An Efficient Algorithm for Ranking Research Papers based on Citation Network

Advance Algorithm Project

National Institute of Technology, Karnataka

October – January, 2021



Submitted to: **Dr. Sidney Rosario**

Submitted on: 22/01/2021

Submitted by:  **MOHD ASIF KHAN (202IT013)**

**ROUNAK MUKHERJEE(202IT023)**

**SHREYA NEEMA (202IT026)**

**DEEPU PATEL (202IT005)**

INDEX

Content

1. Abstract
2. Introduction
3. Research Methodology
4. Results
5. Conclusion and Future Work
6. References

**1.ABSTRACT**

The paper discusses an efficient method to rank the research papers from various fields of research published in various conferences over the years. This ranking method is based on citation network. The importance of a research paper is captured well by the peer vote, which in this case is the research paper being cited in other research papers. We used DBLP dataset which contains information about various research papers from various fields published over the years. Firstly, we calculated the Outlinks i.e. number of papers a particular paper cites and from that we calculated the Inlinks i.e. the number of papers which cite this particular paper. Now we used a modified version of the PageRank algorithm and rank the research papers, assigning each of them an authoritative score. The PageRank algorithm is based on the fact that the quality of a node is equivalent to the summation of the qualities of the nodes that point to it. In this case, quality refers to the score of the research paper. Now if a research paper cites more than one research paper, it is obvious that it has drawn inspiration from various sources and hence its effect on the score of the paper it cites should diminish by a factor equal to the number of papers it cites. Thus, we divided the Inlink score by the number of Outlinks of the Inlink. The iteration finally stops when there is no change in the score of any paper during the iteration. In the code, author has introduced the concept of damping factor which is also taken as inspiration from original PageRank. Now, using the scores of the research papers, we calculated scores for conferences and authors and rank them as well. Author has also considered a new metric in the algorithm which takes into account the time factor in ranking the research papers to reduce the bias against the recent papers which get less time for being studied and consequently cited by the researchers as compared to the older papers. Often a researcher is more interested in finding the top conferences in a particular year rather than the overall conference ranking. Considering the year of publication of the papers, in addition to the paper scores we also calculated the year-wise score of each conference by slight improvisation of the above mentioned algorithm.

**2.INTRODUCTION**

It is a real challenge for a researcher to refer to a particular paper or group of papers on a particular topic which will be best suited for his kind of research considering the pool of papers available to him. Thousands of research papers are published every year and these papers span various fields of research. For a new researcher, it becomes a very difficult task to go through the entire repository of research papers in order to determine the important ones. So the author ranked the paper so that researchers can refer to a particular paper according to the rank of the paper. Given a repository of research papers. But the question is how to know which research papers are important and how to find “important” research papers in repository of data? Which metric to use for determining a research paper is important ? Finding research papers which are important according to a certain metric in efficient time. Here proposed metric is number of citations (inspiration) taken from a paper. The papers which have more citations are considered popular and important in this case. Once a score based on citations is assigned for each paper , it can be used for comparing the Authors who have written paper as well as Journals in which paper got published.

**3.RESEARCH METHODOLOGY**

**3.1. Dataset**

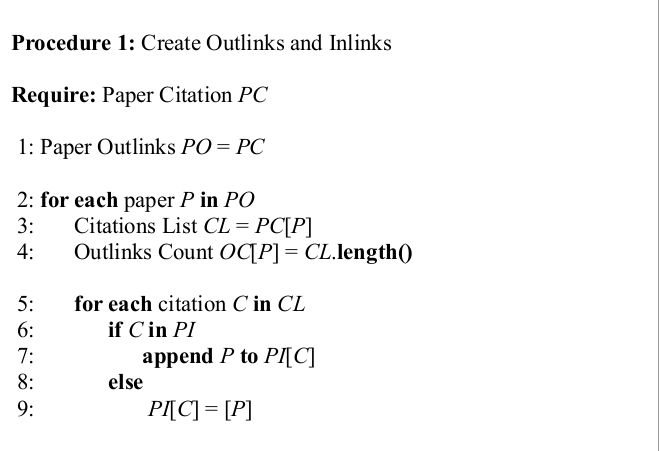
For building the citation network, we used the DBLP XML Records available at http://dblp.uni-trier.de/xml/ . The DBLP dataset contains information about various research papers from various fields published over the years. This information includes the name of the research paper, its author(s), the year of publication, the conference of publication and the list of research papers the given research paper cites. We built the citation network defined as a graph, with each research paper representing a node and the citations representing the edges in the graph, the edges being directed ones, directed from the citing node to the cited node. Each node is uniquely represented by the unique key from DBLP and the node has several attributes viz. research paper title, author(s), year of publication and conference of publication.

