 2007), author citation networks (Radicchi et al. 2009), paper citation networks (Chen et al. 2007; Ma et al. 2008), journal citation networks (Bollen et al. 2006; Leydesdorff 2007, 2009), and author co-citation networks (Ding et al. 2009). The Science Author Rank Algorithm (SARA) presented by Radicchi et al. (2009) is used for analyzing a weighted author citation network. SARA monitors the number of authors along with their indegree and instrength distribution. Indegree is the number of incoming

links of a node *i*, while instrength is a separate measure that takes into account the weight of incoming edges. Behind this form of weighting is an idea that authors own a certain amount of discrete units of credit which they distribute among their neighbors, proportioned to the weight of the directed connection. Ding et al. (2009) have proposed a weighted version of PageRank in which the number of publications or the number of citations of a given author can be used as added weights to an existing PageRank co-citation network. Another weighted version of PageRank, proposed by Yan and Ding (2011) uses citation count as a weight for a co-authorship network.

2.2 Heterogeneous Network-based Methods

Methods discussed so far either consider the weights of citing papers, citing authors and citing journals individually but do not consider these factors simultaneously, as they are applied only to homogenous networks. A number of studies have, however, tried to combine two different types of networks. The Co-Ranking framework (Zhou et al. 2007) combined co-authorship networks and paper citation networks, creating a paper-author matrix. The Future-Rank framework (Sayyadi & Getoor 2009) combines co-authorship and paper citation networks by finding the PageRank values for each article in a paper citation network, and then calculating values for authors in a paper-author matrix using the HITS (Kleinberg 1999) algorithm. Li and Tang (2008) have introduced a heterogeneous network for modeling papers, authors and locations simultaneously in a single heterogeneous network for the purpose of identifying field experts. Yan et al. (2011) have used the same heterogeneous network for measuring prestige.

2.3 Topic-based Methods

Topic-sensitive PageRank (2002) was proposed by Havelilwala, which pre-calculates the PageRank score vectors for 16 different queries selected from the Open Directory Project (ODP). They then added topic-sensitive personalized vectors to the random jump part of the original PageRank formula. At query time, similarity measures of the query with respect to each of these 16 vectors were calculated, and topic-sensitive PageRank vectors were weighted based on similarity. By making PageRank topic-sensitive, Havelilwala resolves the problem of unrelated pages obtaining high ranks by virtue of their high number of in-links.

Pal and Narayan (2005) proposed a web-surfer model in which they made changes to the network part of the PageRank algorithm. In this model, a web-surfer will prefer to choose from links on same topic, and there will be lesser probability to select a link which is on a different topic. This method does not consider the random jump component of PageRank. Richardson and Domingos (2001) proposed an intelligent surfer model in which they have added topic sensitivity to both the network part and the random jump part of the PageRank model. It is also a query specific model in which a surfer will only follow a query-relevant link. This model suffers from efficiency issues due to finding the query-specific scores at runtime.

Yang et al. (2009) employed the Latent Dirichlet Allocation topic modeling algorithm to find the topic distribution for each document in a dataset. In this model, the random surfer not only can randomly jump to new related pages but can also follow links on visited pages that are related to query topics. Limitation of this paper is that it can provide the topic distribution only at the document level but not at the author or journal level. Ding (2011b) applied an extended LDA topic

modeling algorithm and proposed a topic-based PageRank which considers the topic distribution for authors as well as documents. Two methods were proposed for topic-based ranking: (1) a simple combination of LDA and PageRank and (2) a topic-based random walk. The topic-based methods considered the rank of general web entities while Ding's method was specifically designed for the ranking of authors in a co-citation network.

We now summarize important features covered by the methods studied from literature. The weighted methods are capable of measuring the author's productivity (by publications) or quality (by citations). These methods also study the effect of different damping factors. Some methods can also incorporate a temporal dimension which is important for finding changes in interest of scientists, predicting future citations, and giving more weight to recent works. The heterogeneous methods can simultaneously rank related academic entities like authors, papers and journals. These entities have influence on each other, hence ranking them simultaneously is significant. The topic-based methods ranks academic entities with respect to their topic. However most methods described in the literature are dedicated to ranking general web pages, not academic entities. Table 1 provides a summary of the related methods included in this study.

Table 1. Summary of the related work

Reference #	Proposed	Finding	Limitation
Liu et al. 2005	Author rank for a weighted directional network	Co-authorship frequency is used as weighted measure and Author-Rank correlates with PageRank	Topic insensitive and cannot deal with multiple academic entities
Liu et al. 2007	A model and structure for weighted network of research areas (WNRA)	Study of distance, centrality, clustering coefficient and betweenness for requisition papers	Specific for one type of entity (papers), and topic insensitive
Yan & Ding 2009	Centrality measures for impact analysis	Centrality measures are significantly correlated with citation counts. Micro level study of network is useful for impact analysis	Deals with authors only and topic insensitive
Radicchi et al. 2009	Science Author Rank Algorithm (SARA)	Credits are exchanged by authors and Proposed method performs better than Citation count and Balanced citation count	Ranking of papers only and topic insensitive
Chen et al. 2007	PageRank for physical review	Similarities and differences between WWW and citation network. Study of relative importance of papers	Ranking of authors only and topic insensitive
Ding et al. 2009	A weighted PageRank	Proposed method co-relates with PageRank. Effect of different damping factors	Ranking of authors only. Effect of publication venue is ignored. Topic insensitive
Yan & Ding 2011	A weighted PageRank	Proposed method outperforms PageRank. Effect of different damping factors	
Zhou et al. 2007	Method for co-ranking	Co-Ranking of authors and papers is better than counting publications and citations	Effect of publication venue is ignored
Sayyadi & Getoor 2009	Future Rank, with incorporation of publication date	Ranking of authors by predicting their future citations	Topic Insensitive
Li & Tang 2008	Method to integrate temporal information into random walk	Including the time information can improve the ranking performance	
Yan et al. 2011	Prestige by weighted citations	Impact of quantification of citations for measuring prestige	
Havelilwala 2002	Method for ranking of web pages using 16 topic-based vectors	Effect of precomputed topic-based vectors to generate topic-based results	These methods are addressing general web pages, and cannot deal with academic entities
Pal & Narayan 2005	Model for general web surfer w.r.t author's interest	Use of web surfer's history to find his or her topic of interest.	
Richardson & Domingos 2001	Directed surfer model	Use of probabilistic model to find relevance of a query to page	
Yang et al. 2009	Probabilistic model for random walk	Use of topic models for queries, and its integration into random walk	

Ding 2011b	Combination of topic modeling and weighted pagerank	Use of topic modeling for adding weights to PageRank	Cannot address related academic entities simultaneously

The previous studies have either applied weighted methods to rank entities, used homogeneous networks, or ignored the topic information. This study involves topic-based modeling of scholarly networks by simultaneously modeling a heterogeneous network of authors, papers and journals.

