# Citation graph based ranking in Invenio

Ludmila Marian[1], Jean-Yves Le Meur[1], Martin Rajman[2], and Martin Vesely[2]

[1] European Organization for Nuclear Research
CH-1211 Geneve 23, Switzerland
{ludmila.marian, jean-Yves.le.meur}@cern.ch
[2] Ecole Polytechnique Fédérale de Lausanne
LIA, Station 14, CH-1015 Lausanne, Switzerland
{martin.rajman, martin.vesely}@epfl.ch

**Abstract.** Invenio is the web-based integrated digital library system developed at CERN. Within this framework, we present four types of ranking models based on the citation graph that complement the simple approach based on citation counts: time-dependent citation counts, a relevancy ranking which extends the PageRank model, a time-dependent ranking which combines the freshness of citations with PageRank and a ranking that takes into consideration the external citations. We present our analysis and results obtained on two main data sets: Inspire and CERN Document Server. Our main contributions are: (i) a study of the currently available ranking methods based on the citation graph; (ii) the development of new ranking methods that correct some of the identified limitations of the current methods such as treating all citations of equal importance, not taking time into account or considering the citation graph complete; (iii) a detailed study of the key parameters for these ranking methods.

**Key words:** CDS, Invenio, Inspire, citation graph, PageRank, external citations, time decay



## 1 Introduction

Invenio is the integrated digital library system developed at CERN [4], suitable for middle-to-large scale digital repositories (100K-10M records). It is a suite of applications which provides the framework and tools for building and managing an autonomous digital library server. Besides being used to run the CERN Document Server (which is ranked $4^{th}$ in the *Webometrics Top 400* institutional repositories [3]), Invenio has also been chosen by several other important institutions and projects. Among them, the recently launched INSPIRE service, that is meant to become the reference repository for High Energy Physics documents.At CERN, Invenio manages over 500 collections of data, consisting of over 1M bibliographic records [1].
In the setting of this framework, our goal is to develop robust citation ranking methods. We start our analysis from existing citation ranking methods, studying their strengths and their weaknesses. We do an in-depth analysis of the set of

parameters that influence the outcome. We develop novel citation ranking methods in order to overcome the identified drawbacks of the existing ones.
We use as a baseline the Citation Count. This method fails in capturing the differences between citations in importance as well as in publication date. In order to take into account the importance of the citations we study link-based ranking methods (Subsection 4.2). The idea of applying link-based methods to the citations graph is not new (Subsection 2). We found it relevant to re-evaluate the outcome as well as the parameter analysis in the context of our data sets, and using different metrics. By doing this, we discovered several drawbacks generated by different properties of the citation graph (connectivity, completeness and correctness). We correct these drawbacks by developing a novel link-based ranking that accounts for external citations (Section 4.4). In order to take into account the publication date of each citation, i.e. the "freshness" of the citations, we study time-dependent citation ranking methods (Subsections 4.1 and 4.3). Although the decayed time factor was also introduced previously in the literature, our contribution is firstly, applying this method on top of citation counts, introducing the notion of decayed citation counts, and doing an in-depth analysis of the stability of the rankings with respect to the decay factor, and secondly, analyzing time-decayed link-based ranking in the context of our data sets. This lead to the discovery of cycle-induced anomalies, that proven this method unsuited for time inconsistent data sets. These methods bring major improvements over the citation count baseline: by considering the importance of the citations, we can identify modestly cited publications that have a high scientific impact on the research community; on the other hand, by taking into consideration the publication date of each citations, i.e. the "freshness" of the citations, we can identify currently relevant publications, or better said, the "hot trends" of a specific domain that would have not been identified by the citation count method.

## 2 Related Work

In this section we review some of the work that has been conducted in the domains of citation analysis and ranking scientific publications.
In different cases, the citation count is not able to fully capture the importance of a publication, mainly due to the fact that it treats all the citations equally, disregarding their differences in importance and also their creation date. In order to overcome these drawbacks, several studies had been done. P. Chen et al. in [8] apply the Google's PageRank algorithm (proposed by S. Brin, L. Page in [13]) on the citation graph to assess the relative importance of all publications in the Physical Review family of journals from 1893-2003. They prove with different examples that applying PageRank is better at finding important publications then the simple citation count. They also argue about using a different damping factor than the one used in the original PageRank algorithm. The authors extended their work in [5] by introducing a new algorithm, CiteRank, a modification of PageRank, that also accounts for the date of the citations by distributing the

