Direction Awareness in Citation Recommendation

Onur Küçüktunç[1,2], Erik Saule[1], Kamer Kaya[1], Ümit V. Çatalyürek[1,3]
[1] Dept. Biomedical Informatics, The Ohio State University
[2] Dept. Computer Science and Engineering, The Ohio State University
[3] Dept. Electrical and Computer Engineering, The Ohio State University
{kucuktunc,esaule,kamer,umit}@bmi.osu.edu

ABSTRACT

Literature search is an important part of academic research. The increase in the number of published papers each year makes manual search inefficient, hence, automatic methods must be devised. Unfortunately, traditional search engines use keyword-based approaches to solve the search problem which are prone to ambiguity and synonymity. This paper focuses on the problem of extending a set of references using the citation relations between the documents. In particular, we introduce the class of *direction-aware* algorithms which weight the importance of incoming and outgoing edges of the citation graph differently based on user preferences. Using such an algorithm, the user can easily focus her search toward recent developments or traditional papers. We present two direction-aware algorithms and show that they are better suited at solving the problem at hand than state-of-the-art recommendation methods. One of these algorithms is currently deployed in a publicly available web-service called the**advisor**.

1. INTRODUCTION

The academic community has published millions of research papers to date, and the number of new papers has been increasing with time. For example, based on DBLP[1], computer scientists published 3 times more papers in 2010 than in 2000 (see Figure 1-left). With more than one hundred thousand new papers each year, performing a complete literature search became a herculean task. A paper cites 20 other papers on average (see Figure 1-right for the distribution of citations in our data), which means that there might be more than a thousand papers that cite or are cited by the papers referenced in a research article. Researchers typically rely on manual methods to discover new research such as keyword-based search via search engines, reading proceedings of conferences, browsing publication list of known experts or checking the reference list of the paper they are interested. These techniques are time-consuming and only allow to reach a limited set of documents in a reasonable time. Developing tools that help researchers to find unknown and relevant papers will certainly increase the productivity of the scientific community.

[1]statistics based on data acquired from DBLP in Dec'11

DBRank'12, in conjunction with VLDB, August 27, 2012, Istanbul, Turkey

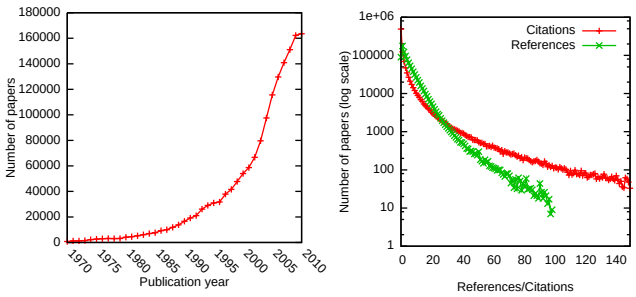


Figure 1: Number of new papers published each year based on DBLP (left), and number of papers with given citation and reference count (right).

Some of the existing approaches and tools for the literature search cannot compete with the size of today's literature. Keyword-based approaches suffer from the confusion induced by different names of identical concepts in different fields. (For instance, *partially ordered set* or *poset* are also often called *directed acyclic graph* or *DAG*). Conversely, two different concepts may have the same name in different fields (for instance, *hybrid* is commonly used to specify software hybridization, hardware hybridization, or algorithmic hybridization). These two problems may drastically increase the number of suggested but unrelated papers.

To alleviate the above mentioned problems, we built a web service called the**advisor**[2]. It takes a bibliography file containing a set of papers, i.e., *seeds*, as an input to initiate the search. The user can specify that she is interested in classical papers or in recent papers. Then, the service returns a set of suggested papers ordered with respect to a ranking function. The service works using only the *citation graph* of known bibliography. In other words, it does not take the textual data into account because our aim is finding all conceptually related and high quality documents even if they use a different terminology. It has been shown that text-based similarity is not sufficient for this task and that most of the relevant informations are contained within the citation graph [18]. Besides, it is plausible that there is already a correlation between citation similarities and text similarities of the papers [16].

RefSeer[3] is another webservice that shares our goals but uses a very different approach. It aims at providing relevant references of an existing text by discovering its main topics and suggests the most famous works within each topic of

[2]`http://theadvisor.osu.edu/`
[3]`http://refseer.ist.psu.edu/`

interest. Therefore, it tends to suggest only very well-cited documents. We believe a citation based approach will be more local and will provide the opportunity of finding papers that are not popular but still highly relevant.