3. Methodology

3.1 Data Collection

We selected the field of information retrieval as our test field. Papers and their cited references were collected from the Web of Science (WoS) from the period from 1956 to 2014. Search strategies were based on the following terms (including plurals and variants) which were determined by checking Library of Congress Subject Headings and consulting several domain experts: INFORMATION RETRIEVAL, INFORMATION STORAGE and RETRIEVAL, QUERY PROCESSING, DOCUMENT RETRIEVAL, DATA RETRIEVAL, IMAGE RETRIEVAL, TEXT RETRIEVAL, CONTENT BASED RETRIEVAL, CONTENT-BASED RETRIEVAL, DATABASE QUERY, DATABASE QUERIES, QUERY LANGUAGE, QUERY LANGUAGES, and RELEVANCE FEEDBACK. In total, we collected 20,359 papers with 44,770 distinct authors, 558,498 citations relationships and 3,270 unique journals.

Self-citations in a citation network can be categorized in three different levels (1) author level (Aksnes 2003; Hyland 2003; Glänzel & Thijs 2004) (2) journal level (Tsay 2006; Krauss 2007; Frandsen 2007) and (3) research group level (Van Raan 2008; Hendrix 2009). The issue of self-citations has been handled in different ways. Some indices, such as Journal Impact Factor (Garfield, 1999) do not consider the issue of self-citations while others like Eigenfactor (Bergstrom et al. 2008; West et al. 2010) exclude journal self-citations only. Measures like the F-index (Katsaros et al. 2009) try to neutralize this effect by finding more exclusive numbers of authors in the citation network. Although self-citation is susceptible to possibility of manipulation, they still can be considered a sincere form of citing activity. If an author, for example, is consistently building upon previous work, citing his or her own work should be considered a vital and expected part of the process. This study considers the effect of self-citation by assigning lower weights to such activity. TH Rank assigns full weight (1) to regular citations, 0.5 to journal self-citations, and 0.25 to author self-citations as in the Heterogeneous Rank model (Yan et al. 2011).

`

3.2 Topic-based Heterogeneous Rank (TH Rank)

Topic-based Heterogeneous Rank (TH Rank) measures the prestige of academic entities in heterogeneous networks. Following assumptions have been followed in order to give weights to citations.

1. Articles are important if they are cited by important articles (Ding & Cronin 2011; Yan et al. 2011; Chen et al. 2007; Ma et al. 2008; Maslov & Redner 2008).
2. Authors have more impact if their articles receive citations from important articles, and similarly articles are also more important if they are cited by more prestigious authors (Zhou et al. 2007; Sayyadi & Getoor 2009; Yan et al. 2011).
3. Journals have higher impact if their articles receive citations from important articles, and similarly articles are also more important if they are being cited by more prestigious journals (Pinski & Narin 1976; Cronin 1984; Davis 2008; Yan & Ding 2010a; Yan et al. 2011) .

TH Rank measures the prestige of an article by measuring the impact of the articles, authors, and journals that cite it. The TH Rank of authors can subsequently be calculated from the TH Rank of his or her publications, and the TH Rank of journals can be calculated by the rank of articles published in it. Topic distribution is also taken into account when measuring rank of each entity. First, we calculate the topic-based ranks of all the papers in our dataset. Then, topic-based personalized vectors are calculated for authors and journals; these vectors consider the topic interest of authors-papers for the case of authors, and journal-papers for journals.

3.2.1 Topic Modeling

Latent Dirichlet Allocation (LDA) provides a probabilistic model for a document's latent topic layer (Blei et al. 2003). For each document *d*, a multinomial distribution θ_d over topics is sampled from a Dirichlet distribution with parameter π. For each word w_{di}, a topic z_{di} is chosen from the topic distribution. A word w_{di}, is generated from a topic-specific multinomial distribution ϕ_{zdi}. One way to conceptualize the reasoning behind LDA is to imagine that before writing a paper the author first selects particular topics and then uses the words that have a high probabilistic association with these topics in the writing of his or her text (Ding, 2011b).

The Author-Conference-Topic (ACT) model proposed by Tang et al. (2008) was applied to capture document content, author interests, and journal topics. A total of 10 topics were extracted using the ACT model. With each topic there is a list of words associated with this topic and the authors, papers, and journals are ranked according to their topic distribution probabilities in each topic. The ACT model calculates the interest distribution of each author with respect to the ten extracted 10 topics (i.e., probability of a topic for a given author: P(t|a)). For example, if there are three topics, author A has a probability of 0.351 for topic 1, 0.298 for topic 2, and 0.351 for topic 3, and sum of topic distribution across 3 topics is 1.0. The ACT model also calculates the author's probability distribution for a given topic (i.e., P(a|t) =P(a)P (t|a)/P(t)). For example, there are 44,770 authors and each author will have a topic distribution on topic 1. The sum of all the author-topic distributions over topic 1 would be 1.0. The average probability for each author-topic distribution over topic 1 would be 1/44770. We also used ACT to calculate topic distribution over documents. We have a total of 20,359 papers and each paper has a distribution over 10 topics which also sums to 1. Similarly we used ACT to generate topic distribution of 3,270 journals across the 10 topics.

3.2.2 Network Structure

The heterogeneous network under consideration is composed of three networks, shown in Figure 1. We begin with a directed paper citations graph $G_P = (V_P, E_P)$, where V_P is the set of papers/articles and the directed edge $(pi, pj) \in E_P$ indicates that article *pi* cites article *pj*.

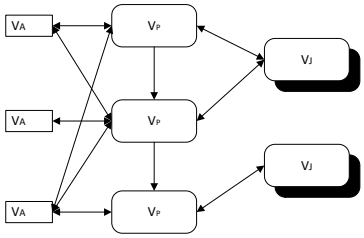

Figure 1. The heterogeneous scholarly network for TH Rank

To represent author-paper relationships, we have a bipartite graph $G_{P\text{-}A} = (V_P \cup V_A, E_{P\text{-}A})$, where V_A is the set of authors, V_P is the set of papers, and their edges are represented as $E_{P\text{-}A}$.

Relationships between articles and their publication venues (journal/conference) are modelled by the bipartite graph $G_{P\text{-}J} = (V_P \cup V_J, E_{P\text{-}J})$, where V_J is the set of publication locations, V_P is the set of papers, and the venue-paper relationship is recorded in $E_{P\text{-}J}$.

We combine these different graphs to form a heterogeneous graph centered by the citation network: $G(V, E) = (V_P \cup V_A \cup V_J, E_P \cup E_{P\text{-}A} \cup E_{P\text{-}J})$.

