A Query-oriented Approach for Relevance in Citation Networks*

Luam C. Totti[1], Prasenjit Mitra[1], Mourad Ouzzani[1], Mohammed J. Zaki[2]
[1]Qatar Computing Research Institute (QCRI), Doha, Qatar
[2]Rensselaer Polytechnic Institute (RPI), Troy, NY, USA
luamct@gmail.com, {pmitra,mouzzani}@qf.org.qa, zaki@cs.rpi.edu

ABSTRACT

Finding a relevant set of publications for a given topic of interest is a challenging problem. We propose a two-stage query-dependent approach for retrieving relevant papers given a keyword-based query. In the first stage, we utilize content similarity to select an initial seed set of publications; we then augment them by citation links weighted with information such as citation context relevance and age-based attenuation. In the second stage, we construct a multi-layer graph that expands the publications subgraph by including links to the authors, venues, and keywords. This allows us to return recommendations that are both highly authoritative, and also textually related to the query. We show that our staged approach gives superior results on three different benchmark query sets.

1. INTRODUCTION

Researchers are constantly faced with the task of gathering comprehensive and up to date lists of publications relevant to their research. Despite the importance of such mapping task and how frequently it is performed, relatively few tools are available to effectively help scientists. While existing search engines provide reasonable search results, oftentimes these results are neither adequate nor satisfactory. In fact, the scholar has to use other strategies to find relevant publications such as rephrasing the query and searching using experts in the field, top venues, and so on. Improving search results will inevitably reduce the time scholars spend on identifying documents to read, thereby increasing their productivity. The speed of research, and the volume and availability of publications make this task even more challenging.

Documents can be modeled not only using content, e.g., words and topics, but, also relationships such as citations, authors, and venues. In this work, we show how we can utilize this network of additional information. We propose a two-stage query-dependent approach for computing relevance in citation networks that combines both the textual content and the connectivity information between the various entities. Given a user query q comprising a phrase or a set of keywords, and a bibliographic dataset D comprising publications, citations, and metadata on each publication, our approach returns a ranked list of publications.

Our approach, called IQRA[1] (anagram of **Q**uery-based **R**elevance **I**n BibliogrA**p**hic Networks), works as follows: First, we utilize content similarity between the query and the documents to select a seed set of highly-relevant publications. Next, we expand our document set by following the citations from these seed documents to a certain (tunable) depth. While substantial work has been done on two-stage document query processing and query expansion [8], we believe our method is one of the first to expand the document set obtained after the first query stage. We further incorporate as a link weight information such as citation context relevance and age-based attenuation. Consequently, our seed set has a reasonably complete and high-quality candidate set of papers, which can then be reranked based on their citations or other metrics. Second, using this expanded document set, we proceed to find the top entities to recommend. If the objective is to obtain paper recommendations, then our strategy IQRA-TC (for IQRA-TopCited) that ranks the expanded document set based on their citations works the best in our experiments. Note that using top-cited documents on the entire repository does not work as well as using the selected top-cited documents from the expanded document set. We also propose a strategy based on a multilayer, relational graph that expands the publications subgraph by including layers with links to publication-related entities, such as authors, venues, and keywords. Our multilayered method IQRA-ML (for IQRA-MultiLayer) ranks all the nodes (across all layers) via a random-walk based procedure that quantifies each node's relevance while providing control over the contributions of each layer. This random walk mimics a manual literature search (as any scholar would do) to retrieve a core set of papers by finding and using related keywords, following the citations to and from a paper in the seed set, trying to find other related papers by important authors, reading the proceedings of the top venues, and so on.

We perform a comprehensive evaluation comparing our novel two-stage method IQRA with several state-of-the-art methods. With the help of experts, we created a gold-standard benchmark for a chosen set of queries. We also use two other benchmark datasets we created based on ti-

*This work was supported in part by NSF Award IIS-1302231.

[1] IQRA also means "read" in Arabic

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890518 .

tles from regular and survey papers. One of our key findings is that once the relevant query-dependent subgraph has been created, our relatively simple and very efficient Iqra-tc method, which ranks the publications layer based on the number of citations clearly outperforms existing approaches, followed closely by our multi-layer approach Iqra-ml.