There are various citation-analysis-based paper recommendation methods depending on a pairwise similarity measure between two papers. Bibliographic coupling, which is one of the earliest works, considers papers having similar citations as related [5]. Another early work, Cocitation, considers papers which are cited by the same papers as related [17]. A similar cites/cited approach by using collaboration filtering is proposed by McNee et al. [14]. CCIDF also considers only common citations, but by weighting them with respect to their inverse frequencies [7].

More recent works define different measures such as Katz which is proposed by Liben-Nowell and Kleinberg for a study on the link prediction problem on social networks [11] and used later for information retrieval purposes including citation recommendation by Strohman et al. [18]. For two papers in the citation network, the Katz measure counts the number of paths by favoring the shorter ones. Lu et al. stated that both bibliographic coupling and Cocitation methods are only suitable for special cases due to their very local nature [12]. They proposed a method which computes the similarity of two papers by using a vector based representation of their neighborhoods in the citation network. Liang et al. argued that most of the methods stated above considers only direct references and citations alone [10]. Even Katz and the vector based method of [12] consider the links in the citation network as simple links. Instead, Liang et al. added a weight attribute to each link and proposed the method Global Relation Strength which computes the similarity of two papers by using a Katz-like approach.

Many works use random walk with restarts (RWR) for citation analysis [3, 6, 9, 13]. RWR is a well known and efficient technique used for different tasks including computing the relevance of two vertices in a graph [15]. It is very similar to the well known PageRank algorithm [1] which is used by both Li and Willett [9] (ArticleRank) and Ma et al. [13] to evaluate the importance of the academic papers. Gori and Pucci [3] proposed an algorithm, called PaperRank, for RWR-based paper recommendation which can also be seen as a Personalized PageRank computation [4] on the citation graph. Lao and Cohen [6] also used RWR for paper recommendation in citation networks and proposed a learnable proximity measure for weighting the edges by using machine learning techniques. As far as we know, none of these works study the recent/traditional paper recommendation problem. The closest work is Claper [19] which is an automatic system that measure how much a paper is classical, allowing to rank a list of paper to highlight the most classical ones.

Our aim in this work is to evaluate the existing algorithms and to explain the new algorithms that power the**advisor**. We introduce a class of parametric algorithms, said to be *direction aware*, which allow to give more importance to either the citation of papers or their references. They make the citation suggestion process easily tunable for finding either recent or traditional relevant papers. In particular we extend two eigenvector based methods into direction-aware algorithms, namely DaRWR and DaKatz. These algorithms are compared to state-of-the-art citation-based algorithms for bibliographic recommendation and their adequation to the problem is studied.

2. PROBLEM AND SOLUTIONS

Let $G = (V, E)$ be the *citation graph*, with n papers $V = \{v_1, \ldots, v_n\}$. In G, each directed edge $e = (v_i, v_j) \in E$ represents a *citation* from v_i to v_j. For the rest of the paper, we use the phrases *"references of v"* and *"citations to v"* as to describe the graph around vertex v (see Figure 2). We use $deg^-(v)$ and $deg^+(v)$ to denote the number of references of and citations to v, respectively.

In this work, we target the problem of paper recommendation assuming that the researcher has already collected a set of papers in the manuscript preparation [18]. Therefore, the objective is to return papers that the given manuscript might cite:

Paper recommendation (PR): Given a set of m seed papers $\mathcal{M} = \{p_1, \ldots, p_m\}$ s.t. $\mathcal{M} \subseteq V$, and a parameter k, return top-k papers which are relevant to the ones in $\mathcal{M}$.

2.1 Random walk with restart

RWR is widely used in many fields. In citation analysis, PaperRank [3] is a method based on random walks in the citation graph G. However, the current structure of G is not suitable for finding recent and relevant papers since such papers have only a few incoming edges. Moreover, since the graph is acyclic, all random walks will end up on old papers. To alleviate this, given a PR query with inputs $\mathcal{M}$ and k, PaperRank constructs a directed graph $G' = (V', E')$ by slightly modifying the citation graph G as follows:

A source node s is added to the vertex set: $V' = V \cup \{s\}$. Back-reference edges (E_b), the edges from s to seed papers (E_f), and restart edges from V to s (E_r) are added to the graph: $E_b = \{(y, x) : (x, y) \in E\}$, $E_f = \{(s, v) : v \in \mathcal{M}\}$, $E_r = \{(v, s) : v \in V\}$, and $E' = E \cup E_b \cup E_f \cup E_r$.