The proposed heterogeneous academic network contains three walks: an intra-class walk within the paper citation network G_P and two inter-class walks, one between papers and authors in $G_{P\text{-}A}$ and the other between papers and journals in $G_{P\text{-}J}$. PageRank is used as the underlying algorithm for the intra-class walk. Let M^P be the *nxn* matrix for the paper citation matrix, where *n* is the number of nodes/articles in the network:

$$M_{i,j}^{P} = \begin{cases} 1 \; if \; paper \; i \; cites \; paper \; j \\ 0 \qquad\qquad otherwise \end{cases} \qquad (1)$$

Furthermore, for any paper *pi* which does not cite any other paper in the dataset, we define $M_{i,j}^{P} = 1$. In this way, we create virtual links from dangling nodes to every other node.

In addition, we define two more adjacency matrices to define inter-class walks on bipartite graphs. M^A is the *nxm* paper-author adjacency matrix, where *n* is the number of articles and *m* is the number of authors:

$$M_{i,j}^{A} = \begin{cases} 1 \; if \; author \; j \; writes \; paper \; i \\ 0 \qquad\qquad otherwise \end{cases} \qquad (2)$$

This matrix is used to link the citing authors to citing articles. Similarly, M^J, is the *nxq* article-journal adjacency matrix, where *n* is the number of articles and *q* is the number of journals:

$$M_{i,j}^{J} = \begin{cases} 1 \; if \; article \; i \; is \; published \; on \; journal \; j \\ 0 \qquad\qquad otherwise \end{cases} \qquad (3)$$

From the ACT model we have the topic distribution probability of papers, authors and journals for each topic. We have normalized them to make sure that each column sums up to 1. For papers we have a *NPT* (Normalized Paper Topic) matrix, which is an *nxT* matrix where *n* is number of papers and *T* is number of topics. Each column of this matrix represents topic probability of a paper in the corresponding topic. Column 1 represents the personalized ranks based on the topic distributions of all papers in topic 1, column 2 represents personalized ranks based on the topic distributions of all papers in topic 2 and so on. These personalized rank vectors are used as initial ranks while ranking papers.

Similarly, we have an *mxT* matrix *NAT* (Normalized Author Topic) where *m* is number of authors and *T* is number of topics, and *qxT* matrix *NJT* (Normalized Journal Topic) where *q* is number of journals and *T* is number of topics. Each column of this matrix represents the topic probability of an author or journal in the corresponding topic, respectively. Column 1 represents the personalized ranks based on topic distributions of all authors/journals in topic 1, column 2 represents the personalized ranks based on the topic distributions of all authors/journals in topic 2 and so on. These personalized rank vectors are used as initial ranks while ranking authors and journals respectively.

3.2.3 Proposed Topic-based Heterogeneous Rank

We consider a heterogeneous network made up of three underlying networks sharing nodes. The proposed TH Rank operates on the whole network, passing information back and forth in an iterative manner between the three subnetworks. The whole process combines the results from one intra-class walk within the paper citation network and two inter-class walks, one between the article and author bipartite graph, and the other between the article and journal bipartite graph. This continues in an iterative manner until convergence. We represent the vector of paper ranks by R^P, the vector of author ranks by R^A, and vector of journal ranks by R^J. The vectors R^P, R^A and R^J are initialized from the corresponding topic vectors from *NPT*, *NAT* and *NJT*, respectively, instead of using the common initialization technique. Calculation of rank in inter-class walks is based on one step of HITS (Kleinberg, 1999). The score of authors is computed by the following formula:

$$R^A = M^{A^T} * R^P \qquad (4)$$

where M^{A^T} is the transpose of paper-author matrix. Formula 4 represents that papers transfer their authority scores to the authors, and authors receive authority scores from all their publications. Similarly, the rank of journals is also calculated by one step of an inter-class walk for paper-journal network. The score of journals is computed by the following formula:

$$R^J = M^{J^T} * R^P \qquad (5)$$

where M^{J^T} is the transpose of paper-journal matrix. We describe the calculation of the ranking of papers R^P as follows:

$$R^P = (\alpha * M^P * R^P) + (\beta * M^A * R^A) + (\gamma * M^J * R^J) + (1 - \alpha - \beta - \gamma) * 1/n \qquad (6)$$

The intra-class and inter-class walks are coupled with α, β, and ϒ which are used to represent the mutual dependence of papers, authors, and journals.
The steps followed by the proposed method are as follows:

1. Assigning topic probability distributions: we use the ACT model to assign topic probability distributions to papers, authors, and venues, rather than the simple 1/n procedure followed by previous methods. This is broken down into the following discrete steps:
    a. The ACT model calculates the probability of a paper for a given topic, the probability of an author for a given topic, and the probability of a conference for a given topic.
    b. Gibbs sampling is used for inference, and the hyper-parameters π, δ, and μ are set at fixed values (π=5.00, δ=0.10, and μ=0.1).
    c. Topic probability distributions for papers are obtained from the file model-final10.theta as column vectors, one column of files representing the topic probability for one topic. These files are normalized and stored in vector R^P for a given topic.
    d. Similarly, the topic probability distributions for authors and venues are attained from files model-final10.theta_ak and model-final10.theta_ck respectively and stored in vectors R^A and R^J.
2. The personalized vector R^P of equation (6) is computed based on the R^P, R^A and R^J scores attained from step 1. In first iteration its value is taken from step 1; in all subsequent iterations the updated values of these vectors from step 3 and 4 are used.
3. Authors attain their scores via the paper-author adjacency matrix and the topic probability distribution of each paper in given topic, as follows: $R^A = M^{A^T} * R^P$ where M^{A^T} is the transpose of paper-author adjacency matrix and R^P is the topic probability of each paper in the given topic. In the first iteration its value is taken from step 1, in all subsequent iterations its value is retrieved from step 2.
4. Journals attain their scores via the paper-journal adjacency matrix and the topic probability distribution of papers in a given topic as follows: $R^J = M^{J^T} * R^P$, where M^{J^T} is the transpose of the paper journal adjacency matrix and RP is the topic probability of each paper in a given topic. In the first iteration its value is taken from step 1, in all subsequent iterations its value is retrieved from step 2.
5. Vectors R^P, R^A and R^J are iteratively calculated until convergence.

4. Results and Discussion

4.1 Values for parameters

There are three parameters that can be used to manipulate the results of the TH Rank algorithm: α, β and ϒ. Here α, β, and ϒ correspond to the paper citation network, author bipartite network and journal bipartite network respectively. In equation 6, we have adjusted the values of α, β and ϒ in such a way that $(1 - \alpha - \beta - \gamma)$ corresponds to the damping factor of the original PageRank algorithm. Manipulating these values gives us variants of the proposed TH Rank method. For β=0 we can achieve a variant of TH Rank in which ranking will be done only for the citation network and the journal network as shown in equation 7.

$$R^P = (\alpha * M^P * R^P) + (\gamma * M^J * R^J) + (1 - \alpha - \gamma) * 1/n$$

$$R^J = M^{J^T} * R^P \qquad (7)$$

Similarly, ϒ=0 results in another variant of TH Rank that ranks only the citation network and authors network as shown in equation 8.

$$R^P = (\alpha * M^P * R^P) + (\beta * M^A * R^A) + (1 - \alpha - \beta) * 1/n$$

$$R^A = M^{A^T} * R^P \quad (8)$$

For both β = ϒ = 0, TH Rank will simply become the Topic-based PageRank model for citation networks only as shown in equation (9).

$$R^P = (\alpha * M^P * R^P) + (1 - \alpha) * 1/n \quad (9)$$

To conduct the experiments in this study we have adjusted the values of α, β and ϒ in such a way that $(\alpha + \beta + \gamma) = 0.85$ so that $(1 - \alpha - \beta - \gamma)$ corresponds to the damping factor values of the original PageRank method.

