FutureRank: Ranking Scientific Articles by Predicting their Future PageRank

Hassan Sayyadi* Lise Getoor†

Abstract

The dynamic nature of citation networks makes the task of ranking scientific articles hard. Citation networks are continually evolving because articles obtain new citations every day. For ranking scientific articles, we can define the popularity or prestige of a paper based on the number of past citations at the user query time; however, we argue that what is most useful is the expected *future* references. We define a new measure, *FutureRank*, which is the expected future PageRank score based on citations that will be obtained in the future. In addition to making use of the citation network, FutureRank uses the authorship network and the publication time of the article in order to predict future citations. Our experiments compare FutureRank with existing approaches, and show that FutureRank is accurate and useful for finding and ranking publications.

1 Introduction

Ranking scientific publications is an extremely challenging task, due in part to the tremendous diversity of topics and disciplines, and in part due to the dynamic nature of the evolving scientific literature network. There is a long history of research in the bibliometrics which tries to measure the impact of a publication. Likewise there is much recent work in ranking of web pages, and many algorithms have been proposed, including the well-known PageRank algorithm [25] and hubs and authorities model [11]. While both of these approaches work well in the context for which they were designed, ranking based on the current citations and links, neither of these approaches attempt to model the dynamic and evolving nature of the networks. In fact, they were originally designed for static networks. Since only the current snapshot of the network is important for these models, crawling the network frequently and recomputing the scores by the models can help to overcome the dynamic nature of the networks. However, for citations network and ranking scientific articles, the previous, current, and future pictures of the network are all important.

In the context of researchers trying to retrieve relevant work to cite, or perform a literature review based upon which to build their research, the publication time of an article plays an important role and should be considered in the ranking. In order to retrieve relevant research, recent work is very important for finding new research directions, new solutions and approaches, and also finding the overlap of their work and other work. In order to do this, researchers need to find good articles published recently.

So the question that we are interested in answering is "*which article is the most useful article at the user query time and should obtain the highest score in a ranking model?*". One of the simplest and oldest approaches to compute the quality of a scientific article is counting the number of citations. Unfortunately, newly published articles do not have many citations to evaluate their quality. The popularity (number of citations) or prestige (PageRank) of a paper is defined based on the current citations at the user query time. But we differentiate current citations from *usefulness* which can be defined as the number of citations in the future. In other words, usefulness is based on the future popularity and prestige. Future popularity is well-defined (but unobserved), and we define future prestige by introducing a measure we refer to as the future PageRank score. The future PageRank score is the traditional PageRank score computed *based only on the citations that will be obtained in the future.*

While this definition of useful makes sense, obviously, it is also problematic in the sense that it is based on the information which is not available at the user query time. So, in order to make use of it, we need to make predictions about the citations which a paper will obtain in the future.

From a network perspective, articles can be seen as nodes in a network and citations show the directed edges among the nodes. Each article node links to another article node in the network if it cites the other article. Hence, the network will be a directed network and the problem is to rank nodes in this network. This is a traditional scientific article ranking problem, while there is more information, in which we can extend the network

*Department of Computer Science,University of Maryland-College Park,Email:sayyadi@cs.umd.edu

†Department of Computer Science,University of Maryland-College Park, Email: getoor@cs.umd.edu

to have a more complex network that can capture more information. As mentioned before, publication time is an important aspect of this network and we are dealing with a dynamic network.

In this work we extend the network by other available information such as authors and the publication time of the articles, and target to predict the future citation of the article in order to have better ranking model.

The rest of this paper is organized as follows: in the next section we review the related work. Section 3 describes *FutureRank*, our proposed model for ranking scientific articles. Then, section 4 shows the experimental results and evaluations. Finally in section 5, we conclude and describe future research directions.

2 Related Work

Ranking scientific articles is an important and challenging problem and there is a long tradition of work on the topic. One of the important early steps in this area was the work of Garfield [6] in 1970's. He proposed a measure for ranking journals and called it the *ImpactFactor*. It is calculated based on a three-year period. For example, the impact factor of year i for a journal j is calculated as follows:

$A =$ the number of times articles published in journal j in years $i-1$ and $i-2$ were cited in indexed journals during year i.

$B =$ the number of articles, reviews, proceedings or notes published in journal j years $i-1$ and $i-2$

$ImpactFactor(j,i) = A/B$. Thus, the impact factor is an approximation of the average number of citations within a year given to the set of articles in a journal published during the two preceding years. He also applied a similar idea of counting citations to evaluate scientists [7].

Based on this early work, many different versions of impact factors have been proposed [16, 26, 8, 12, 13, 14, 15, 21, 22]. However all of the approaches were based on counting citations, and the problem with counting citations is that it is only based on the popularity of the articles but not the prestige which is usually measured by scores similar to PageRank.

After the revolution in ranking web pages by PageRank [25] and its application to different domains ([2, 9, 5, 23, 10, 19, 20]), many researchers explored applying a PageRank approach to the citation network in order to rank scientific articles [3, 4, 18]. Their results confirmed that the *ImpactFactor* finds the popularity while PageRank score shows the prestige.

In addition to analysis of the citation network, research has looked at making use of the co-authorship network. For example, Liu et al. [17] also applied PageRank to the co-authorship network in order to rank scientists.