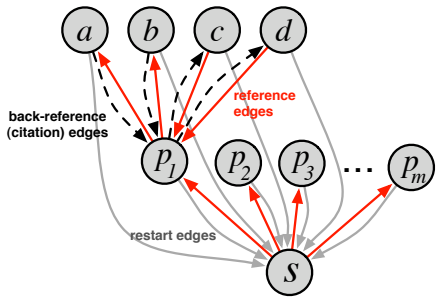


Figure 2: Citation graph with source node s and seed set $\mathcal{M} = \{p_1, \ldots, p_m\}$. The papers a and b are cited by p_1, and c and d cites p_1. Note that there is a corresponding back-reference edge for each reference.

The new directed graph G' has *reference* (red), *back reference* (dashed), and *restart* (gray) edges (see Figure 2). In this model, the random walks are directed towards both references and citations of the papers. In addition, the restarts from the source vertex s will be distributed to only the seed papers in $\mathcal{M}$. Hence, random jumps to any paper in the literature are prevented. We assume that a random walk ends in v continues with a neighbor with a damping factor $d \in (0, 1]$. And with probability $(1 - d)$, it restarts and goes to the source s. Let $R_{t-1}(v)$ be the probability of a random walk ends at vertex $v \neq s$ at iteration $t - 1$. Let $C_t(v)$ be the contribution of v to one of its neighbors at iteration

t. In each iteration, d of $R_{t-1}(v)$ is distributed among its references and citations equally. Hence,

$$C_t(v) = d\frac{R_{t-1}(v)}{deg^+(v) + deg^-(v)}. \quad (1)$$

Initially, a probability score of 1 is given to the source node, meaning that a researcher expands the bibliography starting with the paper itself: $R_0(s) = 1$ and $R_0(v) = 0$ for all $v \in V$, where R_0 is the probability at $t = 0$. PaperRank algorithm computes the probability of a vertex u at iteration t as

$$R_t(u) = \begin{cases} (1-d)\sum_{v\in V} R_{t-1}(v), & \text{if } u = s \\ \sum_{(u,v)\in E} C_t(v) + \frac{R_{t-1}(s)}{|\mathcal{M}|}, & \text{if } u \in \mathcal{M} \\ \sum_{(u,v)\in E} C_t(v), & \text{otherwise.} \end{cases} \quad (2)$$

PaperRank converges when the probability of the papers are stable. Let Δ_t be the difference vector. We say that the process is in a *steady state* when the L2 norm of Δ_t is smaller than a given value ϵ.

2.2 Direction-aware random walk with restart

A random walk with restart is a good way to find relevance scores of the papers. However, the PaperRank algorithm treats the citations and references in the same way. This may not lead the researcher to recent and relevant papers if she is more interested with those. Old and well cited papers have an advantage with respect to the relevance scores since they usually have more edges in G'. Hence G' tends to have more and shorter paths from the seed papers to old papers. We define a *direction-awareness* parameter $\lambda \in [0, 1]$ to obtain more recent results in the top-k documents. We then define two types of contributions of each paper v to a neighbor paper in iteration t:

$$C_t^+(v) = d\lambda\frac{R_{t-1}(v)}{deg^+(v)}, \quad (3)$$

$$C_t^-(v) = d(1-\lambda)\frac{R_{t-1}(v)}{deg^-(v)}, \quad (4)$$

where $C_t^-(v)$ is the contribution of v to a paper in its reference list and $C_t^+(v)$ is the contribution of v to a paper which cites v. Hence, for a non-seed, non-source paper u,

$$R_t(u) = \sum_{(v,u)\in E_b} C_t^+(v) + \sum_{(v,u)\in E} C_t^-(v). \quad (5)$$

For a seed node u, $R_t(u)$ is computed similarly except that each seed node has an additional $\frac{R_{t-1}(s)}{|\mathcal{M}|}$ in the equation. $R_t(s)$ is computed in the same way as (2). With this modification, the parameter λ can be used to give more importance either to traditional papers with $\lambda \in [0, 0.5]$ or recent papers with $\lambda \in [0.5, 1]$. We call this algorithm *direction-aware random walk with restart* (DaRWR).