random surfers exponentially with age, in favor of more recent publications. By this, they try to model the behavior of researchers in search for new information. They test their model on all American Physical Society publications and the set of high-energy physics theory (hep-th) preprints. They find the parameters for their model by trying to maximize the correlation between the CiteRank output and the download history. Also, N. Ma et al., in [12] apply PageRank on the citation graph in order to evaluate the research influence of several countries in the Biochemistry and Molecular Biology fields.
There has been some research activity also in the area of "temporal link analysis", mostly done on WWW pages. In [6] the authors present several aspects and uses of the time dimension in the context of Web IR. K. Berberich et al. [7] argue that the freshness of web content and link structure is a factor that needs to be taken into account in link analysis when computing the importance of a page. They provide a time-aware ranking method and through experiments they conclude on the improvements broth by it to the quality of ranking web pages. They test their approach on the DBLP data set but with the scope of ranking researchers rather then publications.
The task of ranking scientific documents is a complex one and it should not depend only on the citation graph information. In the Invenio framework, there has been significant work done in trying to aggregate different metrics (i.e. the download frequency, the publication date) in order to create a better suited ranking for scientific documents [11], [10].

## 3 Experimental Framework

The experiments were conducted on two data sets of bibliographic data (not completely disjoint): Inspire (`http://hep-inspire.net`) containing 500,000 High Energy Physics (HEP) documents and CERN Document Server (`http://cdsweb.cern.ch`) containing 200,000 CERN documents.
We analyzed three important characteristics of the citation graphs extracted from these data sets: *graph connectivity* (i.e. the number of publications that have no citations, the number of publications that have no references), *graph completeness* (i.e. the number of publications missing from the data set) and *graph correctness* (i.e. if the graph allows cycles). The first two characteristics will be discussed per data set basis, while the third, since it is common for both citation graphs, will be discussed separately.

**Inspire Data Set.** Inspire is a new High Energy Physics information system which will integrate present databases and repositories to host the entire corpus of the HEP literature and become the reference HEP scientific information platform worldwide. It is a common project between CERN, DESY, FERMILAB and SLAC [2]. The Inspire data set contains almost half a million publications, with a total number of 8 million citations. Approximately 25% of the documents are not cited by any other document in the system, while approximately 16%

of the documents have no references. On average, a paper is missing 9 references. We computed this number as the difference between the total number of references displayed for a record and the number of references existing in the database. This 9 missing references/paper, compared with the average number of references that are in the system, 20 references/paper, tell us that although we do not have a complete citation graph, having more than 50% of the references is still better than expected. Also, Inspire is a human edited repository, meaning that the citation extraction is validated by an authorized person.

**CERN Document Server Data Set.** CDS contains the CERN collection of publications [1]. Out of more than 900,000 bibliographic records indexed by CDS we sampled a subset of 200,000 documents with 1,4 million citations. Approximately 20% of these documents are not cited by any other document in the system while 35% of the documents have no references. On average, each document is missing 28 out of 37 references. One reason for this low number of available references is that currently CDS is using an automated references extractor [9]. Since the future of bibliographic repositories is the automation of the data extraction, one must consider these drawbacks in the development and analysis of the ranking methods based on the citation graph. So, since the Inspire data set generates a better citation graph then the CDS data set, in terms of completeness, we will mainly discuss our results on the Inspire data set, but we will also present solutions for less dense data sets.

**Data Correctness** While the intuition is that the citation graph is a directed acyclic graph (DAG), we discovered that this is not true. Since the system contains preprints (drafts of scientific papers that have not yet been published in a peer-reviewed scientific journal) as well as published papers and conference proceedings, it might happen in some cases that future work is cited. On top of this, there are also some cases where a paper is citing itself. We try to eliminate these last types of anomalies as often as possible. Still, the first class of problems is harder to permanently eliminate, and even though theoretically impossible, the "future work" citation is sometimes legitime. For these reasons, we build our algorithms on top of a general directed graph and not on top of DAG.