Zhou et al.[28] used the idea of mutual reinforcement between hubs and authorities introduced in HITS algorithm [11], and they made use of three networks: the citation network, the co-authorship network, and the authors' social network. In the authors' social network, two authors are linked if they published a paper together or attended the same conference together. They used the mutual reinforcement idea, but their model was simulating the PageRank random walk in the combination of these three networks. In fact, the Co-Rank model contains two independent random walks in citations network and authors' social network and a random walk between the two networks on the authorship network. If at any given moment the random walk is on the author side, then it can either make m intra-class steps or k inter-class steps. Similarly, if it is on the document side, then it can either make n intra-class step while another option is to make k inter-class steps:

$$M = \begin{bmatrix} (1-\lambda)(\widetilde{A}^T)^m & \lambda DA^T(AD^TDA^T)^k \\ \lambda AD^T(DA^TAD^T)^k & (1-\lambda)(\widetilde{D}^T)^n \end{bmatrix}$$

in which A denotes the adjacency matrix of authors' social network, D shows the adjacency matrix of document citations, and AD shows the adjacency matrix of authorship network from authors to documents (DA just shows the reverse direction of AD). Then, they concatenated the two vectors of ranks (vector a for articles and vector d for documents) into a vector v such that $v = [a^T, d^T]^T$. Finally, by solving $v = Mv^T$, the final scores are computed. The model provides a co-ranking of articles and authors, and was evaluated based on the author ranking. Nie et al. also consider articles as Web objects and collect Web information for object, then rank Web objects in terms of their popularity and relevance to the user query. [24].

A problem with all of the above approaches is that they rank articles based on the prior popularity or prestige, so that recent articles will always obtain lower scores. Walker et al [27] introduced CiteRank, a method which uses the publication time of the articles and defines a random walk to predict number of future citations and then uses this to rank the articles. CiteRank models the citation process in which researchers start their search from a recent paper or reviews and follow a chain of citations until satisfied. In the proposed model, the probability of jumping to an article is proportional to its publication time which is computed by:

$$\rho_i = e^{-age_i/T_{dir}}$$

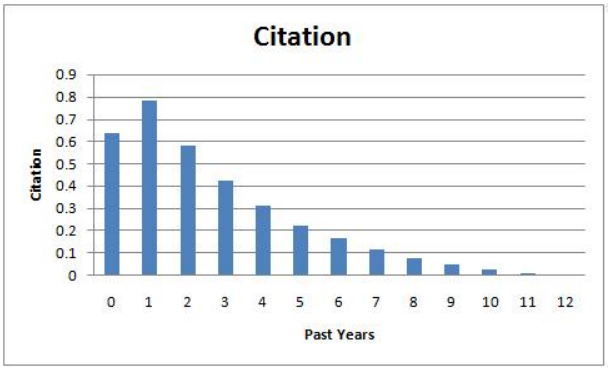

Figure 1: Average number of citations in the dataset(arXiv), which articles obtain, based on the number of years prior from publication date

in which age_i is the age of the i-th article. The CiteRank traffic of the paper is then defined as follows:

$$\overrightarrow{T} = I.\overrightarrow{\rho} + (1-\alpha)W.\overrightarrow{\rho} + (1-\alpha)^2W^2.\overrightarrow{\rho} + ...$$

which shows the probability of encountering an article via all possible paths (W is the adjacency matrix of the citation network). Their experiment shows that the best correlation between the predicted citation by CiteRank and the number of future citations was around 0.68.

3 Our Proposed Model

In this section, we describe our proposed ranking model for scientific articles and authors. As mentioned in introduction, we would like to rank papers based on predicted future citations, as this will help researchers find good articles more easily. In order to do this,we need some measures for defining a useful article. As mentioned in the related work section, one traditional measure is based on the popularity or the number of citations, but a better measure would be the PageRank score, or estimated prestige of the article.

Figure 1 shows, for arXiv, a collection of high energy physics publications[1], average number of citations to articles in the same year (past year = 0), previous year (past year = 1), and so on). As we can see, while the aggregate number of citation increases each year and the number of citations that an article obtains in each individual year decrease exponentially, most references are to articles published in the previous year. Any algorithm which does not take recency of publication date into account in ranking articles is not going to be able to capture this effect.

In addition, useful articles are often written by well-known researchers, and therefore, another source of useful information is the authorship network, from which author reputation and contribution can be extracted. Our model for ranking the scientific articles is based on the following assumptions:

- Important articles are cited by many important articles.
- Good research papers are written by researchers with high-reputations and researchers have high-reputation since they write good research articles. This illustrates mutual reinforcement between articles and the authors.
- Recently published articles are more useful, or in other words, they will obtain more citations in the future.
- Among old papers, recently cited articles are more useful.

There is one situation which at the first glance may not fit to our assumptions, and it is for the case of well-known classic papers which are not very recent but their citations do not decrease, because everybody cites them. Actually, these types of papers will still achieve good ranks based on our assumptions. While these papers are not recent, recent papers tend to have high ranks and many recent papers cite these classic papers. Then, in any authority transfer model, the citing papers will propagate their score to the referenced papers, so the classic papers will still have a high score because of their recent citations.

3.1 Network Structure

Figure 2 shows an abstract representation of a scientific publication network. The network has two types of nodes, papers and authors. In addition, there are two types of edges, *authorship edges*, which are undirected edges between papers and their authors, and *citation edges*, which are directed links from a paper to each of its references. In order to rank nodes in this network, we can see the network as a combination of two networks. The first network only contains the paper nodes and citation edges between them, so PageRank can be used here for authority transfer from articles to their references. The second network is the network of papers and authors but only contains authorship edges. It is obviously a bipartite network, and it can be mapped onto a HITS-type network where articles are *authorities* and authors are *hubs*, and the network can simulate the mutual reinforcement between articles and authors using a HITS-style propagation algorithm. Figure 3 shows the mapping.