2. RELATED WORK

Existing methods for recommending publications can be mapped into three main approaches: content-based [12, 14, 18], graph-based [10, 17, 20], and collaborative filtering methods [9, 18, 19]. Techniques such as PageRank [3] and HITS [7] can also be used to pre-compute the authority scores for publications, which can then be used in conjunction with text similarity to rank documents. See Beel, et al. [2] for a comprehensive survey of around 200 academic papers on research-paper recommendation systems.

Approaches based mainly on textual similarity suffer from the traditional information retrieval issues, such as query ambiguity and difficulty in capturing query semantics [1, 13]. A research topic can often be described by different keywords, or it can be closely related to different topics. These aspects can be hard to model with purely textual methods.

Some techniques incorporate the text around the citations to improve results. Ritche [12] proposed a context enhanced document representation and showed improvements in retrieval performance. He *et al.* [5] presented a context-aware approach to recommend publications to be used in a given citation placeholder. The recent work in [16] identifies important versus non-important citations (and also a finer-grained classification) via a supervised approach. Our approach employs citation contexts differently by boosting relationships between documents according to a context's similarity to the query.

Citation links are used to quantify publication importance and to identify research communities. Walker *et al.* [17] propose a variant of PageRank to account for the fact that publications cannot be updated and therefore only cite older publications. This creates a shift in the relevance flow towards older publications, which is compensated by jump probabilities which decay exponentially based on the age of the papers. We use a similar age decay in our model.

Zhou *et al.* [20] create a low-dimensional embedding of documents by using multiple sources, such as the documents, authors and venues, and the connectivity among them. Ren *et al.* [11] assume that the process of choosing citations for some manuscript varies according to the textual content, authors, and venues; they present a soft clustering approach to account for these behavioral differences. Meng *et al.* [10] incorporate entities such as documents, authors, and topics into a multi-layer graph and measure relevance using a random walk. We improve upon their work along several aspects including network construction, control over the transitions between layers, inclusion of baseline methods for evaluation, and the much larger datasets we employ.

Probabilistic models and topic-based approaches have also been well explored in the context of publications recommendation. Wang and Blei [18] incorporated textual content into the traditional matrix factorization methods by means of a probabilistic model based on topic modeling and content analysis. Tang and Zhang [14] employ topic-based recommendation by learning topic distributions over documents and citations simultaneously. They used ground truth data of relevant documents for each textual context (text around a citation) to train a two-layer Restricted Boltzmann Machine to perform future recommendations, showing performance improvements over a basic language model approach. Our work employs topic modeling in a different manner. By creating a keywords layer connected to the publications we explore both content and structural information derived from the actual keywords found in the documents.

3. IQRA: QUERY-SPECIFIC RECOMMENDATION

Given a query q (keywords or a phrase) and a bibliographic database D (consisting of publications, citations, and metadata), our goal is to return a ranked list of publications.

3.1 IQRA-TC: Paper Recommendation

We incorporate the query q in the first stage of the graph construction by retrieving a small set of publications that are textually similar to the query, and then expanding this set via their citations to create a query-specific subgraph $G_q = (V_q, E_q)$. Formally, given an input query q, the algorithm retrieves a set P_0 of the K most textually similar documents to q according to the cosine similarity between the TF-IDF representations of the query q and each document. Documents are represented by the concatenation of their titles and abstracts after stemming and stop-word removal.

We define the set P_H as the publications that either cite or are cited by a document in P_0 within H hops. The *seed set* of relevant publications is given as $P_s = P_0 \cup P_H$, which makes up the vertex set for the query-specific subgraph G_q, i.e., $V_q = P_s$. The edge set E_q comprises *directed* edges (p_i, p_j, w_{ij}), if publication p_i cites publication p_j, with the link weight w_{ij}. Thus, G_q is a query-dependent subgraph based on both the textual query similarity and the structural properties of the citation network. The link weight is defined as the product

$$w_{ij} = C_{ij} Q_j Y_j$$

where C_{ij} measures the similarity between the citation context and the query, Q_j measures the similarity between p_j and the query, and Y_j represents the age decay.