## 4 Ranking Methods

In this section we study four types of citation ranking algorithms with respect to the baseline algorithm, Citation Count. All algorithms and parameters have been studied in the context of both data sets. We chose to present the results obtained on the Inspire database, since both connectivity and completeness parameters where higher in this case, thus facilitating the evaluation of the outcome. The only exception is the link-based ranking with external citations (Subsection 4.4), developed in particular for data sets with low completeness of graph (in our case, the CDS data set).

Our goal is to develop robust ranking methods based on the citation graph. In order to achieve this, we start from the citation count method, which we consider the baseline algorithm. We gradually add features and study both their positive and their negative impact on the final outcome. The result is four ranking methods, each suited for different types of publication discovery.
We first study the effect of time in the citation graph by applying a time-decay factor to the citation counts. In this context, we study the rank stability with respect to various settings of the time parameter. Since this method does not take into consideration the importance of different citations, we continue our analysis with a link-based ranking. Here we study the correlation between the damping factor and the bias towards older publications. In this case, our goal is to retrieve publications that are and have been of great interest for the community, although, they are modestly cited. In order to take into consideration both the age of citations and their importance, we combine the decay factor with the link-based ranking. The idea behind this method is that it is able to retrieve modestly cited papers that are at the present time of interest for their community. Unfortunately, this method suffers from cycle-induced anomalies. Last, but not the least, in order to overcome the bias of PageRank to incomplete citation graphs we introduce a novel link-based method.

### 4.1 Time-dependent Citation Count

To overcome the fact that the Citation Count method does not take into account the time dynamics of the citation graph we introduce the notion of *time-dependent citation counts*. In this context, the weight of a publication $i$ is defined as: $weight_i = \sum_{j,j \to i} e^{-w(t_{present} - t_j)}$ where $t_{present}$ is the present time and $t_j$ is the publication date for document $j^{th}$.
Furthermore, this introduces the time decay parameter ($w \in (0, 1]$), which quantifies the notions of "new" and "old" citations (i.e. publications with ages less than the time decay parameter would be considered "new"; publications with ages larger than the time decay parameter would be considered "old"). The larger the time decay parameter is, the faster we "forget" old citations.

**Results.** Since the time decay factor, $w$, is the only quantifier for the "freshness" of the results, we analyzed its impact on the stability of the final rankings and also on the stability of certain ranges of ranks.
In order to find out if the adding a time decay has a global impact on the ranking (i.e. the tail is promoted to the head and the other way around) or if it is rather local (i.e. there are certain windows in the ranking where there is some reshuffling) we measured the "*locality of changes*".
Let us consider $s$ as being the *stability factor*: $s = \frac{|\{d | rank_d(t), rank_d \in window\}|}{windowSize}$ where $rank_d(t), rank_d$ are the ranks of publication $d$, the first when using a time-dependent ranking method and the last when using the non-decayed ranking method.
Using the stability factor we want to determine what windows of the ranking

are suffering the most from different time decay parameters. For this we are building dynamic windows as follows: we are splitting the rank range by consecutive powers of 2, until we either reach a rank window of size less than 100 or the stability factor goes below a certain minimum threshold (0.3 in our experiments). We should mention that we remove the publications with 0 citations (125k documents), since their weight and rank will not be influenced by any ranking method. We constructed a chart for each value of the time decay factor (1 year, 2 years, 5 years, 10 years, 20 years, 40 years) (Figure 1). The interpretation of these charts is that whenever we have a zone with a lot of activity (a lot of points), that zone is quite stable at a high level and needs to be broken into small intervals to reach the instability threshold. On the other hand, when we have a zone with low activity, that means that the stability of the corresponding window is low also at a high level, so if we would split it in smaller windows, the stability will drop even lower then the threshold. From the Figure 1 we observe that the head of the ranks is usually more stable than the rest. Also, even with such a large time decay as 40 years, the ranks are still reshuffled, but in small windows.