4.2 Topic

Ten topics were extracted from dataset using the ACT model, out of which three were selected for experimentation. Table 3 shows the selected topics and the words associated with each topic. We have also assigned labels to these topics. Table 4 shows the words associated with the remaining 7 topics that were not selected. The ACT model was used to calculate the probability distribution of papers, authors and journals simultaneously. The probability of a paper for a given topic P(p|t) produces the most related papers to the given topic. Similarly, the probability of an author for a given topic P(a|t) identifies the most productive authors for the given topic, and the probability of a journal for a given topic P(j|t) gives us the most productive journals with respect to this given topic.

Table 3. Top 10 words associated with selected topics

Topics	**Associated Words**
Multimedia IR	Image, texture, data, content, color, visual, video, shape, learning, feature,
Medical IR	Data, medical, biomedical, care, patient, database, clinical, system, health, analysis
Database & Query Processing	Query, processing, data, relational, language, xml, object, spatial, database, search,

Table 4: Words associated with remaining 7 topics

Associated Words
Language, document, search, query, system, relevance, text, evaluation, analysis, searching
Query, data, database, language, relational, object, system, spatial, network, temporal
Search, web, semantic, library, online, analysis, system, digital, research, knowledge
Web, semantic, image, document, knowledge, data, ontology, design, framework, management
System, data, storage, analysis, computer, automated, protein, chemical, document
Data, image, music, classification, analysis, recognition, learning, content, neural, algorithm
Document, fuzzy, web, semantic, image, approach, knowledge, memory

4.3 Papers

In the current heterogeneous network, papers are the central entities, while authors and journals are connected to papers via bipartite networks. Authors' ranks are dependent on the papers they have presented, and similarly, journal ranks are dependent on the papers they have published.

Papers are therefore the most important entity of network. Topic probability distributions from the ACT model can reveal most papers related to a given topic. Table 5 shows the top 10 papers retrieved by TH Rank for selected topics, along with results of a general citation network which does not consider topic differences ordered by their Without-Topic Ranking (WTR). We can see in the TH ranking that the paper "*Query by image and video content - the qbic system*" placed first in the Multimedia IR field. It was also ranked first in the WTR due to the effect of network topology and citations. In the TH ranking, "*Similar-shape retrieval in shape data management*", "*Unifying keywords and visual contents in image retrieval*", "*Semantics in visual information retrieval*" and "*Chabot - retrieval from a relational database of images*" have moved up 1, 2, 3 and 4 positions, respectively, as compared to their respective positions at 3, 5, 7, and 9 in WTR. The paper "*From pixels to semantic spaces: advances in content-based image retrieval*" was ranked 15th in the WTR but moved to position 7 in the multimedia IR TH ranking. The paper "*Developments in automatic text retrieval*" dropped to position 8 from position 2 in the WTR. These prominent papers in field of Multimedia IR were also important nodes in the overall network, and they consequently appear in the top ranks of the WTR. The most prominent changes were with the papers "*Image retrieval using nonlinear manifold embedding*" and "*Scale invariant image matching using triplewise constraint and weighted voting*," which entered the top ten of the TH ranking from positions 68 and 86, respectively, in the WTR.

Unlike Multimedia IR, prominent nodes in Medical IR field as identified by TH ranking were not given top positions in WTR. One exception is the paper "*Developments in automatic text retrieval*," which, due to its important position in citation network, occupies the top position in medical IR TH ranking and second position in WTR. We can see from Table 5 that papers "*Gene clustering by latent semantic indexing of medline abstracts*" and "*Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families*" are ranked 2 and 3 in the medical IR ranking while they were ranked 19 and 20, respectively, in the WTR. The papers "*Informatics in radiology (inforad) - benefits of content-based visual data access in radiology*", "*Ebimed - text crunching to gather facts for proteins from medline*", "*Automated retrieval of ct images of liver lesions on the basis of image similarity: method and preliminary results*", "*Genecards: a novel functional genomics compendium with automated data mining and query reformulation support*", "*Resolving abbreviations to their senses in medline*" and "*Biowarehouse: a bioinformatics database warehouse toolkit*" occupy positions 5, 6, 7, 8, 9 and 10 in the medical IR TH ranking, while they are ranked 12, 28, 22, 38, 35 and 52, respectively, in the WTR. Due to effect of TH ranking's topic-based methods, these papers appear in top 10 of topic-specific rankings. We can clearly see that TH Rank has successfully altered the results of the WTR to bring relevant papers from a certain topic into view.

From Table 3 we can see words associated with the Database & Query processing topic. Table 5 shows the papers associated with these words. "*Comparative analysis of five xml query languages*," which occupies top position in Database & Query processing was ranked 6 in the WTR. "*Query by image and video content - the qbic system*" has moved one position down because the ACT model has assigned it slightly less probability in database & query processing when compared to the multimedia IR topic. The paper "*On supporting containment queries in relational database management systems*" has gained position 3, whereas in the WTR it ranked 11th. The

papers "*A query language for biological networks*", "*Reasoning on regular path queries*", "*Multi-dimensional scattered ranking methods for geographic information retrieval*", "*Efficient implementation techniques for topological predicates on complex spatial objects*", "*Picture query languages for pictorial data-base systems*" and "*Algorithms for nearest neighbor search on moving object trajectories*" have moved to positions 4,5, 7, 8, 9 and 10 respectively from positions 32, 57, 56, 79, 85 and 75 in the WTR. We can see that topic-based ranking allows the topic specific papers to get visibility and they emerge from lower positions when topic is accounted for. This visibility is not possible in a general without-topic ranking.