Citation Context: We enforce query relevance of cited papers by incorporating the *citation context*, which comprises the words in the sentence in which the citation occurs. Let publication p_i cite publication p_j, and let f_{ij} denote the set of words comprising the citation context for p_j in document p_i. We define the text similarity function between the context f_{ij} and the query q as $S(f_{ij}, q) = \cos\big(\text{tf-idf}(f_{ij}), \text{tf-idf}(q)\big)$ which is the TF-IDF based cosine similarity between the arguments. The values for all $S(f_{ij}, q)$ are then normalized by the maximum value $S_{\max}$ to obtain the normalized similarity: $s_{ij} = S(f_{ij}, q)/S_{\max}$. Finally, we define $C_{ij} = \exp\left(\omega(1 - s_{ij})\right)$, where $\omega \geq 0$ controls the importance of the query; for instance ω close to 0 makes C_{ij} close to 1, diminishing the effect of the query similarity values, s_{ij}.

Query Similarity: We define Q_j as the query relevance of p_j, defined analogously to the citation context, as follows: $Q_j = \exp\left(\sigma(1 - b_j)\right)$, where b_j is the normalized cosine similarity of p_j to q using TF-IDF, and $\sigma \geq 0$ controls the relevance between q and p_j.

Age Decay: Citation networks are susceptible to aging effects since a publication typically does not cite another publication that is published later in time. We define the age attenuation factor Y_j for publication p_j as follow: $Y_j = \exp(-\gamma(y_c - y_j))$, where y_j is the publication year of p_j, y_c is the current year and $\gamma \geq 0$ is a parameter to control the age decay.

IQRA-TC Algorithm: Given the query-specific subgraph G_q, IQRA-TC simply ranks each node by the number of times it is cited within G_q, with the top-cited k papers constituting the query result.

3.2 IQRA-ML: Multi-Layer Recommendation

For holistic recommendation using all entities, in the second stage, we incorporate the authors, venues and keywords layers, to yield a multi-layer subgraph G_q, as shown in Fig. 1.

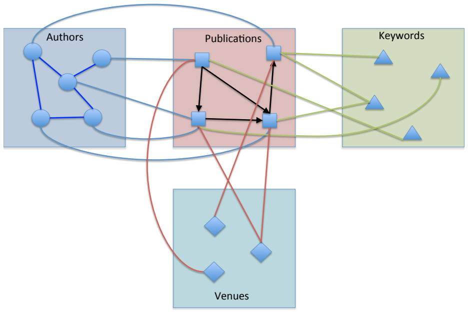

Figure 1: Query-specific multi-layer graph G_q. Each entity type comprises a layer, with connections between layers defined by natural relationships (citations, authorship, venues, relevant words).

Authors Layer: Author credibility and expertise can play an important role when looking for relevant papers. For each paper in the seed set P_s, we fetch the authors and add an *undirected and unweighted* edge (p_i, a_j) if a_j is an author of p_i. *Note that by unweighted we always mean a weight of* 1. Next, we add internal *undirected* edges (a_i, a_j, w_{ij}) between author nodes, which represent co-authorship relationships. The weight is given as $w_{ij} = 1 + \log_{10} N_{ij}$, where N_{ij} is the number of publications the authors have together.

Venues Layer: The venue can play a role in the credibility and relevance of a publication. The venues layer is assembled by creating a node v_j for each distinct venue, and adding an *undirected and unweighted* edge (p_i, v_j) if p_i appears in venue v_j.

Keywords Layer: For the final layer, we include as nodes the relevant author-defined keywords, which can yield a high quality set of topics. First, we define a vocabulary of relevant keywords, denoted V_k, as the set of all extracted keywords that appear at least five times, which removes the less representative ones. Next, we calculate TF-IDF scores for 1-grams, 2-grams and 3-grams for each publication, restricted to our vocabulary. We add a keyword node k_j to the graph, and a corresponding *undirected and weighted* edge (p_i, k_j, w_{ij}) with weight w_{ij} = tf-idf(p_i, k_j), provided the similarity is above some threshold, i.e., if tf-idf$(p_i, k_j) \geq \theta_k$.

Relevance Computation: To compute the relevance of each node in the multi-layer subgraph G_q, we modify the PageRank algorithm [3] to account for the contribution of each layer. Let $R(u_i)$ denote the relevance of node $u_i \in G_q$. The relevance vector R across all nodes can be computed via power-iteration, as follows:

$$A = \alpha \cdot \frac{1}{n_q} \cdot \mathbf{1} + (1-\alpha) \cdot W^T$$
$$R \approx A^t r, \qquad t = 1, 2, \ldots$$

where α is the random jump (or teleportation) probability, $n_q = |V_q|$ is the total number of nodes in G_q, $\mathbf{1}$ is the $n_q \times n_q$ matrix of all ones, and W is the weight matrix for the graph G_q, i.e., $W_{ij} = w_{ij}$ is the weight on the edge from u_i to u_j. The matrix A must be column stochastic to ensure convergence and the initial relevance vector r is typically chosen with all of its entries set to $1/n_q$. The final ranking vector R is the dominant eigenvector of A.