We also analyzed the effects of different values of the time decay factor on

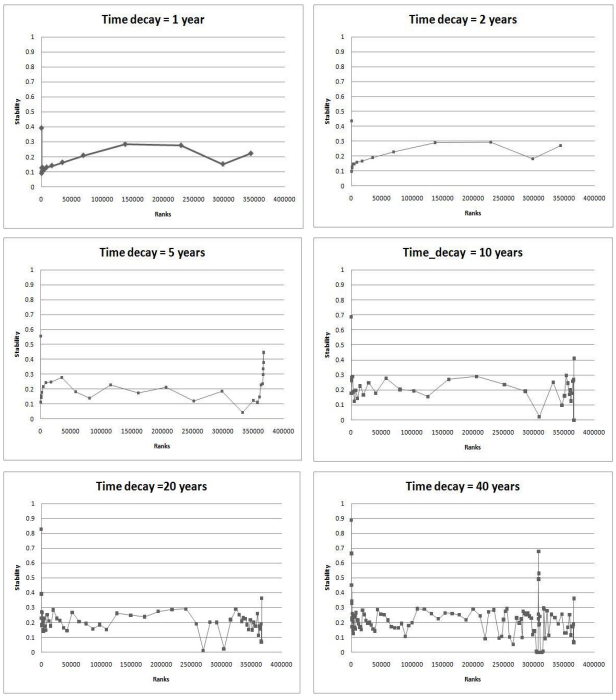


**Fig. 1.** Stability of Time-dependent Citation Count

the number of publications promoted/demoted. We discovered that that the *time-depending ranking methods are promoting more publications than demoting*. Secondly, and as a consequence of the first observation, the *time-depending ranking methods are demoting strongly than promoting*.
Based on this analysis one can choose between either having a strong time decay, which will boost really new publications, or having a weaker time decay, which will still boost publications with newer citations, but will also take into account old citations.
Adding even a weak time decay factor, the time-dependent ranking can still differentiate between an old publication that acquired a large number of citations over a long period of time, and a new publication, that, although important for the scientific community, did not have enough time to acquire as many citations as the old one, in the favor of the latter. Still, this method inherits one major shortcoming from the Citation Count method, i.e. it does not take into consideration the different importance of each citation. To overcome this, we developed the Time-dependent Link-based Ranking as a combination of Time-dependent Citation Counts and Link-based Ranking (Subsection 4.3).

### 4.2 Link-based Ranking: PageRank on the Citation Graph

The PageRank algorithm [13] is based on a random surfer model, and may be viewed as a stationary distribution of a Markov chain.
The PageRank model assigns weight to documents proportional with the importance of the documents that link to them:

$$PR(p_i) = \frac{1-d}{n} + d \sum_{j, p_j \to p_i} \frac{PR(p_j)}{deg(j)} \tag{1}$$

where $PR(p_i)$ is the PageRank score of paper $i$ and $deg(j)$ is the out-degree of node $j$ (i.e. total number of documents cited by paper $j$). $d$ is called *damping factor* and in the literature concerning the web graph it usually has values in $[0.85, 1)$. It is a free parameter that controls the performance of the PageRank algorithm, preventing the overweighing of older publications. This raking models the behavior of a user moving from paper to paper in the document collection [8]. At each moment in time the user can either follow a randomly chosen reference from the current document, with the probability $d$, or he can restart the search, from a uniformly randomly chosen publication with a probability of $1 - d$. For the WWW it is considered that on average, the users follow 6 continuous links, until they get bored and restart the search. In [8] the authors consider that a researcher will only follow on average 2 links on the citation graph, until the search is restarted. This is why they propose a damping factor of 0.5. In order to verify their hypothesis, we tested three different values for the damping factor: 0.50, 0.70, 0.85.
**Results.** The calculation of the Spearman's rank correlation coefficients generated with the three chosen values for the damping factor, showed us that, at the global scale, the differences between rankings are almost undetectable (the

lowest correlation, with a value of 0.996 was between d=0.50 and d=0.85). So in order to choose the best $d$ we have to dig dipper. For this, we looked at the distribution of the ranks over the time. Since we know that a higher damping factor is boosting old papers rather than new ones, we are interested to see if we can detect this kind of behavior also for our data. For this, we plotted the distribution in time for the Top 100 papers ranked with PageRank. The results are displayed in Figure 2. Indeed, we can see that for a damping factor of 0.5, the age of the top 100 papers decreases. If for a damping factor of 0.85 we have a large concentration of top papers in the 1970-1980 period, when decreasing the $d$, we see a shifting of the top papers towards the 1990-2000 period. Since we wish to have a ranking that is not biased towards older publications, we also conclude that a value of 0.5 for the damping factor is better suited.