Networks are often represented as sociograms or adjacency matrices, and we use a matrix representation

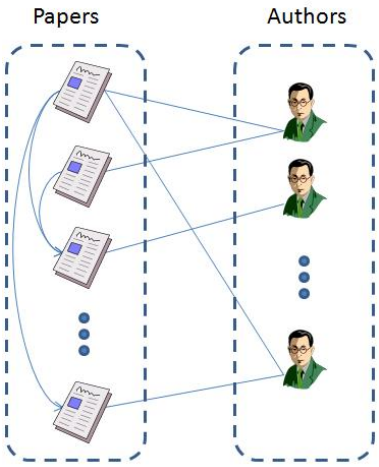

Figure 2: An example of the scientific article network. There are two types of nodes, papers (rectangles) and authors (circles) and two types of edges, authorship edges, between authors and papers, and citation links, between papers.

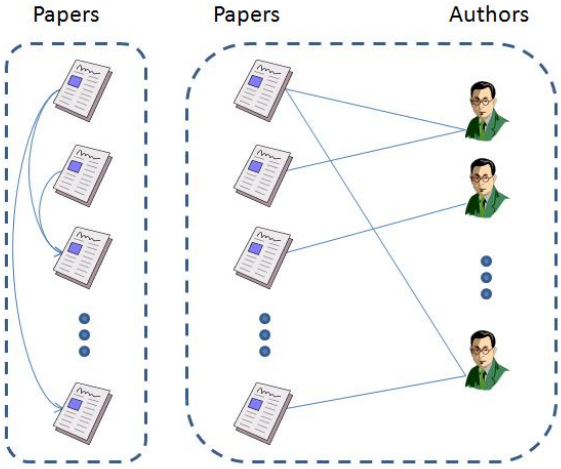

Figure 3: Network Decomposition: A single mode network of citations and a bipartite network of authorship

as well and store ranking scores in vectors. If P is the set of papers and A is the set of authors, the matrix M^C is the $|P| \times |P|$ citation matrix where

$$M^C_{i,j} = \begin{cases} 1 & \text{if } p_i \text{ cites } p_j; \\ 0 & \text{otherwise;} \end{cases}$$

Furthermore, for any paper p_i which does not cite any other article in the dataset, we define $M^C_{i,j} = 1$, for all j. Essentially, we create virtual links from dangling nodes to every other node.

In addition, we define the matrix M^A which is the $|P| \times |A|$ authorship matrix:

$$M^A_{i,j} = \begin{cases} 1 & \text{if } a_i \text{ is the author of } p_j; \\ 0 & \text{otherwise;} \end{cases}$$

3.2 Proposed Ranking Algorithm: FutureRank

Since two networks share nodes, we cannot compute rankings for each of the networks individually with different models. Instead, we propose a new ranking algorithm, which we refer to as *FutureRank*, which operates on both networks, passing information back and forth between the networks. The ranking algorithm is an iterative algorithm which runs one step of PageRank, one step of HITS and combines their results. It then repeats the above steps until convergence. We denote the vector of paper scores by R^p, and the vector of author ranks by R^a. The score of authors is computed by the following formula:

$$R^A = M^A * R^P$$

which is the hub score in the authorship network. In other words, articles transfer their authority score to their authors, and an author collects the authority score of all of his/her publications. However, the rank of papers is computed by a more complicated formula as follows:

$$\begin{aligned} R^P &= \alpha * M^C * R^C \\ &+ \beta * M^{A^T} * R^A \\ &+ \gamma * R^{Time} \\ &+ (1 - \alpha - \beta - \gamma) * [1/n]. \end{aligned}$$

This formula is the weighted sum of the following three scores:

- $M^C * R^C$ is the PageRank score in the citation network,
- $M^{A^T} * R^A$ is the authority score in the authorship network (M^{A^T} shows the transposition of M^A),
- The last term, R^{Time}, is a "personalized" PageRank vector. In the original PageRank model, the personalized vector is a pre-computed score vector to rank the results in favor of user preferences, where the default value is $\frac{1}{n}$ for all nodes (n is the number of nodes in the network). The values in our personalized vector are pre-computed based on the current or query time, $T_{current}$, the publication time of the papers, and favor recently published papers:

$$R^{Time}_i = e^{-\rho * (T_{current} - T_i)}$$

where T_i is publication time of p_i. $T_{current} - T_i$ shows the number of years prior from the publication time of p_i.

The initial value of R_i^P is $\frac{1}{|P|}$ and similarly the initial value of each R_i^A is $\frac{1}{|A|}$. This initialization keeps the sum of the paper ranks equal to 1, as well as the sum of author ranks. This property will hold after each iteration too, since the computation performs an authority propagation and sum of the weights, $\alpha + \beta + \gamma + (1 - \alpha - \beta - \gamma)$, is equal to one.

4 Evaluation

In this section, we evaluate several variants of our proposed FutureRank method on a collection of scientific papers. We compare with several alternate, previously proposed approaches, and evaluate according to several performance criteria. We conclude with a discussion of running times and convergence.

4.1 Data Set We evaluated our algorithms on a real dataset of scientific articles, the arXiv (hep-th) dataset[1]. The dataset contains articles published on high energy physics from 1993 to 2003. The dataset contains approximately 28,000 articles, with 350,000 citations. There are about 15,000 authors. We extracted authors from the description file for each article, and in the extraction phase, two authors are considered identical if their full names match; this is a strong matching criteria, and more sophisticated author name resolution strategies should improve our performance. There were a few citations from an article to another article published later; these were removed.

4.2 Evaluation Setup For evaluation, we split the dataset into two sets: the first set, the query data, contains all papers published before 2001 and the second part, evaluation data, contains all papers published in or after 2001. We used the first set in our experiments for computing the ranks, and then used the second partition as the future data for evaluating the ranking.