Note that DaRWR (5) has the *probability leak* problem when a paper has no references or citations. If this is the case, some part of its score will be lost at each iteration. For such papers, we distribute the whole score from the previous iteration towards only its references or citations.

2.3 Katz and direction awareness

The direction awareness can be also adapted to other similarity measures such as the graph-based Katz distance measure [11] which was used before for the citation recommendation purposes [18]. With Katz measure, the similarity score between two papers $u, v \in V$ is computed as

$$Katz(u, v) = \sum_{i=1}^{L} \beta^i |paths^i_{u,v}|, \quad (6)$$

where $\beta \in [0, 1]$ is the decay parameter, L is an integer parameter, and $|paths^i_{u,v}|$ is the number of paths with length i between u and v in the graph with paper and back-reference edges $G'' = (V, E \cup E_b)$. Notice that the path does not need to be elementary, i.e., the path $uvuv$ is a valid path of length 3. Therefore the Katz measure might not converge for all values of β when $L = \infty$. β needs to be chosen smaller than the larger eigenvalue of the adjacency matrix of G''. And in practice L is set to a fixed value (in our experiment $L = 10$). In our context with multiple seed papers, the relevance of a paper v is set to $R(v) = \sum_{u\in\mathcal{M}} Katz(u, v)$.

We extend the Katz distance by using direction awareness to weight the contributions to references and citations differently with the λ parameter as in DaRWR:

$$DaKatz(u, v) = \sum_{i=1}^{L} \left[\lambda\beta^i |Rpaths^i_{u,v}| + (1-\lambda)\beta^i |Cpaths^i_{u,v}|\right],$$

where $|Rpaths^i_{u,v}|$ (respectively, $|Cpaths^i_{u,v}|$) is the number of paths in which the last edge in the path is a reference edge of E (respectively, a citation edge of E_b).

3. EXPERIMENTS

3.1 Dataset collection

The retrieval of bibliographic information and citation graph generation is a difficult task since academic papers are generally copyrighted and they are accessible through publishers' digital libraries. Therefore, we limited our study to data with license that explicitly allow data mining.

We retrieved information on 1.9M computer science articles (as of March 2012) from DBLP[4] [8], 740K technical reports on physics, mathematics, and computer science from arXiv[5], and 40K publications from HAL-Inria[6] open access library. This data is well-formatted and disambiguated; however, it contains very few citation information (less than 470K edges). CiteSeer[7] is used to increase the number of paper-to-paper relations of computer science publications, but most of its data are automatically generated [2] and are often erroneous. We mapped each document in CiteSeer to at most one document in each dataset with the title information (using an inverted index on title words and Levenshtein distance) and publication years. Using the disjoint sets, we merged the papers and their corresponding metadata from four datasets. The papers without any references or incoming citations are discarded. The final citation graph has about 1M papers and 6M references, and is currently being used in our service.

[4] http://www.informatik.uni-trier.de/~ley/db/
[5] http://arxiv.org/
[6] http://hal.inria.fr/
[7] http://citeseerx.ist.psu.edu/

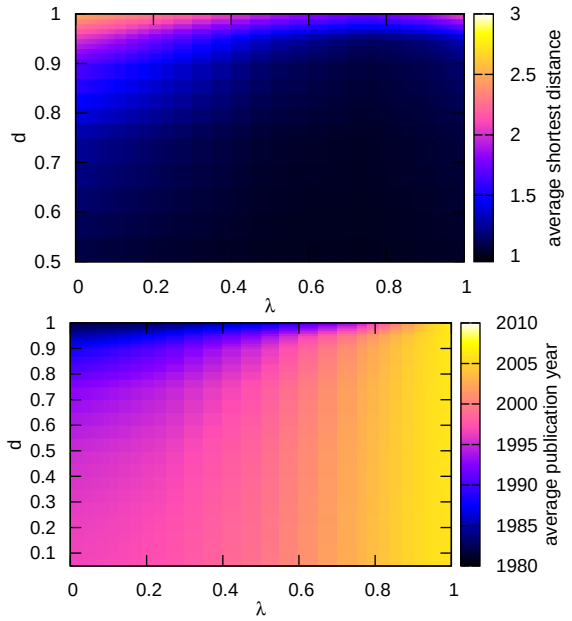


Figure 3: Average shortest distance from seed papers (top) and publication year (bottom) of top-10 recommendations by DaRWR based on d and λ.