The main advantage of this ranking method is that it weighs each publication

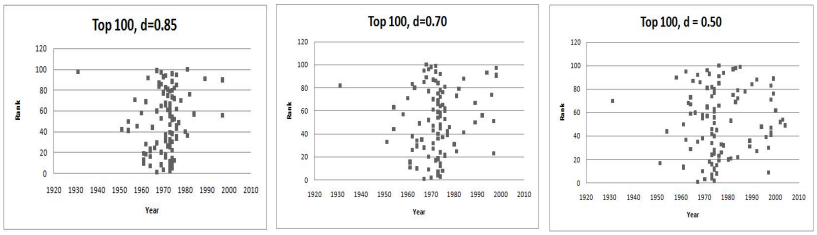


**Fig. 2.** Time distribution for the top 100 publication, ranked with PageRank

based on the importance of its citations. In this way, the quality is preferred over the quantity. We can say that it associates to each publication an "all-time achievement" rank (Table 1).

### 4.3 Time-dependent Link-based Ranking

The idea of the Time-dependent Link-based Ranking method is to distribute the random surfers exponentially with age, in favor of more recent publications. Every researcher, independently, is assumed to start his/her search from a recent paper or review and to subsequently follow a chain of citations until satisfied. In this way the effect of a recent citation to a paper is greater than that of an older citation to the same paper. This method was also presented in [5].

We consider the weight of each publication as being inversely proportional with its age: the younger the publication is, the more its citations will value. In this case, the initial probability of selecting the $i^{th}$ paper in a citation graph will be given by: $p_i = e^{-w(t-t_i)}$, where $t$ is the present time, $t_i$ is the publication date for document $i^{th}$ and $w$ is what we call the *time decay parameter*.

Adding the time decay to equation (1), we obtain:

$$PR(i,t) = \sum_{x=1}^{n} \left( \frac{1-d}{n} \times p_x(t) \right) + d \sum_{j,j \to i} \left( \frac{PR(j)}{deg(j)} \times p_j(t) \right)$$

| CC | Publication | Rank | RCC |
|---|---|---|---|
| 6565 | A Model of Leptons: Weinberg, Steven (1967) | 1 | 1 |
| 3023 | Confinement of Quarks: Wilson, Kenneth G., (1974) | 2 | 26 |
| 3671 | Weak Interactions with Lepton-Hadron Symmetry: Glashow, S.L., (1970) | 3 | 9 |
| 5351 | CP Violation in the Renormalizable Theory of Weak Interaction: Kobayashi, Makoto, (1973) | 4 | 2 |
| 2379 | Ultraviolet Behavior of Nonabelian Gauge Theories: Gross, D.J., (1973) | 5 | 44 |
| 2472 | Radiative Corrections as the Origin of Spontaneous Symmetry Breaking: Coleman, Sidney R., (1973) | 6 | 40 |
| 2390 | Reliable Perturbative Results for Strong Interactions?: Politzer, H.David, (1973) | 7 | 43 |
| 1978 | Pseudoparticle Solutions of the Yang-Mills Equations: Belavin, A.A., (1975) | 8 | 56 |
| 3556 | Maps of dust IR emission for use in estimation of reddening and CMBR foregrounds: Schlegel, David J., (1997) | 9 | 13 |
| 2332 | Axial vector vertex in spinor electrodynamics: Adler, Stephen L., (1969) | 10 | 47 |

**Table 1.** Top 10 publication by PageRank, when damping factor = 0.50 (CC = Citation Count, RCC = Rank by Citation Count)

$p_x(t)$ is the probability of initial selecting the $x^{th}$ node in the citation graph. **Results.** Analyzing the ranking results, we discovered in Top 100 cases of older publications, with a modest number of citations, which, due to the fact that they acquired some of these citations recently, are ranked higher compared with the PageRank score, and so, they are easier to be discovered by the researchers. This is exactly the outcome we were hoping to see. Unfortunately, we also discovered some anomalies (Table 2).