**3.2.Citation Graph**

Our main focus here is to rank the research papers. The rankings for conferences and authors are derived from the research paper ranks. We will build the citation network defined as a graph, with each research paper representing a node and the citations representing the edges in the graph, the edges being directed ones, directed from the citing node to the cited node. Each node has several attributes viz. research paper title, author(s), year of publication and conference of publication. We define the citation graph G = (V, E) comprising a set V of nodes, which each node Ni representing a research paper Ri and a set E of directed edges, with each edge E(i,j) directed from the citing node Ni to the cited node Nj.

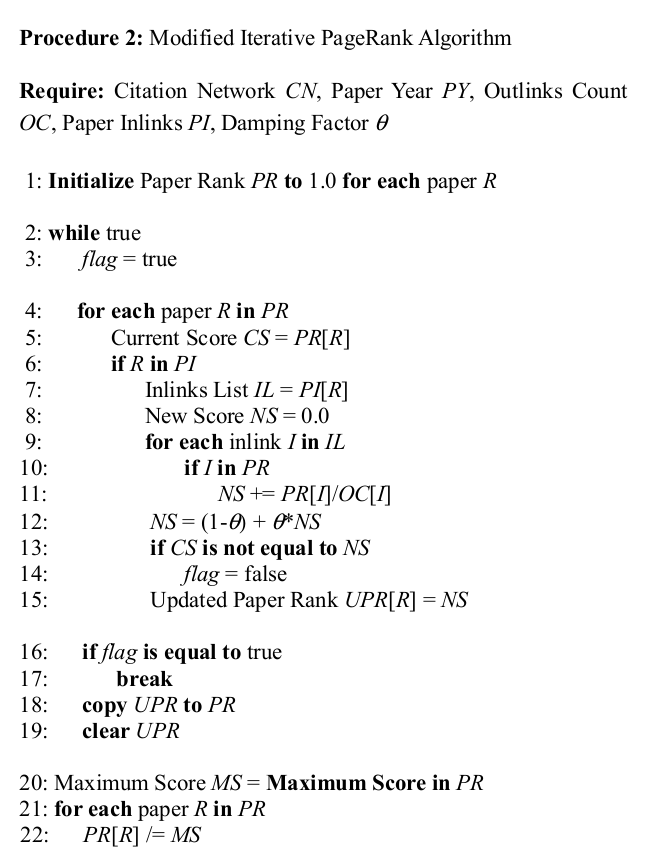
**3.3. InLink/OutLink**

Outlinks: From a given node N, link all the nodes N i that the node N cites. Inlinks: To a given node N, link all the nodes N j that cite the node N. To calculate Outlinks and Inlinks -



**3.4.Modified iterative page rank algorithm**

The iterative PageRank algorithm starts with initializing all the candidates to a constant value, generally unity and then it iteratively modifies each candidate's score depending on the score of the candidates that point towards it. It stops when all the candidate scores converge, i.e. become constant. Step 1 initializes the score of each paper to unity. Step 2 to 19 is the iterative calculation of the authoritative score for each paper. The iteration stops when there is no change in the score of any paper during the iteration. This is signified by no change in value of flag set to true in step 3; flag is set to false in step 13 and 14 if there is a change in the score of any paper during the update step. From step 4 to 15, for each paper, a new authoritative score is calculated based on the scores of the inlinks in the previous iteration (Step 9 to 11). The PageRank algorithm is based on the fact that the quality of a node is equivalent to the summation of the qualities of the nodes that point to it. In this case, quality refers to the score of the research paper. Now if a research paper cites more than one research paper, it is obvious that it has drawn inspiration from various sources and hence its effect on the score of the paper it cites should diminish by a factor equal to the number of paper it cites. This fact is used in step 11 by dividing the inlink score by the number outlinks of the inlink. Step 12 modifies the score calculated above to incorporate the damping factor. The use of damping factor is required to prevent the scores of research papers that do not have any inlinks from falling to zero. The value of damping factor is set to 0.85 [1]. The formula for calculating the score at a given iteration is: NS = 0.15 + 0.85 \* ∑ (PR[I]/OC[I]) . In step 20 to 22, we normalize the scores to scale down the scores within the range [0, 1]. Thus, finally we get the authoritative scores of each research paper based on the score of the research papers that have cited it. The iterative page rank algorithm-