3.2 Parameter tests

Before comparing the different methods presented in the paper, we study the impact of the damping factor d and the direction-awareness parameter λ on the papers recommended by the DaRWR algorithm. In particular, we want to verify that changing these parameters allows the user to obtain suggestions that are either closer to or farther away from the seed papers $\mathcal{M}$, and to obtain suggestions that are either recent or more traditional. To verify these effects, a source paper published between 2005 and 2010 is randomly selected and its references are used as the seed papers. We use the top-10 results as the set of recommended papers $\mathcal{R}$. The test is repeated for 2500 distinct queries that satisfy the given constraints.

Figure 3 (top) shows the impacts of d and λ on the average year of $\mathcal{R}$ and average shortest distance in the citation graph between $\mathcal{R}$ and $\mathcal{M}$. When d increases, the probability that the random research jumps back to the source node s is reduced. It allows reaching vertices distant from s to be reached more often. λ makes little difference in the average distance to the seed papers. However, setting a higher value of d should allow to find relevant papers whose relation to the seeds are not obvious.

Figure 3 (bottom) also shows that increasing d leads to earlier papers since they tend to accumulate more citations. But for a given λ, varying the damping factor do not allow to reach a large diversity of time frames. The direction-awareness parameter λ can be adjusted to reach papers from different years with a range from late 1980's to 2010 for almost all values of d. In our online service, the parameter λ can be set to a value of user's preference. It allows the user to obtain recent papers by setting λ close to 1 or finding older papers by setting λ close to 0.

Overall, first-level papers are often returned for $d < 0.8$; yet many papers at distance 2 and more appear. Also, it is possible to choose between traditional papers (by setting $\lambda < 0.4$) or recent papers (by setting $\lambda > 0.8$).

3.3 Experimental settings

We test the quality of the recommended citations by different methods in three different scenarios.

The **hide random** scenario represents the typical use-case where a researcher is writing a paper and trying to find some more references. To simulate that, a source paper s with enough references ($20 \leq deg^+(s) \leq 100$) is randomly selected from the papers published between 2005 and 2010. Then we remove s and all the papers published after s from the graph (i.e., $G_s = (V_s, E_s)$ where $V_s \subset V \setminus \{s\}$ and $\forall v \in V_s$, $year[v] \leq year[s]$) to simulate the time when s was being written. Out of $deg^+(s)$, 10% of the references are randomly put in the hidden set H, and the rest is used as the seed papers (i.e., $\mathcal{M} = \{v \notin H : (s, v) \in E\}$). We compute the citation recommendations on $\mathcal{M}$ and report the mean average precision (MAP) of finding hidden papers within the top-50 recommendations for 2500 independent queries.

The **hide recent** scenario represents another typical use-case where the author might be well aware of the literature of her field but might have missed some recent developments. Here, the hidden set H only contains the most recent references. Again, MAP of finding hidden papers within the top-50 recommendations is reported for each query. Similarly, we define **hide earlier** where the hidden papers are the oldest publications.

The methods we proposed are compared on the three scenarios against widely-used citation based approaches: bibliographic coupling [5], Cocitation [17], CCIDF [7], PaperRank [3] and the original Katz distance [11]. The algorithms and the parameters that lead to the best accuracy in different experiments are summarized in Table 1.

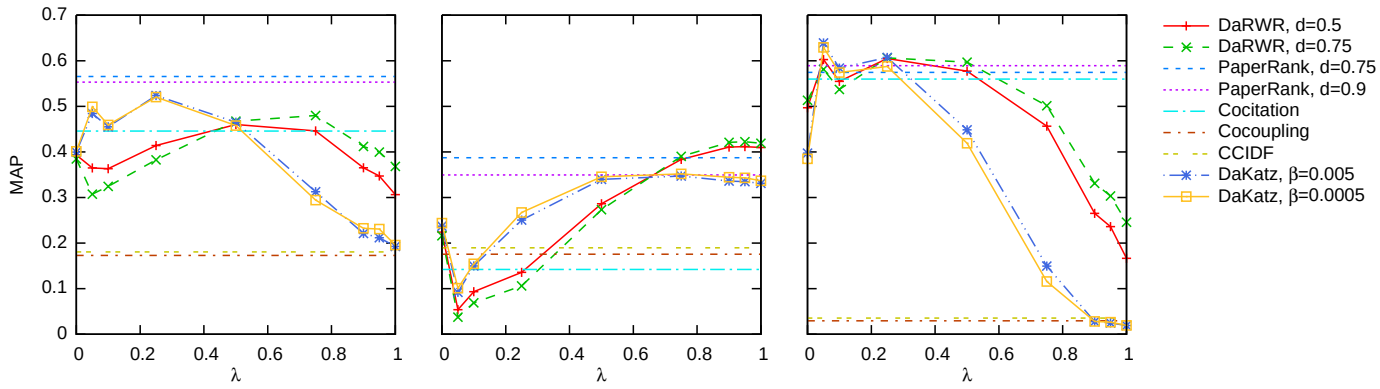
Table 1: Parameters used in the experiments.