The two publications presented in Table 2 have less than 20 citations, and thus,

| Citations | Publication | Rank | Rank by CC |
|---|---|---|---|
| 19 | Gauge symmetry and supersymmetry of multiple M2-branes: Bagger, Jonathan (2007) | 31 | 90786 |
| 18 | Comments on multiple M2-branes: Bagger, Jonathan (2007) | 32 | 94900 |

**Table 2.** Snapshot from Top 100 publications by Time-dependent PageRank (CC = Citation Count)

are ranked really low with the Citation Count ranking method. How is it possible to be so highly ranked with the new ranking method? Further investigations showed that the problem comes from the fact that these two papers are citing each other, and thus, are part of a cycle. Because of this and of the link-based

ranking which iteratively propagates the weight in the graph, when a strong time decay factor is used (in our case, a 5 year time decay), the newly published documents that are part of a cycle accumulate artificial weight. Unfortunately, this makes the time-dependent link-based ranking method unsuitable for data sets that allow cycles. As discussed previously, even the bibliographic data sets can allow cycles due to certain inconsistencies in the publication dates or in the listing of references. Since some of the publications are not dated, the identification/removal of the cycles is almost impossible due to the computational overhead. Because of this and of the link-based ranking which iteratively propagates the weight in the graph, when a strong time decay factor is used, the newly published documents that are part of a cycle accumulate artificial weight. Thus, this method is not suited for data sets that allow cycles.

### 4.4 Link-based Ranking with External Citations

As we saw in Section 3, the Inspire data set is missing on average 9 out of 30 references per paper while the CDS data set is missing on average 28 out of 37 references per paper. While for the Inspire data, these missing links represent just a small percentage, for the CDS data they represent almost 75%. In the context of applying the PageRank algorithm, this means that instead of distributing the weight to 37 references, a node is distributing its weight only to 9. This further means that these 9 papers receive much more weight then expected. So, we end up with a phenomena of "artificial inflation of weights".

For fixing this error we developed a new ranking method that accounts for the external citations. This new method assumes the existence of an "external authority" that accumulates weight from all the nodes in our graph, proportionally with the missing citations, and also feeds back into the network a certain percentage of its weight. With this method, we assure the correct propagation of the weight through the network.

The "external authority" (EA) node is controlled by two parameters, $\alpha$ and $\beta$. Each publication will contribute to the EA's weight with $\frac{\beta \times max\{1, ext_i\}}{\beta \times max\{1, ext_i\} + int_i}$, where $ext_i$ is the number of external citations for publication $i$, and $int_i$ is the number of internal citations. On the other hand, EA contributes to all publications with $\frac{\alpha}{n}$ weight, where $n$ is the total number of publications in the repository. Intuitively, $\alpha$ quantifies how much of the external weight is re-injected into the network and $\beta$ represents the fraction between an external citation and an internal one. We consider that, if a publication is not in the data set, it means that it values less for the repository then the ones already inserted in the database.

**Results.** In order to analyze how $\alpha$ and $\beta$ influence the final outcome of the ranking we calculated the Spearman Correlation Coefficient (SCC) between our new ranking method with different settings of $\alpha$ and $\beta$ (between 0 and 1 with 0.1 step), and the PageRank, for the CDS data set.

Table 3 presents the aggregated results after 200 experiments (for each $\alpha, \beta \in (0, 1)$, with a step of 0.1). Our experimental analysis showed that $\alpha$ only influences the rate of convergence of the iterative algorithm (with the best convergency rate obtained for $\alpha = 0.5$) and has little impact on the general reordering

| $\alpha$ | $\beta$ | SCC with PageRank | SCC with Citation Count |
|---|---|---|---|
| $\alpha \in (0,1)$ | $\beta = 0.1$ | 0.97 | 0.89 |
| $\alpha \in (0,1)$ | $\beta = 0.2$ | 0.94 | 0.91 |
| $\alpha \in (0,1)$ | $\beta = 0.3$ | 0.92 | 0.92 |
| $\alpha \in (0,1)$ | $\beta = 0.4$ | 0.91 | 0.92 |
| $\alpha \in (0,1)$ | $\beta = 0.5$ | 0.89 | 0.93 |
| $\alpha \in (0,1)$ | $\beta = 0.6$ | 0.88 | 0.93 |
| $\alpha \in (0,1)$ | $\beta = 0.7$ | 0.87 | 0.93 |
| $\alpha \in (0,1)$ | $\beta = 0.8$ | 0.87 | 0.93 |
| $\alpha \in (0,1)$ | $\beta = 0.9$ | 0.86 | 0.93 |
| $\alpha \in (0,1)$ | $\beta = 1.0$ | 0.85 | 0.93 |