In order to construct the evaluation, we can view the system from the stand point of a user in 2001, who is searching for research papers. At this point, the only available information is the information in the first partition of the dataset. As the definition of usefulness, the most useful paper for that user should be the paper that will obtain the highest *future PageRank*, e.g., the highest number of citations after 2001. For the evaluation, we compute *future PageRank* as the PageRank scores of articles only based on the citations in and after 2001. In order to do this, we create a new network of all papers from 1993 to 2003, but only make use of the citations edges originating from the articles published in and after 2001.

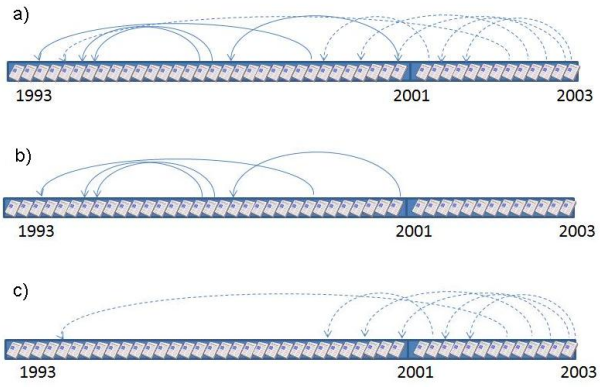

Figure 4: (a) Hypothetical figure of entire Dataset in which the horizontal line shows the time line and edges show the citations of older articles by earlier articles (b) The full dataset; the horizontal line shows the time line and edges show the citations of articles by later articles; (c) the query data, which only contains edges originating before 2001, and (c) the evaluation data, which only contains edges originating after 2001.

To illustrate the setup, figure 4(a) shows a hypothetical figure of a full dataset. The horizontal bar in the figure shows the time line, and each arrows show a citation from an article to another one. The figure shows the information available in the query dataset and the dashed arrows show the future citations which are available only in the evaluation data. The query set and evaluation set are also shown separately in Figure 4(b) and 4(c), respectively. Then, the PageRank in the evaluation network is computed. The PageRank scores of articles published prior 2001 are the true scores, which we are using as the evaluation data and call it *future PageRank*. The goal of *FutureRank* is to predict a ranking which is consistent with this score. However, the prediction should be done based only on the information available to at the user query time, which in this case is only the nodes and edges prior to 2001.

Furthermore, as discussed earlier, it is important to model the effect of publication time in the article ranking. For this, in section 3.2, we introduced the personalized vector R^{Time}. In order to find the best value of ρ for computing the value of R^{Time}, we find the best exponential trend line, which fits Figure 1. Ignoring the data for *past years*= 0, the best exponential function which matches the data in Figure 1 is

$$c * e^{-0.62 * x},$$

so the best value of ρ will be -0.62. Interestingly, this result confirms exactly the result of [27], which $T_{dir} = 1.6 = \frac{1}{0.62}$ is the best value. The authors of [27], ran CiteRank for all possible values of T_{dir}, then found $T_{dir} = 1.6$ years as the best value, while we simply find the best trend line fit to the dataset. Interestingly, both works obtain the same value, while we can find the best value simply by counting the citations in different years for each dataset *without* making use of the evaluation data, while they ran the ranking process and needed test data to find the value of the parameter which gave the best precision.

4.3 Ranking: Evaluation and Approaches For evaluating the ranking, we use two approaches: 1) the precision-recall curve, and 2) the Spearman's rank correlation between the rankings provided by the models and the future PageRank computed on the test data.

We compared and evaluated four different versions of *FutureRank* with previous works:

- **FutureRank**: Our proposed model which use all available information (author, citation and publication time):

$$\begin{aligned} R^P &= \alpha * M^C * R^C \\ &+ \beta * M^{A^T} * R^A \\ &+ \gamma * R^{Time} \\ &+ (1 - \alpha + \beta + \gamma) * [1/n] \end{aligned}$$

$$R^A = M^A * R^P$$

- **FutureRank(CT)**: A variant of our proposed model which only uses citation and publication time, but does not use the author information ($\beta = 0$).

$$\begin{aligned} R^P &= \alpha * M^C * R^C \\ &+ \gamma * R^{Time} \\ &+ (1 - \alpha - \gamma) * [1/n] \end{aligned}$$

$$R^A = M^A * R^P$$

  This is the same information that CiteRank[27] uses. As it is clear from the formula, this version is not using the authorship network and the mutual reinforcement between authors and publications.

- **FutureRank(CA)**: A variant of our proposed model which only uses citation and authorship information ($\gamma = 0$).

$$\begin{aligned} R^P &= \alpha * M^C * R^C \\ &+ \beta * M^{A^T} * R^A \\ &+ (1 - \alpha - \beta) * [1/n] \end{aligned}$$

$$R^A = M^A * R^P.$$

  This model do not use the publication time of the articles, so it uses the same information as CoRank[28].

- **PageRank**: which is the traditional PageRank Model for $\alpha = 0.9$, and random jump with probability of 0.1 ($\alpha = .9, \beta = 0$, and $\gamma = 0$).

$$\begin{aligned} R^P &= \alpha * M^C * R^C \\ &+ (1 - \alpha) * [1/n] \end{aligned}$$

$$R^A = M^A * R^P$$

Furthermore, in order to evaluate the algorithms for finding the precision and also computing the precision-recall curve, we need to construct ground truth. Ground truth, especially for ranking algorithms, is challenging. In order to evaluate, we took the top 50 papers sorted by future PageRank as the true result set. We measure the precision of our algorithm by comparing this top 50 with the top 50 of returned by our proposed algorithms. We compute the precision as:

$$Precision = \frac{|FutureRank_{top50} \cap Future\ PageRank_{top50}|}{50}$$

which is the precision of the top 50 of *FutureRank* results. Later, we also evaluate with different values for the size of this set.