Method	Random	Recent	Earlier
Katz$_\beta$	$\beta = 0.0005$		$\beta = 0.005$
DaKatz	$\beta = 0.005$ $\lambda = 0.25$	$\beta = 0.0005$ $\lambda = 0.75$	$\beta = 0.005$ $\lambda = 0.05$
PaperRank	$d = 0.75$	$d = 0.75$	$d = 0.9$
DaRWR	$d = 0.75$ $\lambda = 0.75$	$d = 0.75$ $\lambda = 0.95$	$d = 0.75$ $\lambda = 0.25$

3.4 Results

Accuracy: Figure 4 presents a comparison of all the methods on there scenarios. Many algorithms are represented as horizontal lines since they are not direction aware. The first remark is that Cocoupling and CCIDF perform poorly on all four scenarios. Cocitation performs the worst in the hide recent scenario and performs reasonably good but not the best in the other scenarios. These methods only consider counting and weighting distance-2 papers from the seeds, and they are outperformed by the eigenvector-based methods which take whole graph into account.

Notice that PaperRank performs well overall but for different values of the damping parameter d. The performance of DaKatz is significantly varying with the parameter set, but it is important to notice that the variations with the direction-awareness parameter are similar to the one observed on DaRWR. The results of Katz are not explicitly presented but can be read on DaKatz when $\lambda = 0.5$. Notice that DaKatz is always a better method that Katz. PaperRank achieves the best results when the query is generic (on



(a) hide random (b) hide recent (c) hide earlier

Figure 4: Mean average precision of the algorithms on three experiments based on λ and other parameters. Note that Katz is equal to DaKatz at $\lambda = 0.5$.

Table 2: Results of the experiments with mean average precision and 95% confidence intervals.

	hide random			hide recent			hide earlier		
	mean	interval		mean	interval		mean	interval	
DaRWR	48.00	46.80	49.20	42.22	**40.95**	**43.50**	60.64	59.48	61.80
P.R.	56.56	**55.31**	**57.80**	38.75	37.50	40.00	58.93	57.76	60.10
DaKatz	52.39	51.18	53.60	35.18	33.96	36.40	63.93	**62.76**	**65.10**
Katz$_\beta$	46.33	45.16	47.50	34.56	33.42	35.70	44.19	42.97	45.40
Cocit	44.60	43.39	45.80	14.22	13.25	15.20	55.97	54.64	57.30
Cocoup	17.28	16.36	18.20	17.56	16.61	18.50	2.93	2.57	3.30
CCIDF	18.05	17.11	19.00	18.97	17.94	20.00	3.55	3.10	4.00

the hide random scenarios); however direction-aware methods lead to higher accuracy when the query is targeted.

The accuracy being close to each other, we report in Table 2 the 95% confidence interval for the best parameters of each method on the three scenarios. In each scenario, the confidence interval of the method that performs best does not intersect with any other interval. It indicates that their dominance is statistically significant.

Coverage: The previous experiments show a statistically significant but little difference in accuracy between the best method and the runner-up. We investigate whether the recommended documents are similar or different. Table 3 presents the intersection matrix of the different methods on the three scenarios for a limited number of queries. Each method's parameters are set to optimize the accuracy. The diagonal of the matrix shows the actual accuracy of the methods. Other values show the MAP of the intersection of the solutions of the two corresponding methods. DaKatz clearly dominates Katz. Cocitation and CCIDF recommend different documents (up to 8%). DaRWR and PaperRank can show significant differences (up to 5.6%) as well. In each scenario, the best algorithm can not be improved more than 7% using the solution of another algorithm.