In our multi-layer relevance computation, we decompose the matrix A into several submatrices $A^{(xy)}$ whose rows are nodes from layer x and whose columns are nodes from layer y. For example, $A^{(pp)}$ is the transition matrix of all edges exclusively between publications, while $A^{(ap)}$ comprises the submatrix for the edges from authors to publications. The relevance value of a publication can then be given by the sum of contributions from nodes of the other layers weighted by a corresponding layer parameter ρ_{xy}, as follows:

$$R(p_i) = (1-\alpha)\left[\rho_{pp} \sum_{p_j \to p_i} A^{(pp)}_{ji} R(p_j) + \rho_{ap} \sum_{a_j \to p_i} A^{(ap)}_{ji} R(a_j)\right.$$
$$\left. + \rho_{vp} \sum_{v_j \to p_i} A^{(vp)}_{ji} R(v_j) + \rho_{kp} \sum_{k_j \to p_i} A^{(kp)}_{ji} R(k_j)\right] + \alpha \cdot s_i$$

where s_i is the normalized similarity between the query q and the publication p_i defined as $s_i = S(q, p_i) \Big/ \sum_{j=1}^{n_q} S(q, p_j)$, where $S(q, p_i)$ is the TF-IDF based cosine similarity between q and p_i. This equation gives the relevance for the publication nodes; relevance for nodes in the other layers can be computed in a similar manner. The key observation is that by keeping the transitions between and within the different layers separated, we control the flow between each layer by tuning the corresponding ρ_{xy} parameters. Also note that we incorporate node-specific teleportation parameters $\alpha \cdot s_i$ that allow us to bias the random jumps based on query relevance.

The relevance for publications can still be solved by power iterations over the decomposed matrix A shown below:

$$A = \begin{pmatrix} \rho_{pp} A^{(pp)} & \rho_{ap} A^{(ap)} & \rho_{vp} A^{(vp)} & \rho_{kp} A^{(kp)} \\ \rho_{pa} {A^{(ap)}}^T & \rho_{aa} A^{(aa)} & \mathbf{0} & \mathbf{0} \\ \rho_{pt} {A^{(vp)}}^T & \mathbf{0} & \rho_{vv} A^{(vv)} & \mathbf{0} \\ \rho_{pk} {A^{(kp)}}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

As depicted in Fig. 1, G_q has inter-layer edges only between publications and other layers (authors, venues and keywords). Intra-layer edges exist within publications and within authors. Finally, the only non-symmetric submatrix is $A^{(pp)}$ since citations are directed, whereas all other interactions are undirected. Correspondingly, in the transition matrix A, several sub-matrices $A^{(xy)}$ are $\mathbf{0}$, which means that there is no interaction between those layers x and y. To further reduce the ρ_{xy} parameters, we make some intuitive assumptions. First, we set $\rho_{ap} = \rho_{pa}$, $\rho_{vp} = \rho_{pv}$ and $\rho_{kp} = \rho_{pk}$, since there is no apparent reason to control the flow differently in each direction between layers. We also assume that $\rho_{aa} + \rho_{ap} = 1$, since a random walker can either stay in the authors layer or leave it. Analogously for the venues layer, $\rho_{vv} + \rho_{vp} = 1$. A random walker in the publications layer has four possible moves – go to any of the other three layers or remain in the publications layer. Therefore, $\rho_{pp} + \rho_{pa} + \rho_{pv} + \rho_{pk} = 1$. By applying these constraints and simplifying the subscripts in the parameters, we are left with four ρ values to be set: $(\rho_p, \rho_a, \rho_v, \rho_k)$, which intuitively represent the importance of each layer in the relevance calculation. The final, simplified transition matrix A is:

$$A = \begin{pmatrix} \rho_p A^{(pp)} & \rho_a A^{(ap)} & \rho_t A^{(vp)} & \rho_k A^{(kp)} \\ \rho_a {A^{(ap)}}^T & (1-\rho_a)A^{(aa)} & \mathbf{0} & \mathbf{0} \\ \rho_v {A^{(vp)}}^T & \mathbf{0} & (1-\rho_v)A^{(vv)} & \mathbf{0} \\ \rho_k {A^{(kp)}}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

IQRA-ML Algorithm: This method uses the full multi-layer graph G_q, and runs multi-layer relevance ranking using the simplified transition probability matrix A from above. After computing its dominant eigenvector R, the top k papers, authors, venues, and keywords are returned in decreasing order of rank $R(u_i)$ within each layer.