4.4 Effect of Parameters on Precision We began by investigating the sensitivity of the performance of the *FutureRank* to different settings for the parameters which weight the citation (α), author (β) and publication time information (γ).

Figure 5 shows the precision of *FutureRank* for different values of α, β, and γ. The x-axis shows the value of γ and the vertical axis shows the value of α. Since $\alpha + \beta + \gamma$ is always equal to 1, at any point in the heatmap, the value of β is $1 - \alpha - \gamma$ (The top right triangle of the map is empty, because the sum of α,β, and γ cannot be more than 1). The lighter the color in the heatmap, the higher the precision.

Figure 5 shows the possible configurations of *FutureRank*. For example, the accuracies shown on each edge of the heatmap triangle show the combination of using only two type of information. All values on the

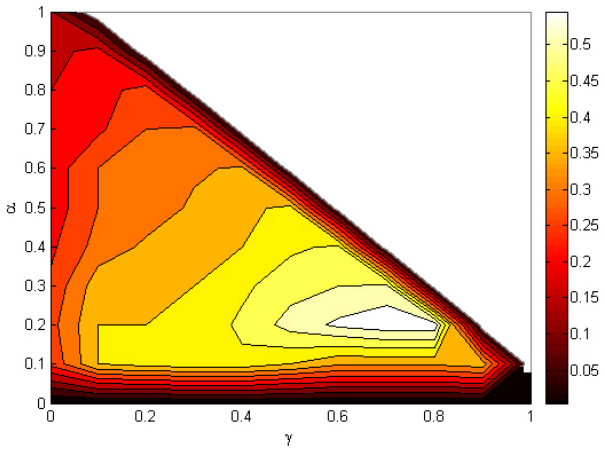

Figure 5: The precision of *FutureRank* for different settings of the three weighting parameters, α, β, and γ. In any point in the figure, the value of β is equal to $1 - \alpha - \gamma$.

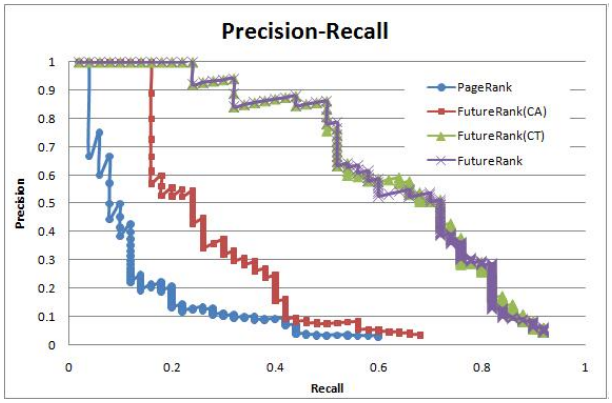

Figure 6: Precision-Recall based on precision in top 50 results.

horizontal edge obtained for $\gamma = 0$, so it means the horizontal edge shows all possible configurations of *FutureRank(CA)*,while the hypotenuse shows all possible configurations of *FutureRank(CT)*. Each corner also shows the precision of *FutureRank* which only use one type of information (on of α, β, and γ is equal to 1 and the two others are zero). The nice observation is that the space has a single optimal region, rather than a more complex collection of optimal configurations.

The highest precision of *FutureRank* is obtained at $\alpha = 0.19$, $\beta = 0.02$ and $\gamma = 0.79$. However, one should take care in interpreting these values. This combination of values does not mean the effect of time R^Time is three times more important than the citation. This happens because we take the weighted sum of three ranks and all of these ranks are between 0 and 1, but because the sum of all scores in each vector R^P, R^A, and R^{time} must sum to 1 individually, the ranges of scores in the different vectors are different. For example, the sum of author scores should be 1 as well as the sum of papers scores, but there are about twice as many papers as authors. This means that the average score of papers will be less than the average score of authors.

4.5 Further Comparison of Proposed Algorithms Next, we evaluated the effectiveness of the proposed algorithms in further detail. We begin by exploring the precision-recall trade-offs for each. Figure 6 shows the precision-recall curves of four models:

1. **FutureRank**: Here the best curve is obtained by $\alpha = 1.9$, $\beta = 0.02$, and $\gamma = 0.79$
2. **FutureRank(CT)**: Here the best curve is obtained by $\alpha = 0.2$,$\beta = 0$, and $\gamma = 0.8$.
3. **FutureRank(CA)**: Here the best curve is obtained by $\alpha = 0.2$, $\beta = 0.8$,, and $\gamma = 0$.
4. **PageRank**

As we can see from the diagram, *FutureRank* provides significant improvement in ranking scientific articles; the first 25% of retrieved articles by *FutureRank* are correct, which means the precision is 100% among them.

The focus of the CoRank algorithm [28] was on ranking authors, so authors of the paper do not present an evaluation for article rankings. *FutureRank(CA)* uses the same information as CoRank.

While the results in both figures 5 and 6 are based on the top 50 results, figure 7 shows the precision of *FutureRank* top k results for different values of k.

We were also interested in comparing more directly to CiteRank [27]. The authors of CiteRank did not report any precision-recall numbers but they computed the correlation coefficient between the number of citations in the evaluation set (in the future) and the citation traffic estimation by CiteRank. The best correlation between the predicted citation by CiteRank and the number of future citations was around 0.68, while the correlation between the *FutureRank* score and the future PageRank score is 0.83. This is encouraging, but these correlation values are not comparable since they are correlation of two different measures.