Citation patterns: The large variation of the accuracy when the direction-awareness parameter varies indicates that searching for old papers is inherently different than searching for recent papers or arbitrary papers. We believe that traditional papers and recent papers cite and are being cited differently. To qualify this difference, we study the properties of the suggestions returned by the methods and compare them to the properties of the actual references within the papers. We argue that an appropriate method should suggest

Table 3: Intersection matrix of the results for hide random (i), recent (ii), and earlier (iii) experiments.

(i)	DaRWR	P.R.	DaKatz	Katz$_\beta$	Cocit	Cocoup	CCIDF
DaRWR	**44.76**	41.54	40.54	34.13	31.96	11.95	12.61
P.R.		**51.97**	44.98	39.03	33.58	13.50	14.21
DaKatz			**51.89**	39.55	37.57	14.07	13.69
Katz$_\beta$				**42.73**	29.48	14.71	14.10
Cocit					**43.25**	10.46	8.95
Cocoup						**16.37**	11.64
CCIDF							**16.98**

(ii)	DaRWR	P.R.	DaKatz	Katz$_\beta$	Cocit	Cocoup	CCIDF
DaRWR	**40.14**	32.15	30.86	30.25	10.02	14.78	17.05
P.R.		**34.91**	27.34	27.75	11.31	14.17	16.30
DaKatz			**35.31**	33.23	9.05	15.95	16.79
Katz$_\beta$				**34.51**	9.54	16.05	16.72
Cocit					**13.50**	5.92	5.58
Cocoup						**17.39**	13.43
CCIDF							**19.22**

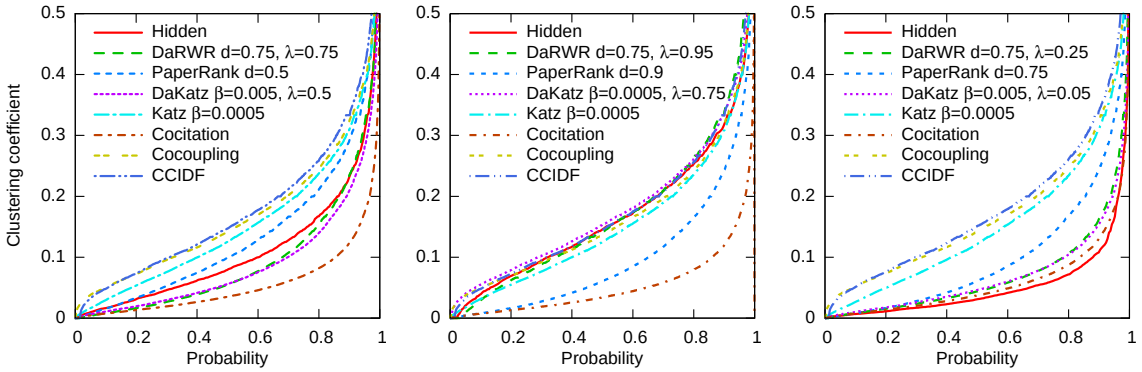
(iii)	DaRWR	P.R.	DaKatz	Katz$_\beta$	Cocit	Cocoup	CCIDF
DaRWR	**60.87**	52.39	56.56	40.10	47.31	2.17	2.306
P.R.		**57.99**	53.98	40.53	48.69	2.65	2.75
DaKatz			**63.84**	41.34	50.81	2.45	2.63
Katz$_\beta$				**42.09**	38.27	2.80	2.78
Cocit					**54.97**	2.43	2.16
Cocoup						**2.91**	2.04
CCIDF							**3.19**

papers having patterns resembling the properties of the papers it is designed to find. The *clustering coefficient* [20] C_v of a paper v can be used to qualify the citation patterns. It is computed as:

$$C_v = \frac{|\{(i,j) \in E \mid i,j \in N_v \cup \{v\}\}|}{|N_v| \times (|N_v|+1)},$$

where N_v is the set of neighbor papers of v which either cite v or are cited by v. Intuitively, clustering coefficient indicates how close of being a clique a vertex and its neighbors are.