4. EXPERIMENTAL EVALUATION

Our evaluation was performed on an Intel Core i5 2.5GHz processor, with 12GB memory and 1TB disk. We used the CiteSeerX [4] dataset (denoted CSX; extracted on June 2012), which initially contained over four million publications. We assume that the CiteSeerX disambiguation solution for publications and authors, although not perfect, is correct. After eliminating publications with parsing errors, we extracted the enclosing sentence around citations as the citation contexts. We employ regular expressions to find the location of citations throughout the publication, and match each citation to its corresponding reference. If a paper is cited multiple times, we concatenate all of its contexts into a single overall context. The CSX dataset was obtained by removing all duplicate publications and authors, as well as publications that were not cited even once, which can be retrieved using text similarity only. The combined multi-layer network has 898K nodes, and about 8.43 million edges. See Table 1 for statistics; each paper on average cites 5.49 and is cited by 6.48 papers. We also tried the ArnetMiner [15] dataset, but the results were similar, so due to space constraints, we will only report results on CSX.

Node-Type	#Nodes	Edge-Type	#Edges
CiteSeerX (CSX)			
#pubs	657,119	pubs-pubs	2,730,547
#authors	186,682	authors-authors	654,717
#keywords	191,333	pubs-keywords	4,288,284
#venues	3,128	pubs-authors	753,437

Table 1: Statistics on the Entities and Links

4.1 Query Sets

We use three different query sets for evaluation. We have made these queries available publicly at:
`https://github.com/zakimjz/IQRA`.

Manual Set: We asked domain experts in the data analytics group at QCRI to choose a query and a set of 20-30 relevant papers. We further asked them to group the papers into two categories: R1 for highly relevant and R2 for other relevant publications. We collected 9 queries and responses (see Table 2). The "schema matching" query was submitted by two different experts, with different responses.

Set	#Queries	Query examples
Manual	9	subgraph pattern mining, data exchange sentiment analysis, subspace clustering schema matching (2), record linkage graph clustering, spectral clustering
Surveys	100	monte carlo tree search text clustering privacy preserving data publishing
Citations	200	tutorial multiple view geometry expressive power deep architectures representing cyclic human motion using functional analysis

Table 2: Query Set Sizes and Sample Queries

Citations Set: We randomly selected a publication and used its title (after removing stop-words) as the query and its reference section as the ground-truth. We assume that the authors have done due diligence in the literature review and have cited the most relevant publications. We selected 200 queries from CSX (see Table 2 for examples). We remove all the query set papers from the bibliographic dataset before querying. In addition, we also avoid selecting publications as ground-truth if there exists another very similar publication, such as the journal version of a conference paper. Such pairs share multiple citations in common, and can therefore artificially inflate the relevance.

Surveys Set: Survey authors are more likely to have done a comprehensive search to highlight the work in a given area, more so than regular papers that may be limited due to space or editorial policy on the number of citations they can include. Therefore, we searched for paper titles containing the term *survey* and then manually selected actual survey articles. The query was extracted from the title by removing stop-words and other irrelevant words (including the word "survey"). For instance, the title *A Survey of Text Summarization Techniques* simply becomes the query *text summarization*. We selected 100 surveys from CSX (see Table 2 for examples). The cited papers are used as the ground-truth. As before, all the survey papers in the query set are removed from the bibliographic dataset before testing and evaluation.

4.2 Baseline Methods and Parameter Settings

For IQRA a user can tune the algorithm using several parameters. For example, a user may want papers directly relevant to a query, while another may want distantly related papers; one may want recent papers, while another the earliest works. Our model allows such tuning via the ρ parameters. However, for comparison, we use the following parameters values tuned on an independent tuning query set comprising 100 (random) publications from CSX (also available at `https://github.com/zakimjz/IQRA`): $K = 20$, $H = 1$, $\omega = 0.5$, $\sigma = 0.3$, $\gamma = 0.01$, $\alpha = 0.3$, and $\theta_k = 1.0$. The ρ parameters are set to the default value of 0.25.