Hence, instead, we choose another measure for comparison, the Spearman's rank correlation between the rankings provided by two models and the ranking by future citations in the evaluation data. The CiteRank article showed the highest Spearman's rank correlation

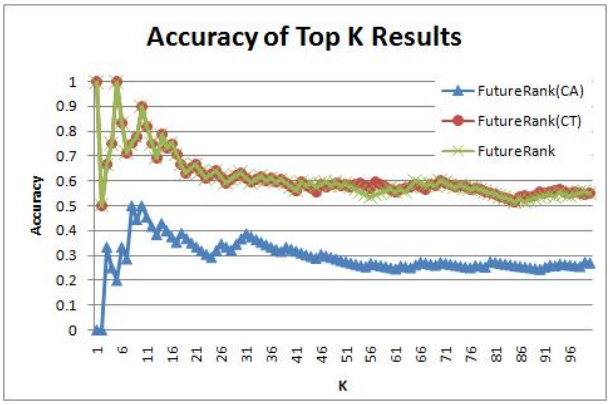

Figure 7: The precision of top k results of *FutureRank* comparing to the top k results of future PageRank

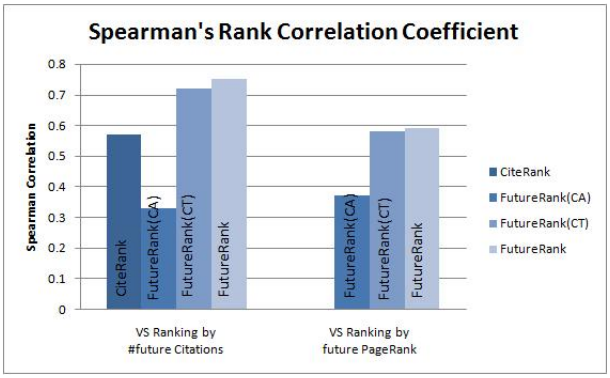

Figure 8: Correlation between the ranking on the validation data and the results of CiteRank and *FutureRank*

between the CiteRank ranking and the ranking by the number of citations in the future was 0.57, while the highest correlation between *FutureRank* and the ranking by the number of citations in the future is 0.75 which shows a significant improvement (In the evaluation data, we sorted articles by their publication date if two articles had the same number of citations in the evaluation data). We also computed the correlation between the *FutureRank* ranking and the ranking by the future PageRank, which is, in our minds, a more desirable measure. The obtained correlation is 0.59, which is still more than the correlation obtained by CiteRank for the simpler measure of the number of future citations.

The parameter settings for correlation values are shown in Figure 8 are as followings:

1. **FutureRank**: Here the best correlation is obtained by $\alpha = 0.4$, $\beta = 0.1$, and $\gamma = 0.5$
2. **FutureRank(CT)**: Here the best correlation is obtained by $\alpha = 0.5$,$\beta = 0$, and $\gamma = 0.5$.
3. **FutureRank(CA)**: Here the best correlation is obtained by $\alpha = 0.65$, $\beta = 0.35$,, and $\gamma = 0$.

The highest correlation between *FutureRank(CA)* and the number of future citation is 0.33 which confirms its precision recall curve. This shows the importance of publication time which approaches like CoRank do not use. Furthermore, the highest correlation between the number of citations and *FutureRank(CT)*, which only used the same information that CiteRank uses, is 0.72. In fact, *FutureRank(CT)* is a configuration of *FutureRank*($\beta = 0$) which runs the PageRank algorithm with a personalized vector to favor recently published articles. Even this simple model of FutureRank, obtains much higher correlation compared to the CiteRank algorithm. This also shows that in term of precision-recall, *FutureRank* will significantly outperform CiteRank, although the precision-recall curve of CiteRank was not available for comparison. In fact, we obtained the best precision-recall curve and the best correlation value with different parameter configurations for each model.

Figure 6 showed that the precision-recall curves of *FutureRank(CT)* and *FutureRank* are very similar and cross each other several times. Although the top results of both configurations of our model are very similar, the correlations obtained for *FutureRank* is slightly better than those of *FutureRank(CT)*.

The top 20 results of *FutureRank* within PageRank and *FutureRank(CA)* is shown in Table1, which can be compared to future PageRank, the "ground truth" ranking, on the evaluation data.

For *FutureRank* and future PageRank, the number of citations a paper obtained before 2001 (part of the available information to the *FutureRank*), the number of citations a paper obtain in and after 2001 (the future information which was not available to the *FutureRank* model), as well as the publication date are shown in that table. Since the top results of *FutureRank* and *FutureRank(CT)* are almost equal, results of *FutureRank(CT)* are not shown in the Table1. A case in point, while article 9906064, "*An Alternative to Compactification*", has only 414 citations before 2001, it obtained a good rank (3) by *FutureRank*. Considering that it published recently on 1999 and obtained this number of citations in less than two years, it confirms that the paper will obtain many citations in the future and is a very useful paper at the user query time (in 2001). Both the number of citations (617 citations) which this paper obtain after 2001 and the fact that it obtains the second position in the ranking by future PageRank confirm the accuracy of *FutureRank*.