Figure 5 presents the cumulative density function of the clustering coefficients of the documents suggested by each algorithm and of the hidden papers in three scenarios. First of all, the trace of the hidden papers is different in the three scenarios. When the hidden papers are early papers, they are typically well cited and their clustering coefficients are low. (It is unlikely that a large neighborhood forms a clique since the outgoing degree is typically small.) Recent papers have a higher clustering coefficient and their neighborhoods are small. This confirms that clustering coefficient can be used to distinguish old and recent papers.



(a) hide random (b) hide recent (c) hide earlier

Figure 5: Clustering coefficient of the suggested citations for the experiments.

None of the methods matches the trace of the hidden paper in the **hide random** scenario. PaperRank matches its trace at the beginning of the curve, while DaRWR and DaKatz match it at the end. In the **hide recent** scenario, most algorithms have a trace similar to the one of the hidden papers beside PaperRank and Cocitation. Notice that CCIDF and Cocoupling exhibit a trace similar to the hidden papers despite suffering from a low accuracy: they find recent papers but not the relevant ones. In **hide earlier** scenario, DaRWR, DaKatz and Cocitation have patterns similar to hidden papers. The rest of the algorithms have patterns different from the hidden papers. In all cases, CCIDF has citation patterns similar to the one of Cocoupling.

This analysis shows that direction-aware algorithms can be tuned to reach a variety of citation patterns, allowing them to match the patterns of recent or old documents. However, having a similar trace is an important property but it is not enough to reach a high precision.

4. CONCLUSION AND FUTURE WORK

In this paper, we present direction-aware algorithms for citation recommendation. They allow to tune the search for finding more recent or more traditional documents. We developed two algorithms based on the direction-aware model, namely DaKatz and DaRWR. In our experiments, direction-aware algorithms outperform the existing algorithms when the objective is to find either traditional or recent papers. We deployed one of the algorithms in our web service, the**advisor**, which allows any researcher to upload a bibliography file and obtain suggestions. We believe that our service will become a tool of major interest for researchers.

As future work, we are planning to weight edges differently based on how many times a paper cites an other. We will also improve the amount and quality of the bibliographic data, and conduct an intensive user study to obtain a real-world evaluation of the system.

Acknowledgments

This work was partially supported by the U.S. Department of Energy SciDAC Grant DE-FC02-06ER2775 and NSF grants CNS-0643969, OCI-0904809 and OCI-0904802. The authors also would like to thank DBLP, CiteSeer, arXiv and HAL-INRIA for making their data publicly available.

5. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of WWW*, 1998.

[2] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *Proc. of ACM Conf. Digital Libraries*, 1998.

[3] M. Gori and A. Pucci. Research paper recommender systems: A random-walk based approach. In *Proc. of IEEE/WIC/ACM Web Intelligence*, 2006.

[4] G. Jeh and J. Widom. Scaling personalized web search. In *Proc. of WWW*, 2003.

[5] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[6] N. Lao and W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81:53–67, 2010.

[7] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *Computer*, 32:67–71, 1999.

[8] M. Ley. DBLP - some lessons learned. *PVLDB*, 2(2):1493–1500, 2009.

[9] J. Li and P. Willett. Articlerank: a PageRank-based alternative to numbers of citations for analyzing citation networks. *Proc of ASLIB*, 61(6), 2009.

[10] Y. Liang, Q. Li, and T. Qian. Finding relevant papers based on citation relations. In *Proc. of WAIM*, 2011.

[11] D. Liben-Nowell and J. M. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–31, 2007.

[12] W. Lu, J. Janssen, E. Milios, N. Japkowicz, and Y. Zhang. Node similarity in the citation graph. *Knowl. Inf. Syst.*, 11:105–129, 2006.

[13] N. Ma, J. Guan, and Y. Zhao. Bringing pagerank to the citation analysis. *Inf. Process. Manage.*, 44:800–810, 2008.

[14] 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 *Proc. of ACM Computer Supported Cooperative Work*, 2002.

[15] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proc. of ACM KDD*, 2004.

[16] G. Salton. Associative document retrieval techniques using bibliographic information. *J. ACM*, 10:440–457, 1963.

[17] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.*, 24(4):265–269, 1973.

[18] T. Strohman, W. B. Croft, and D. Jensen. Recommending citations for academic papers. In *Proc. of SIGIR*, 2007.

[19] Y. Wang, E. Zhai, J. Hu, and Z. Chen. Claper: Recommend classical papers to beginners. In *Proc. of Fuzzy Systems and Knowledge Discovery*, 2010.

[20] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, pages 440–442, June 1998.