We compared IQRA-TC and IQRA-ML against several state-of-the-art competing methods listed below:

TF-IDF and BM25: These are based purely on textual similarity. For TF-IDF, we simply rank the publications using the cosine similarity between the query and the document (using title and abstract). For BM25, we use the Okapi BM25 scoring function [6] instead of TF-IDF.

TopCited: We retrieve documents from the entire dataset D containing the query terms, and then rank them by the number of times they are cited.

	Manual Set		Surveys Set		Citations Set	
Method	MAP@20	NDCG@20	MAP@20	NDCG@20	MAP@20	NDCG@20
Iqra-tc	0.273 ± 0.116	**0.530 ± 0.146**	**0.197 ± 0.184**	**0.356 ± 0.231**	**0.119 ± 0.132**	**0.251 ± 0.196**
Iqra-ml	**0.284 ± 0.152**	0.525 ± 0.167	0.173 ± 0.173	0.326 ± 0.222	0.103 ± 0.117	0.226 ± 0.183
Okapi BM25	0.056 ± 0.034	0.207 ± 0.110	0.056 ± 0.078	0.143 ± 0.149	0.024 ± 0.044	0.076 ± 0.097
TF-IDF	0.063 ± 0.046	0.197 ± 0.110	0.057 ± 0.083	0.146 ± 0.152	0.025 ± 0.044	0.080 ± 0.098
TopCited	0.004 ± 0.004	0.034 ± 0.034	0.010 ± 0.026	0.038 ± 0.073	0.009 ± 0.021	0.032 ± 0.060
CiteRank	0.002 ± 0.004	0.015 ± 0.031	0.007 ± 0.024	0.027 ± 0.063	0.005 ± 0.016	0.020 ± 0.048
PageRank (pre)	0.026 ± 0.043	0.089 ± 0.091	0.015 ± 0.032	0.049 ± 0.081	0.008 ± 0.020	0.032 ± 0.058
PageRank (pos)	0.001 ± 0.002	0.006 ± 0.016	0.007 ± 0.023	0.027 ± 0.062	0.005 ± 0.016	0.019 ± 0.047
PageRank (G_q)	0.166 ± 0.097	0.400 ± 0.143	0.126 ± 0.121	0.270 ± 0.187	0.077 ± 0.097	0.191 ± 0.159
GoogleScholar	0.247 ± 0.214	0.428 ± 0.250	0.033 ± 0.054	0.106 ± 0.130	0.014 ± 0.023	0.057 ± 0.077
ArnetMiner	0.177 ± 0.124	0.332 ± 0.183	0.020 ± 0.045	0.065 ± 0.107	0.001 ± 0.005	0.003 ± 0.017

Table 3: Performance Comparison for Manual, Surveys, and Citations Query Sets on the CSX Dataset

PageRank: We consider three variants: (i) *Pre-Filter:* Run PageRank on the entire bibliographic dataset D, and then retain only those publications that contain the query terms. (ii) *Post-Filter:* Retrieve publications that contain the query terms, and then run PageRank for ranking. (iii) G_q: Run PageRank on the publications layer in our query-specific subgraph G_q. The teleportation parameter was set to $\alpha = 0.3$.

CiteRank [17]: CiteRank uses personalized teleportation factors for each node and also uses age attenuation. See [17] for details; the attenuation parameter was $\tau = 2.6$ as suggested by the author.

GoogleScholar (`scholar.google.com`): We query Google Scholar and retain only those publications found in our bibliographic dataset D, and return the top-k papers.

ArnetMiner (`www.arnetminer.org`): For each query, we extract the top-k results from ArnetMiner.

We also compared with HITS [7] and the CiteSeerX web service (`citeseerx.ist.psu.edu`), but the results were similar to PageRank and are not shown. Unfortunately, we are not able to compare to ClusCite [11], since its Matlab-based code did not finish running even after three days on the CSX dataset.

4.3 Evaluation Metrics

We use MAP and NDCG as evaluation metrics [8].

Mean Average Precision (MAP): MAP is defined as

$$MAP@k = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \left(\frac{1}{\min(m,n)} \sum_{i=1}^{k} P@i \right)$$

where $|Q|$ is the number of queries, k denotes the number of top items, n is the number of results returned, m is the number of relevant items, and $P@i$ is precision at the top i items, i.e., the fraction of relevant papers in the top-i returned results. If either m or n is 0, then $AP@k$ is also 0.