For purposes of comparison and evaluation we have included the topic-based method proposed by Ding (2011b) as another baseline. This method involves the topic-based ranking of authors in a co-citation network. Methods for author co-citation networks can also be applied on a paper citation networks, and so we applied this method as a baseline and compared the results of the paper ranking with the TH Rank method with results attained by method of Ding (2011b). Table 6 shows the topic-based results for ranking of papers using Ding's method as a baseline. From a cross-comparison of Tables 5 and 6 we noticed that papers ranked by TH Rank either have authors prominent in the given subfield or are published in an important journal. Consider the paper "*Gene clustering by latent semantic indexing of medline abstracts*," which was ranked highly by the TH Rank method in the medical IR topic. This paper was published in "*Bioinformatics*," a top journal in medical IR and Kirsch H, one of its authors, is also a renowned medical IR researcher. Similarly, the article "*Similar-shape retrieval in shape data management*" is published in the top journal "*Computer*" and some of its authors, such as Mehrotra R and Gary JE, are well-known authors in multimedia IR. Likewise, the paper "*Comparative analysis of five xml query languages*" was published in a top journal in the Database & Query processing topic *"Sigmod record."* This paper was authored by Bonifati A and Ceri S who are present among the top 10 authors of the field. Such patterns are not identified for the papers ranked by the baseline method. The baseline method can identify topic-based papers but cannot simultaneously rank authors and journals

Paper citation networks play a very important role for the ranking of related entities such as authors and journals. The importance of a research article can be conceptualized as a count of peer votes, which in this case are the citations received by other research papers. A heterogeneous network with a paper citation network acting as its hub greatly enriches the amount of extractable information as it creates links between similar authors and similar journals. In an author-paper bipartite network authors are only weakly linked via the papers they coauthor; the case is the same in journal-paper bipartite networks. Ranking a heterogeneous paper-author-journal network at the topic level increases the level of granularity and enhances the ranking of allied bodies (i.e., the authors and journals). The currently proposed method is capable of generating separate rank vectors for a given paper in all selected topics, giving each paper its due visibility according to its topic probability in that particular topic.

Table 5. Top 10 papers ranked by WTR and TH Rank for selected topics

Rank	**WithoutTopic Rank (WTR)**	**Multimedia IR**	**Medical IR**	**Database & Query Processing**
1	Query by image and video content - the qbic system	Query by image and video content - the qbic system	Developments in automatic text retrieval	Comparative analysis of five xml query languages

2	Developments in automatic text retrieval	Similar-shape retrieval in shape data management	Gene clustering by latent semantic indexing of medline abstracts	Query by image and video content - the qbic system
3	Similar shape retrieval in shape data management	Unifying keywords and visual contents in image retrieval	Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families	On supporting containment queries in relational database management systems
4	Information retrieval systems	Semantics in visual information retrieval	Information retrieval in digital libraries: bringing search to the net	A query language for biological networks
5	Unifying keywords and visual contents in image retrieval	Chabot - retrieval from a relational database of images	Informatics in radiology (inforad) - benefits of content-based visual data access in radiology	Reasoning on regular path queries
6	Comparative analysis of five xml query languages	Automatic-indexing and content-based retrieval of captioned images	Ebimed - text crunching to gather facts for proteins from medline	Chabot - retrieval from a relational database of images
7	Semantics in visual information retrieval	From pixels to semantic spaces: advances in content-based image retrieval	Automated retrieval of ct images of liver lesions on the basis of image similarity: method and preliminary results	Multi-dimensional scattered ranking methods for geographic information retrieval
8	Automatic-indexing and content-based retrieval of captioned images	Developments in automatic text retrieval	Genecards: a novel functional genomics compendium with automated data mining and query reformulation support	Efficient implementation techniques for topological predicates on complex spatial objects
9	Chabot - retrieval from a relational database of images	Image retrieval using nonlinear manifold embedding	Resolving abbreviations to their senses in medline	Picture query languages for pictorial data-base systems
10	Websom - self-organizing maps of document collections	Scale invariant image matching using triplewise constraint and weighted voting	Biowarehouse: a bioinformatics database warehouse toolkit	Algorithms for nearest neighbor search on moving object trajectories

Table 6: Top 10 papers ranked by Baseline method for selected topics

	Multimedia IR	**Medical IR**	**Database & Query Processing**
1	Retrieval of images of man-made structures based on projective invariance	Are oral clefts a consequence of maternal hormone imbalance? evidence from the sex ratios of sibs of probands	A pictorial query language for querying geographic databases using positional and olap operators
2	Scale invariant image matching using triplewise constraint and weighted voting	Francisella tularensis novicida proteomic and transcriptomic data integration and annotation based on semantic web technologies	Incorporating language processing into java applications: a javacc tutorial
3	An image retrieval system by impression words and specific object names - iris	Refined repetitive sequence searches utilizing a fast hash function and cross species information retrievals	A filter flow visual querying language and interface for spatial databases
4	Using relevance feedback with short-term memory for content-based spine x-ray image retrieval	An xml transfer schema for exchange of genomic and genetic mapping data: implementation as a web service in a taverna workflow	A data model and data structures for moving objects databases
5	Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization	Semi-automated curation of protein subcellular localization: a text mining-based approach to gene ontology (go) cellular component curation	Efficient implementation techniques for topological predicates on complex spatial objects
6	Image feature descriptor based on shape salience points	Kipar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways	Caching and incrementalisation in the java query language
7	Semi-supervised spectral hashing for fast similarity search	Kidney transplantation search filters for pubmed, ovid medline, and embase	Xquery formal semantics state and challenges
8	Kernel-based metric learning for semi-supervised clustering	Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span	Spatio-temporal data handling with constraints
9	Investigating the behavior of compact composite descriptors in early fusion, late fusion and distributed image retrieval	Biomart and bioconductor: a powerful link between biological databases and microarray data analysis	Investigating xquery for querying across database object types
10	Feature integration analysis of bag-of-features model for image retrieval	Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry	Unifying temporal data models via a conceptual-model

4.4 Authors

Author rank is calculated by an inter-class walk between the paper-citation network and the bipartite author network. The topic sensitivity built into this method allows authors related to a topic to appear as top authors within a given topic. Table 7 shows top 10 authors ranked by TH Rank in selected topics (Multimedia IR, Medical IR and Database & Query Processing). Salton G occupies the first position in both the Multimedia IR TH ranking and the WTR. In multimedia IR Huang TS has risen to the 2nd position from the 9th in the WTR listing. Huang TS has a very strong profile showing special interest in multimedia IR. The ranking of Del Bimbo A as 3rd in the TH ranking indicates his specific research interests in Multimedia IR and Image Processing. He was ranked 26th by WTR. Pala P, who was ranked 27th in the WTR, has moved to position 4 in the TH ranking due to his research on pattern recognition and models for multimedia information retrieval. The research interests of Zhou XS include multimedia information retrieval and image processing and his rank is consequently improved by 3 places from the WTR to the TH ranking. Srihari RK belongs to multimedia databases and information retrieval research groups. Her rank has moved one position down as compared to her rank in WTR, as more authoritative authors in Multimedia IR have taken her place. For similar reasons, Mehrotra R and Gary JE have moved 5 and 6 positions down, respectively, in the Multimedia IR TH ranking when compared to their WTR positions. Vasconcelos N has gained 2 positions due to his interest in multimedia, computer vision, image processing and machine learning. The most interesting name in top-10 list is Ip HHS, who occupied the 339th position on the WTR, entered in top 10 in TH ranking due to his interests in multimedia retrieval, machine learning, pattern recognition, and computer graphics.