**Table 3.** Spearman Correlation Coefficient between PageRank/Citation Count and Ranking with External Citations (The SCC between the PageRank and the Citation Count is 0.81)

while $\beta$ is the one that makes a difference in the outcome of the ranking method. For $\beta \in [0.1, 0.5)$ the outcome of the new ranking method is highly correlated with the PageRank results, and less correlated with the Citation Count results. On the other hand, for $\beta \in [0.5, 1]$ the correlation with the PageRank method drops, while the correlation with the Citation Count remains approximately constant. We advise for the use of a $\beta$ lower than 0.5 since in this case the results will be less correlated with the citation counts and enough correlated with the PageRank as to assume that the artificial inflation problems are resolved. We believe Link-based Ranking with External Citations to be a better candidate than Citation Count or PageRank for the task of ranking scientific publications because: (i) it inherits from PageRank its ability to take into account the citations with weights representing their importance, and thus, fixing one of the main shortcomings of the Citation Count method; (ii) it further corrects one of PageRank's shortcomings, namely the artificial inflation of some of the weights. In the end, our new ranking method is enough correlated with the PageRank method as to assume that it inherits its usefulness and in the same time it corrects its shortcomings.

## 5 Conclusions

The Citation Count is a very popular measure of the impact of a scientific publication. Unfortunately, it has two main disadvantages: it gives all the citations the same importance and it does not take into account time. These drawbacks motivated our study of alternative approaches: Time-dependent Ranking methods and Link-based Ranking methods. The time-dependent ranking methods were developed to take into account the time dynamics of the citation graph. More precisely, we first introduced time- dependent citation counts, taking into consideration the lifetime of the citations. Finally, we combined the link-based ranking with the time-dependent citation counts , creating the Time-dependent

Link-based Ranking. Unfortunately, this algorithm is not well suited for the citation graphs that are not DAG, due to the fact that it tends to overweight the young publications that are part of a cycle. The link-based ranking methods were developed to take into account the importance of the citing papers. We started with the PageRank algorithm originally designed for ranking web pages. In order to make it better suited for the bibliographic citation graph, we first modified the setting of the damping factor. Furthermore, we adjusted the PageRank model by adding an "external authority" node that represents a place holder for all the missing citations. In particular, this additional node prevents some publications from getting artificially boosted simply because of the incompleteness of the citation graph. We believe Link-based Ranking with External Citations to be a better candidate than Citation Count or PageRank for the task of ranking scientific publications.
In terms of future work, we plan to carry out a study on combining the above mentioned ranking methods that are based on citations with other ranking methods that are available in the CDS Invenio software, notably the download counts, word similarity, and reputation measures such as the Hirsch Index.

**Acknowledgments.** We would like to thank our *CDS Invenio* colleagues for their continuous support during this project. Also, we would like to thank Travis Brooks, Inspire Project coordinator, for fruitful discussions.

## References

1. Cds invenio. `http://cds.cern.ch/`.
2. Inspire project. `http://inspire.cern.ch`.
3. Webometrics. `http://repositories.webometrics.info/top400_rep_inst.asp`.
4. M. Gracco J-Y. Le Meur N. Robinson T. Simko A. Pepe, T. Baron and M. Vesely. Cern document server software: the integrated digital library. `http://doc.cern.ch/archive/electronic/cern/preprints/open/open-2005-018.pdf`, 2005.
5. D. Walker et al. Ranking scientific publications using a simple model of network traffic. 2006.
6. E. Amitay et al. Trend detection through temporal link analysis. In *J. of the American Society for Information Science and Technology*, pages 1–12, 2004.
7. K. Berberich et al. T-rank: Time-aware authority ranking. In *WAW*, pages 131–142, 2004.
8. P. Chen et al. Finding scientific gems with google. In *J.Informet. 1*, pages 8–15, 2007.
9. J.-Y. Le Meur J.-B. Claivaz and N. Robinson. From fulltext documents to structured citations: Cern's automated solution. 2001.
10. M. Rajman M. Vesely and J.-Y. Le Meur. The d-rank project: Aggregating rankings for retrieval of scientificpublications in the hep domain. 2008.
11. M. Rajman M. Vesely and J.-Y. Le Meur. Using bibliographic knowledge for ranking in scientific publication databases. pages 201–212, 2008.
12. Y. Zhao N. Ma, J. Guan. Bringing pagerank to the citation analysis. pages 800–810, 2007.
13. L. Page S. Brin. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.