Normalized Discounted Cumulative Gain (NDCG): NDCG considers the rank and the relative relevance of each item to measure retrieval effectiveness. It is defined as:

$$NDCG@k = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \left(\sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2 (i+1)} \Big/ IDCG@k \right)$$

where r_i is the relevance of the item found at rank position i, and $IDCG@k$ is an ideal ranking of results such that $NDCG@k = 1$ if a perfect ranking is returned (top relevant item first, followed by second most relevant item, and so on). For the Manual query set we have $r_i = 2$ for more relevant and $r_i = 1$ for less relevant documents. For the other query sets, $r_i = 1$ if the paper is relevant and $r_i = 0$ if not.

4.4 Retrieval Performance Comparison

Comparative performance results on the CSX dataset are shown in Table 3. For each metric, we report the average as well as the standard deviation. We observe that across all three query sets, Iqra-tc is the best, followed closely by Iqra-ml.

For the Manual query set, GoogleScholar is also quite effective, although Iqra-tc and Iqra-ml are still better. This is perhaps not very surprising since many of the experts in fact used GoogleScholar as a means to search for relevant papers in addition to using their own expert knowledge and ranking of the papers (which need not have matched those obtained from GoogleScholar). GoogleScholar did not fare too well on the Surveys or the Citations query set. Also, ArnetMiner generally performs worse than GoogleScholar on all three query sets. As such both GoogleScholar and ArnetMiner have access to the entire bibliographic dataset, including the citations for the paper corresponding to the query being searched. On the other hand, we remove the query paper and its citation links. Even then, our methods are able to retrieve a more relevant set of papers.

PageRank-based approaches, including pre/post and CiteRank do not perform very well. However, when we run PageRank on the publications layer from our query-specific subgraph G_q, the performance is much better. Purely text-based methods like BM25 and TF-IDF do better than PageRank (pre/post) on all three query sets. These results indicate that query relevance is a very important characteristic in related publication retrieval. Further, the fact that Iqra-ml does well indicates that the multiple layers can play an important role in the relevance flow, resulting in more relevant publications.

The average performance on the Surveys set is much lower for all methods compared to the Manual query set, whereas the performance on the Citations set is even lower. Further, we observe that both experts and Surveys may be more biased towards top-cited papers. These two effects can be explained by the fact that Citations queries (i.e., regular paper titles) are usually very specific, and papers cite only a limited number of relevant publications. The Survey queries are usually more general, and typically survey papers cite many more papers. Authors of surveys almost never miss highly-cited papers. For Citations, this effect is less; while regular papers may cite highly cited papers, they also tend to cite other more recent papers with a limited initial citation count. However, as shown for the Manual query set, experts also tend to consider citation counts when ranking papers, though they probably consider other factors such as author and venue reputation. In these cases, the strength of the multi-layer approach is most evident. Whereas Iqra-tc performs the best, interestingly, TopCited (on the entire network) performs rather poorly. This finding indicates that our approach of selecting the documents similar to the query, followed by expansion using the citations, really helps in focusing the attention to relevant papers.

4.5 Timing Comparison

Table 4 shows the average times and the standard deviation across different queries for each of the query sets. We can see that Iqra-tc is among the fastest methods, and Iqra-ml also has good performance.

Query sets	Manual	Surveys	Citations
Iqra-tc	3.1 ± 1.5	**1.9 ± 0.7**	**2.4 ± 1.4**
Iqra-ml	7.9 ± 11.1	7.2 ± 2.4	6.1 ± 2.8
Okapi BM25	**1.8 ± 1.9**	4.7 ± 3.2	6.7 ± 4.0
TF-IDF	**1.8 ± 1.9**	4.7 ± 3.2	6.9 ± 4.0
TopCited	2.5 ± 3.4	5.2 ± 3.5	7.3 ± 4.2
CiteRank	20.8 ± 37.6	10.9 ± 3.2	12.7 ±3.8
PageRank (pre)	4.3 ± 5.2	13.2 ± 10.1	18.2 ± 12.6
PageRank (pos)	4.7 ± 1.9	7.6 ± 3.2	9.4 ± 3.8
PageRank (G_q)	3.1 ± 1.5	**1.9 ± 0.7**	**2.4 ± 1.4**

Table 4: Timing Comparison on CSX (in seconds)

4.6 Effect of Layers

We also investigated the effect of the different layers in our model. Table 5 shows the MAP@20 scores for various layers using the Manual, Surveys and Citations query sets on CSX. We observe that, as expected, the publications layer (P) plays a major role. However, adding authors (A), venues (V) and keywords (W) helps boost the performance even further. The PAWV model combines all of the layers and performs the best.