Rebholz-Schuhmann D was first in the Medical IR TH ranking, while in the WTR his rank was 31. His academic profile indicates strong research interest in biomedical informatics. He also heads a research group which processes biomedical data resources and biomedical scientific literature. Kangarloo H, Etzold T, Gaudan S, Kirsch H, Huang Y, Shah SP, Xu T and Ouellette BFF took positions 43, 85, 91, 30, 115, 116, 117 and 119 respectively in WTR, and now are ranked 2, 3, 4, 5, 7, 8, 9 and 10 within Medical IR ranking due to their specific research interests at the intersection of bioinformatics and information retrieval.

In the WTR Ceri S and Bonaifati A were ranked 11th and 12th respectively. Due to effect of topic-based ranking in a heterogeneous network, they managed to supersede Salton G in the Database and Query Processing topic. Similarly Clementini E, Leser U, Yao Y, Gehrke J, Calvanese D, De Giacomo G and Lenzerini M rose to positions 4, 5, 6, 7, 8, 9 from positions 95, 556, 35, 36, 224, 225 and 226. Only Salton G, who has achieved great overall prominence in the entire field of IR thanks to his well-known contributions to the field, remains in a top-10 position in the transition from the WTR to the Database and Query Processing topic ranking. Leser U provides an example of a researcher who is recognized as a top researcher in this topic but is missed by the WTR ranking.

We extracted an author citation network from our data and applied Ding's method (Ding 2011b) to serve as a baseline for comparison with the proposed method. Table 8 shows the results retrieved by the baseline method for the selected topics. Ding's method uses the topic distribution of authors to topically rank and is therefore capable of finding subfield experts. Comparing Tables 7 and 8,

we see that in Multimedia IR topic both the proposed and baseline methods both ranked Salton G as the top researcher. This is a reasonable result since Salton G has had a long history of producing very influential and highly cited work in IR. The TH Rank ranked Huang TS as 2nd in contrast to a ranking of 7th on the baseline ranking. TH Rank identified Del Bimbo A, Pala P and Ip HHS as top researchers, while the baseline method did not. These researchers published papers that are highly related to Multimedia IR and were published in core journals for this field. In the case of Medical IR, TH Rank identifies Rebholz-Schuhmann D, Kangarloo H, Etzold T, Huang Y, Shah SP, Xu T and Ouellette BFF as top researchers, while the baseline does not. Profiles of these authors demonstrate that these researchers are highly active in the Medical IR field and have published several articles in top Medical IR journals. Like in the WTR ranking, the baseline method failed to acknowledge the topic-specific contributions of Ceri S and Bonafati A in the Database & Query Processing topic. Overall we conclude that the simultaneous ranking of authors, papers, and journals is capable of identifying nuances missed by the baseline method. Specifically, it acknowledges that while there are individuals such as Salton who have achieved field-wide prominence and have high overall citation counts from prestigious authors and journals, such prominence should not necessarily eclipse the different levels of contributions certain authors have contributed to certain subfields.

Table 7. Top 10 authors ranked by WTR and TH Rank for the selected topics

Rank	Without-Topic Rank (WTR)	Multimedia IR	Medical IR	Database & Query Processing
1	Salton, G	Salton, G	Rebholz-Schuhmann, D	Ceri, S
2	Mehrotra, R	Huang, TS	Kangarloo, H	Bonifati, A
3	Gary, JE	Del Bimbo, A	Etzold, T	Salton, G
4	Swets, JA	Pala, P	Gaudan, S	Clementini, E
5	Srihari, RK	Zhou, XS	Kirsch, H	Leser, U
6	Schatz, BR	Srihari, RK	Salton, G	Yao, Y
7	Ogle, VE	Mehrotra, R	Huang, Y	Gehrke, J
8	Zhou, XS	Vasconcelos, N	Shah, SP	Calvanese, D
9	Huang, TS	Gary, JE	Xu, T	De Giacomo, G
10	Vasconcelos, N	Ip, HHS	Ouellette, BFF	Lenzerini, M

Table 8: Top 10 authors ranked by Baseline method for the selected topics

	Multimedia IR	Medical IR	Database & Query Processing
1	Salton, G	Zhou, XS	Salton, G
2	Mehrotra, R	Huang, TS	GARY, JE
3	GARY, JE	Vasconcelos, Nuno	Mehrotra, R
4	SWETS, JA	Bonifati, A	SWETS, JA
5	Srihari, RK	Ceri, S	Srihari, RK
6	Schatz, BR	Chen, HC	Schatz, BR
7	Huang, TS	Pala, P	OGLE, VE
8	OGLE, VE	Del Bimbo, A	Zhou, XS
9	Zhou, XS	Gaudan, S	Huang, TS
10	Vasconcelos, Nuno	Kirsch, H	Vasconcelos, Nuno

Figures 2, 3 and 4 show the effects of TH Rank on the relative positioning of authors in comparison to topic-insensitive ranking. From here we can see that TH Rank can successfully and effectively identify prominent authors within a given field that would otherwise not show up in top ranks of a topic-insensitive ranking.

When scrutinizing the topology of an academic field, ranking of authors is perhaps the most relevant metric and topic-sensitivity is vital in separating overall prominence from topic-specific prominence. This should be an important consideration in institutional processes that evaluate and assess academic performance or distributes promotion or tenure. The proposed method is capable of generating topic-based vectors for each author, giving his or her relevant position in each topic.