Table 1: Top 20 articles retrieved by *FutureRank*, which are published before 2001

arXiv ID	Title	Publication Date	#Citations before 2001	#Citations after 2001	PageRank	FutureRank(CA)	FutureRank	Future PageRank
9711200	The Large N Limit of Superconformal Field Theories and Supergravity	11/28/1997	1540	874	10	3	1	1
9802150	Anti De Sitter Space And Holography	2/23/1998	1137	638	28	6	2	4
9906064	An Alternative to Compactification	6/9/1999	414	617	92	129	3	2
9802109	Gauge Theory Correlators from Non-Critical String Theory	2/17/1998	1054	587	37	8	4	5
9908142	String Theory and Noncommutative Geometry	8/23/1999	471	673	131	26	5	3
9407087	Monopole Condensation	7/19/1994	1082	217	1	1	6	10
9610043	M Theory As A Matrix Model: A Conjecture	10/8/1996	922	277	14	7	7	8
9510017	Dirichlet-Branes and Ramond-Ramond Charges	10/5/1995	937	218	4	5	8	14
9711162	Noncommutative Geometry and Matrix Theory: Compactification on Tori	11/21/1997	449	339	106	91	9	7
9905111	Large N Field Theories	5/17/1999	352	455	174	80	10	6
9503124	String Theory Dynamics In Various Dimensions	3/21/1995	981	133	2	4	11	46
9408099	Monopoles	8/19/1994	856	150	6	2	12	25
9510135	Bound States Of Strings And p-Branes	10/19/1995	675	100	13	9	13	77
9510209	Heterotic and Type I String Dynamics from Eleven Dimensions	10/30/1995	546	242	40	18	14	9
9611050	TASI Lectures on D-Branes	11/11/1996	594	107	76	23	15	29
9409089	The World as a Hologram	9/20/1994	266	161	95	31	16	15
9711165	D-branes and the Noncommutative Torus	11/24/1997	297	159	211	108	17	33
9204099	The Black Hole in Three Dimensional Space Time	5/8/1992	291	89	69	37	18	62
9410167	Unity of Superstring Dualities	10/25/1994	672	76	5	12	19	94
9603142	Eleven-Dimensional Supergravity on a Manifold with Boundary	3/22/1996	314	180	167	70	20	17

Table 2: Top 20 authors retrieved by *FutureRank* and the number of their articles in the dataset

Rank	Name	# of Publications
1	Edward Witten	100
2	Ashoke Sen	89
3	A.A. Tseytlin	111
4	Zurab Kakushadze	63
5	Joseph Polchinski	47
6	Juan M. Maldacena	21
7	Donam Youm	55
8	Nathan Seiberg	45
9	Cumrun Vafa	78
10	John H. Schwarz	47
11	Michael R. Douglas	52
12	Andrew Strominger	65
13	Nathan Berkovits	59
14	P.K. Townsend	56
15	Sergei V. Ketov	51
16	Miao Li	52
17	C.N. Pope	129
18	Shinichi Nojiri	94
19	N. Seiberg	23
20	Ichiro Oda	37

Finally, the top 20 authors retrieved by *FutureRank* and the number of their publication are listed in Table2.

4.6 Running Time and Convergence We ran our experiments on a machine with *Intel(R) Core(TM)2 Duo T7300 2GHz* CPU and 2 GB memory. We tested for convergences of our algorithms if the difference between the computed scores between two consecutive iterations was less than some threshold *minDifference*. The difference of two consecutive steps' scores is computed as follow:

$$Difference = \sum_{p_j \in P} (R_j^{P^i} - R_j^{P^{i-1}})^2 + \sum_{a_j \in A} (R_j^{A^i} - R_j^{A^{i-1}})^2$$

While the convergence rate was different for difference values of the model parameters, the model converges very fast in most cases. Figure 9 shows the difference between the score values in two consecutive steps. *FutureRank* and *FutureRank(CT)* have very similar behavior and apparently, both of them converge much faster than *FutureRank(CA)*.

We also show the top 50 results precision in Figure 10. It shows the precision converges after three iterations and there is no change in the top results, which means even four iterations are enough, while the results reported in CiteRank article were obtained after 20 iterations. Hence, in addition to better precision and correlations, in terms of running time, *FutureRank* is much faster than the CiteRank.

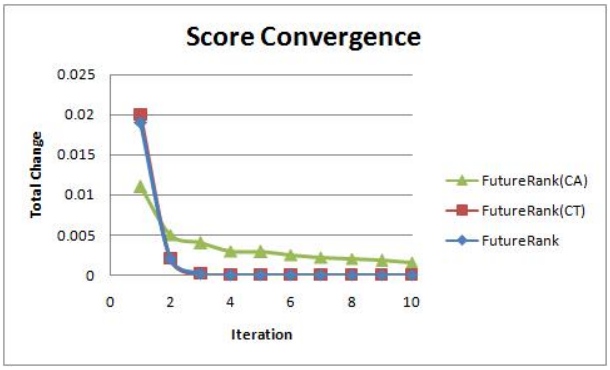

Figure 9: Score Convergence. The vertical axis shows the difference between the score values in two consecutive steps, so zero difference shows the convergence

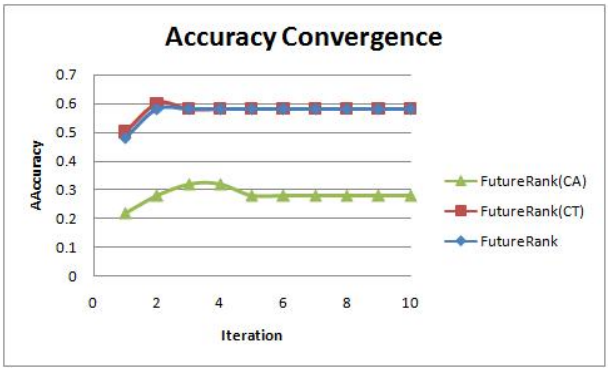

Figure 10: Precision of top 50 results after each iteration

5 Conclusion and Future Work

In this paper, we presented the *FutureRank* algorithm which is able to combine information about citations, authors and publication time to effectively rank scientific articles by predicting their future ranking of a paper. While the goodness of a paper at any time can be measured by the number of the citations that it has obtained, the number of the citations that a paper will obtain in the future measures how useful the paper is at the user query time and is a better criteria to use for retrieving articles that will help researchers in their work. Our experimental evaluation has shown the precision of *FutureRank*, and *FutureRank* achieves significant improvement over other recently proposed algorithms.