Layers	Manual Set	Surveys Set	Citations Set
P	0.188 ± 0.094	0.137 ± 0.129	0.077 ± 0.092
PA	0.205 ± 0.099	0.151 ± 0.144	0.086 ± 0.094
PV	0.187 ± 0.099	0.137 ± 0.130	0.077 ± 0.092
PW	0.205 ± 0.094	0.162 ± 0.151	0.093 ± 0.102
PAV	0.212 ± 0.099	0.151 ± 0.146	0.085 ± 0.094
PAW	0.216 ± 0.094	**0.171 ± 0.160**	0.098 ± 0.105
PWV	0.205 ± 0.093	0.163 ± 0.150	0.094 ± 0.104
PAWV	**0.223 ± 0.103**	**0.170 ± 0.161**	**0.097 ± 0.105**

Table 5: Effect of Different Layers (MAP@20): Publications (P), Authors (A), Venues (V), and KeyWords (W).

5. CONCLUSION

We have proposed a two-step approach for entity relevance and recommendation given a user-specified query. Instead of performing a query-independent search, we show that our strategy of staged query-dependent layer selection is much more effective. This is mainly due to two reasons, namely, fast pruning of irrelevant data, and query-dependent ranking propagation. Results on benchmark query sets show that our approach is more effective than existing methods. Our main conclusion is that for finding the most relevant citations for a paper, our top-cited method Iqra-tc serves well. The multi-layer approach Iqra-ml is a close second, but it has the potential for a more thorough literature survey by suggesting related entities like authors, venues and keywords. Showing the effectiveness of these extra layers is part of our ongoing work.

References

[1] C. Basu, H. Hirsh, and W. W. Cohen. Technical paper recommendation: A study in combining multiple information sources. *Journal of Artificial Intelligence Research*, 14:231–252, 2001.

[2] J. Beel, B. Gipp, S. Langer, and C. Breitinger. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, pages 1–34, 2015.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[4] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *ACM Conference on Digital Libraries*, 1998.

[5] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. Context-aware citation recommendation. In *International Conference on World Wide Web*, 2010.

[6] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 1 & 2. *Information Processing & Management*, 36(6):779–808, 809–840, 2000.

[7] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5):604–632, 1999.

[8] C. D. Manning, P. Raghavan, and H. SchÃijtze. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, UK, 2008.

[9] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. On the recommending of citations for research papers. In *ACM Conference on Computer Supported Cooperative Work*, 2002.

[10] F. Meng, D. Gao, W. Li, X. Sun, and Y. Hou. A unified graph model for personalized query-oriented reference paper recommendation. In *ACM Conference on Information and Knowledge Management*, 2013.

[11] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. Cluscite: Effective citation recommendation by information network-based clustering. In *ACM International Conference on Knowledge Discovery and Data Mining*, 2014.

[12] A. Ritchie, S. Robertson, and S. Teufel. Comparing citation contexts for information retrieval. In *ACM Conference on Information and knowledge management*, 2008.

[13] T. Strohman, W. B. Croft, and D. Jensen. Recommending citations for academic papers. In *ACM Conference on Research and Development in Information Retrieval*, 2007.

[14] J. Tang and J. Zhang. A discriminative approach to topic-based citation recommendation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2009.

[15] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, 2008.

[16] M. Valenzuela, V. Ha, and O. Etzioni. Identifying meaningful citations. In *AAAI Workshop on Scholarly Big Data*, 2015.

[17] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a simple model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, (06):P06010, 2007.

[18] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *ACM International Conference on Knowledge Discovery and Data Mining*, 2011.

[19] N. Zheng and Q. Li. A recommender system based on tag and time information for social tagging systems. *Expert Systems with Applications*, 38(4):4575–4587, 2011.

[20] D. Zhou, S. Zhu, K. Yu, X. Song, B. L. Tseng, H. Zha, and C. L. Giles. Learning multiple graphs for document recommendations. In *International Conference on World Wide Web*, 2008.