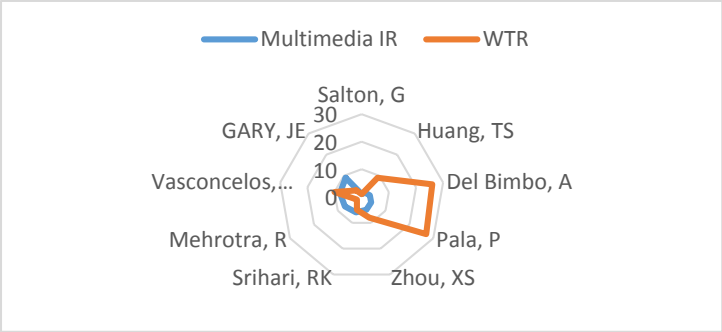

Figure 2: Relative positioning of authors in Multimedia IR with respect to WTR

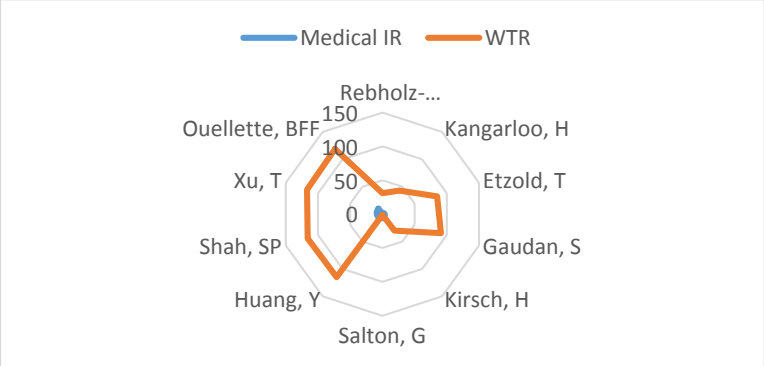

Figure 3: Relative positioning of authors in Medical IR with respect to WTR

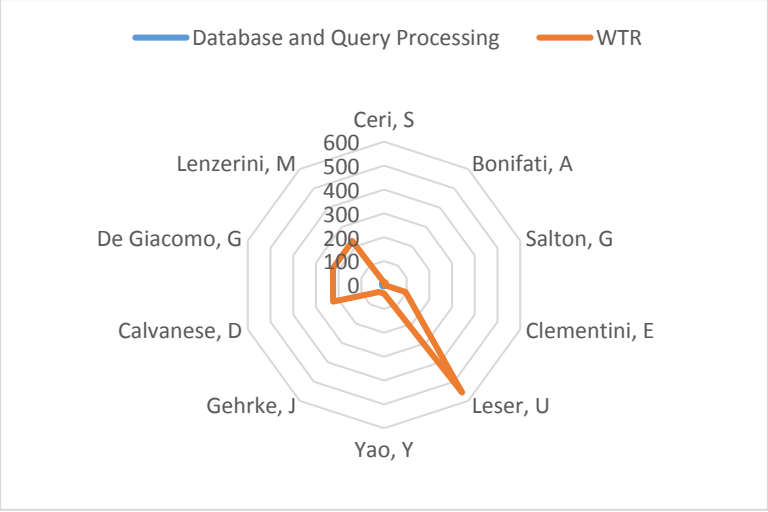

Figure 4: Relative positioning of authors in the Database & Query processing topic with respect to WTR

4.5 Journals

As with author rank, journal rank is calculated by an inter-class walk between the paper-citation network and the bipartite journal network. Topic sensitivity in ranking of journals allows topical journals to appear in rankings within that topic. Table 9 shows the top 10 journals in the WTR and the topic-sensitive Multimedia IR, Medical IR and Database & Query processing rankings. The journal identified by the abbreviation *Computer* occupied first place in the WTR and multimedia IR rankings, while in the Medical IR rankings it is replaced by the journal *Bioinformatics*. The *Bioinformatics* journal occupied the 3rd position in the WTR, but did not appear in the top 10 of the Multimedia IR ranking. *IEEE Multimedia* was promoted from the 6th position in WTR to the 2nd position in Multimedia IR ranking, while it does not appear at all in the top 10 spots of the Medical IR ranking. The journal *Neurocomputing* gained two levels in the Multimedia IR ranking when compared to its WTR rank, and does not appear in the top 10 of the Medical IR ranking. The *Displays* journal, absent in the WTR, attains the 4th position in the multimedia IR ranking. *Science* magazine, which covers a wide range of scientific topic, occupie 2nd place in the topic-insensitive ranking but drops to 6th place in all topic-sensitive rankings. Similar patterns are observed for the topic Database & Query Processing for the ranking of journals. *Geoinformatica*, *ACM sigplan notices*, *Algorithmica*, *Bit* and *Neural networks* are absent in the WTR but appear in the topic-sensitive listing.

Topic-based ranking of journals brings forward journals that have achieved prominence within a particular topic. Considering the topic of an article as well as the topic of citing journals is important for calculating the overall citation impact of a given article. Impact factor, while a useful metric, cannot on its own assist researchers investigating quality and impact within a particular topic or subfield.

Table 9. Top 10 journals ranked by WTR and TH Rank for selected topics

Rank	Without-Topic Rank (WTR)	Multimedia IR	Medical IR	Database & Query Processing
1	Computer	Computer	Bioinformatics	Sigmod record
2	Science	IEEE multimedia	BMC bioinformatics	Bioinformatics
3	Bioinformatics	Neurocomputing	Computer	Geoinformatica
4	Sigmod record	Displays	Radiology	Computer
5	Neurocomputing	Neural networks	Drug safety	ACM sigplan notices
6	IEEE multimedia	Science	Science	Science
7	BMC bioinformatics	Journal of electronic imaging	Radiographics	Algorithmica
8	Radiographics	Information systems	Neuroinformatics	IEEE multimedia
9	Neural networks	Radioengineering	Oncogene	Bit
10	Libri	Pattern recognition	Radiologe	Neural networks

5. Conclusion and Future Work

We contend that the ranking of academic entitie like papers, authors or journals with respect to their topic is an important method in light of the growing interest in the role of context in citation research. Most widely-used ranking algorithms, however, do not simultaneously consider the potential influences of all three of these academic entities. In this study, we proposed a Topic-based Heterogeneous Rank algorithm which is capable of combining information about citations, authors and journals to effectively rank academic entities in a heterogeneous environment with respect to their topics. The ACT model is used to extract topics and to associate these topics with

their respective papers, authors and journals, uniting all three entity types in a heterogeneous network. Unlike in existing methods, which initialize entities with equal values, we used probabilities from the ACT model as weighted vectors for use in the PageRank algorithm. The study constructs a heterogeneous scholarly network in which there is one intra-class walk and two inter-class walks. For the intra-class walk, papers interact with other papers via citation links. For inter-class walks, authors interact with papers via the paper-author adjacency matrix, and journals interact with papers via the paper-journal adjacency matrix.

The results show that TH Rank can effectively find the most relevant papers, authors and journals for a given field when compared to general ranking methods that are insensitive to topic. We have selected three topics, Multimedia IR, Medical IR and Database & Query Processing to demonstrate the ranking results of TH Rank in this paper. We observe that TH Rank gives the objects their due attention with respect to their topics. In the future, we will further evaluate the proposed approach on all topics.

One of the main limitations of the proposed method is the computational complexity and high memory usage of the proposed algorithm. We plan on investigating ways to improve the efficiency of the algorithm in future research. Furthermore, the proposed method is insensitive to time of publication. We hope to improve the application of this method vis-à-vis large dataset covering long periods of time by making it possible to compute and compare the dynamic rankings of authors, papers, and journals over the passage time. We consider time of publication and time of citation to be integral factors in calculating impact, since older papers have more time to get noticed while newer papers, irrespective of their quality or innovation, usually do not accumulate citations until reaching a certain threshold of visibility.

References:

Aksnes, D.W., 2003. A macro study of self-citation. Scientometrics 56 (2), 235–246.