The precision-recall curve showed that using publication time information significantly outperforms the traditional PageRank and *FutureRank(CA)*. We also get a much higher correlation score, in addition to faster model convergence, and a practical solution to estimate the parameter model as compared to the CiteRank.

For future work, we plan to test the *FutureRank* on additional datasets in order to determine how robust it is to the different values of parameters α, β, and γ in different datasets. We also plan to investigate extending the network by adding additional node types, for example information about venues, such as conferences and journals. Furthermore, the *FutureRank* model is a general model which can be applied on different problems in the different areas. For example, the news domain is an interesting domain which has similar network structure. We can create a bipartite network of news articles and named-entities. The mutual reinforcement between news articles and named-entities mentioned is like the mutual reinforcement between scientific articles and authors. In addition, the number of related news items shows the importance of a news articles which is similar to the effect of citation in scientific article ranking.

Acknowledgments

This work was funded in part by NSF Grants No. 0746930 and 0430915.

References

[1] www.cs.cornell.edu/projects/kddcup/datasets.html.

[2] K. Berberich, S. Bedathur, M. Vazirgiannis, and G. Weikum. *Buzzrank ... and the trend is your friend.* In WWW06: Proceedings of the 15th international conference on World Wide Web, pages 937–938, New York, NY, USA, 2006. ACM.

[3] J. Bollen, M. Rodriquez, and H. V. de Sompel. *Journal status.* Scientometrics, 69(3):669–687, 2006.

[4] P. Chen, H. Xie, S. Maslov, and S. Redner. *Finding scientific gems with Google.* Journal of Informetrics 1, 2007.

[5] N. Eiron, K. S. McCurley, and J. A. Tomlin. *Ranking the web frontier.* In WWW04: Proceedings of the 13th international conference on World Wide Web, pages 309–318, New York, NY, USA, 2004. ACM.

[6] E. Garfield. *Citation analysis as a tool in journal evaluation.* Science, 178:471–479, 1972.

[7] E. Garfield. *How to use citation analysis for faculty evaluations and when is it relevant?(part 1).* Current Contents, pages 5–13, 1983.

[8] W. Glnzel. *The need for standards in bibliometric research and technology.* Scientometrics, 35:167–176, 1996.

[9] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. *Combating web spam with trustrank.* In VLDB04: Proceedings of the Thirtieth international conference on Very Large Data Bases, pages 576–587, 2004.

[10] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. *The eigentrust algorithm for reputation management in p2p networks.* In WWW03: Proceedings of the 12th international conference on World Wide Web, pages 640–651, New York, NY, USA, 2003. ACM Press.

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

[12] M. Koenig. *Determinants of expert judgement of research performance.* Scientometrics, 4:361–378, 1982.

[13] M. E. D. Koenig. *Bibliometric indicators versus expert opinion in assessing research performance.* Journal of the American Society for Information Science, 34:136–145, 1983.

[14] R. Kostoff. *Performance measures for government-sponsored research: Overview and background.* Scientometrics, v36, 281-292.

[15] Lawani and Bayer. *Validity of citation criteria for assessing the influence of scientific publications: New evidence with peer assessment.* Journal of the American Society for Information Science, 34:59–66, 1983.

[16] S. Lehmann, A. D. Jackson, and B. E. Lautrup. *Measures and mismeasures of scientific quality.* 2005.

[17] X. Liu, J. Bollen, M. L. Nelson, and H. V. de Sompel. *Co-authorship networks in the digital library research community.* Inf. Process. Manage., 41(6):1462–1480, 2005.

[18] N. Ma, J. Guan, and Y. Zhao. *Bringing pagerank to the citation analysis.* Inf. Process. Manage., 44(2):800–810, 2008.

[19] R. Mihalcea and P. Tarau. *TextRank: Bringing order into texts.* In Proceedings of EMNLP04and the 2004 Conference on Empirical Methods in Natural Language Processing, 2004.

[20] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert. *Generank: Using search engine technology for the analysis of microarray experiments.* BMC Bioinformatics, 6:233, 2005.

[21] Narin. *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity.* Computer Horizons, 1976.

[22] Narin and Hamilton. *Bibliometric performance measures.* Scientometrics, 36, 1996.

[23] B. Neate, W. Irwin, and N. Churcher. *Coderank: A new family of software metrics.* In ASWEC06: Proceedings of the Australian Software Engineering Conference, pages 369–378, Washington, DC, USA, 2006. IEEE Computer Society.

[24] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. *Object-level ranking: bringing order to web objects.* In WWW '05: Proceedings of the 14th international conference on World Wide Web, pages 567–574, New York, NY, USA, 2005. ACM.

[25] L. Page and et al. *The pagerank citation ranking: Bringing order to the web.* Technical report, Stanford Digital Library Technologies Project, 1998.

[26] F. Pinski, Gabriel; Narin. *Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics.* Information Processing and Management, pages 297–312, 1976.

[27] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. *Ranking scientific publications using a simple model of network traffic.* CoRR, abs/physics/0612122, 2006.

[28] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles. *Co-ranking authors and documents in a heterogeneous network.* 7th IEEE International Conference on Data Mining (ICDM07), 2007.