Bergstrom, C.T., West, J.D., Wiseman, M.A., 2008. The Eigenfactor™ metrics. J. Neurosci. 28 (45), 11433–11434.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bollen, J., Rodriquez, M.A., Van de Sompel, H., 2006. Journal status. Scientometrics 69 (3), 669–687.

Chen, P., Xie, H., Maslov, S., Redner, S., 2007. Finding scientific gems with Google's PageRank algorithm. J. Informetr. 1 (1), 8–15.

Cronin, B., 1984. The citation process. The role and significance of citations in scientific communication. Lond. Taylor Graham 1984 1.

Davis, P.M., 2008. Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? J. Am. Soc. Inf. Sci. Technol. 59 (13), 2186–2188.

Ding, Y., 2011a. Applying weighted PageRank to author citation networks. J. Am. Soc. Inf. Sci. Technol. 62 (2), 236–245.

Ding, Y., 2011b. Topic-based PageRank on author cocitation networks. J. Am. Soc. Inf. Sci. Technol. 62 (3), 449–466.

Ding, Y., Cronin, B., 2011. Popular and/or prestigious? Measures of scholarly esteem. Inf. Process. Manag. 47 (1), 80–96.

Ding, Y., Yan, E., Frazho, A., Caverlee, J., 2009. PageRank for ranking authors in co-citation networks. J. Am. Soc. Inf. Sci. Technol. 60 (11), 2229–2243.

Ding, Y., Rousseau, R., & Wolfram, D. (Eds.) (2014). Measuring scholarly impact: Methods and practice. Springer.

Fiala, D., Rousselot, F., Ježek, K., 2008. PageRank for bibliographic networks. Scientometrics 76 (1), 135–158.
Frandsen, T.F., 2007. Journal self-citations—Analysing the JIF mechanism. J. Informetr. 1 (1), 47–58.
Garfield, E., 1999. Journal impact factor: a brief review. Can. Med. Assoc. J. 161 (8), 979–980.
Glänzel, W., Thijs, B., 2004. The influence of author self-citations on bibliometric macro indicators. Scientometrics 59 (3), 281–310.
Haveliwala, T.H., 2002. Topic-sensitive pagerank, in: Proceedings of the 11th International Conference on World Wide Web. ACM, pp. 517–526.
Hendrix, D., 2009. Institutional self-citation rates: A three year study of universities in the United States. Scientometrics 81 (2), 321–331.
Hyland, K., 2003. Self-citation and self-reference: Credibility and promotion in academic publication. J. Am. Soc. Inf. Sci. Technol. 54 (3), 251–259.
Katsaros, D., Akritidis, L., Bozanis, P., 2009. The f index: Quantifying the impact of coterminal citations on scientists' ranking. J. Am. Soc. Inf. Sci. Technol. 60 (5), 1051–1056.
Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. J. ACM JACM 46 (5), 604–632.
Krauss, J., 2007. Journal self-citation rates in ecological sciences. Scientometrics 73 (1), 79–89.
Leydesdorff, L., 2007. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. J. Am. Soc. Inf. Sci. Technol. 58 (9), 1303–1319.
Leydesdorff, L., 2009. How are new citation-based journal indicators adding to the bibliometric toolbox? J. Am. Soc. Inf. Sci. Technol. 60 (7), 1327–1336.
Liu, J.-G., Xuan, Z.-G., Dang, Y.-Z., Guo, Q., Wang, Z.-T., 2007. Weighted network properties of Chinese nature science basic research. Phys. Stat. Mech. Its Appl. 377 (1), 302–314.
Liu, X., Bollen, J., Nelson, M.L., Van de Sompel, H., 2005. Co-authorship networks in the digital library research community. Inf. Process. Manag. 41 (6), 1462–1480.
Li, Y., Tang, J., 2008. Expertise search in a time-varying social network, in: Web-Age Information Management, 2008. WAIM'08. The Ninth International Conference on. IEEE, pp. 293–300.
Ma, N., Guan, J., Zhao, Y., 2008. Bringing PageRank to the citation analysis. Inf. Process. Manag. 44 (2), 800–810.
Maslov, S., Redner, S., 2008. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. J. Neurosci. 28 (44), 11103–11105.
Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank citation ranking: Bringing order to the web.
Pal, S.K., Narayan, B.L., 2005. A web surfer model incorporating topic continuity. Knowl. Data Eng. IEEE Trans. On 17 (5), 726–729.
Pinski, G., Narin, F., 1976. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Inf. Process. Manag. 12 (5), 297–312.
Radicchi, F., Fortunato, S., Markines, B., Vespignani, A., 2009. Diffusion of scientific credits and the ranking of scientists. Phys. Rev. E 80 (5), 056103.
Richardson, M., Domingos, P., 2001. The Intelligent surfer: Probabilistic Combination of Link and Content Information in PageRank., in: NIPS. pp. 1441–1448.
Sayyadi, H., Getoor, L., 2009. FutureRank: Ranking Scientific Articles by Predicting their Future PageRank., in: SDM. SIAM, pp. 533–544.
Sun, Y., Han, J., 2013. Meta-path-based search and mining in heterogeneous information networks. IEEE Tsinghua Science and Technology, 18(4), 329-338.
Tang, J., Jin, R., Zhang, J., 2008. A topic modeling approach and its integration into the random walk framework for academic search, in: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, pp. 1055–1060.
Tsay, M., 2006. Journal self-citation study for semiconductor literature: synchronous and diachronous approach. Inf. Process. Manag. 42 (6), 1567–1577.

Van Raan, A.F., 2008. Self-citation as an impact-reinforcing mechanism in the science system. J. Am. Soc. Inf. Sci. Technol. 59 (10), 1631–1643.
West, J.D., Bergstrom, T.C., Bergstrom, C.T., 2010. The Eigenfactor MetricsTM: A network approach to assessing scholarly journals. Coll. Res. Libr. 71 (3), 236–244.
Yan, E., Ding, Y., 2009. Applying centrality measures to impact analysis: A coauthorship network analysis. J. Am. Soc. Inf. Sci. Technol. 60 (10), 2107–2118.
Yan, E., Ding, Y., 2010. Weighted citation: An indicator of an article's prestige. J. Am. Soc. Inf. Sci. Technol. 61 (8), 1635–1643.
Yan, E., Ding, Y., 2011. Discovering author impact: A PageRank perspective. Inf. Process. Manag. 47 (1), 125–134.
Yan, E., Ding, Y., Sugimoto, C.R., 2011. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. J. Am. Soc. Inf. Sci. Technol. 62 (3), 467–477.
Yang, Z., Tang, J., Zhang, J., Li, J., Gao, B., 2009. Topic-level random walk through probabilistic model, in: Advances in Data and Web Management. Springer, pp. 162–173.
Zhou, D., Orshanskiy, S.A., Zha, H., Giles, C.L., 2007. Co-ranking authors and documents in a heterogeneous network, in: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, pp. 739–744.