**3.5. Time-independent Ranking**

The above algorithm is fairly sufficient for ranking the research papers. But there is one important aspect that has not been considered yet – time. It is quite obvious how time is a very important factor when it comes to ranking the research papers. The older research papers obviously have more time to be studied by the researchers all over the world and consequently be cited in various research papers. The newer ones fall behind in this aspect. To make our algorithm time-independent, we need to have some normalization of the scores based on time. Time-independent Scoring To make the algorithm time-independent, we need a time-dependent metric that can be used in our formula for the authoritative score in the algorithm. The metric we propose to use is the Average Number of Citations per Paper in a year, i.e. the average of the total number of citations of the research papers published in each year. This metric is suitable for normalization with respect to time as it captures the fact that an older paper has more time to be cited by researchers in comparison to the recent papers. We need to modify the formula for calculation of NS in the procedure 2, to incorporate time-independence. For each research paper, we have the year of its publication. Using the year of publication of all the research papers, we will pre-compute the total number of citations for each year and the number of research papers published in each year. Using them, we will determine the average number of citations per paper for each year.

The following pseudo-code needs to be added to the procedure 2, before step 1.

0a: for each paper R in PY b: year Y = PY[R] c: Year Citation Count YCC[Y] += OC[R] d: Year Paper Count YPC[Y] += 1 e: for each year Y in YCC f: Average Year Citations Count AYCC[Y] = YCC[Y]/YPC[Y] Once we have the average number of citations per year, we need to modify the formula for calculating the authoritative score. We will simply divide the NS by the average number of citations in the year of publication of the paper concerned. The change required in procedure 2 is in step 12: 12a: year Y = PY[R] b: NS = (1-θ) + θ\*NS/AYCC[Y]

**3.6.. Ranking Conferences**

The rank of a conference depends on the quality of research papers it publishes. This is the key idea behind our algorithm for ranking conferences. Using the authoritative scores for the research papers, we will rank the various conferences by calculating a cumulative authoritative score for each conference. The score for a conference will essentially be the average score of all the research papers published in that conference. The formula for the score of a conference is: Conference Score CS = Σ(Paper Score PS)/( Number of Papers published in the Conference NPC[C]) .An extension of the ranking of the conferences is ranking the conferences year-wise and not as a whole. For example, we might want to know whether VLDB1990 was better than SIGMOD1992 in terms of the quality of research papers published or was it vice-versa or we might want to know how the quality of research papers published in KDD is varying over the years. For this, we need to make a small change in the way the conference score is being calculated. Earlier we used just the conference name as the key for scoring the conference score. Now we concatenate the year of publication to the conference of publication to the conference name and use this concatenated term is the key. Thus we have keys based on year-wise granularity like KDD1999, ICDE1991, etc. The formula for the score of the conference remains the same:

Conference Year-wise Score CYS = Σ(Paper Score PS)/( Number of Papers published in the Conference in a given year NPCY[C][Y])

**3.7. Ranking Authors**

The rank of an author depends not only on the quality of research papers but also the conference he/she publishes his/her papers in. For ranking authors, thus we will use the scores of the research papers published by the author as well as the score of the conference in which the paper was published. This would incorporate the fact that publication in a comparatively better conference is more difficult. The authoritative score for the authors thus is essentially a weighted mean of the scores of the research paper the author has published, with the score of the conference as the weight. The formula for the score of an author is:

Author Score = Σ(Paper Score PS \* Conference Score CS)/(Number of Papers published by the Author NPA[A]) .

**4.RESULTS**

We implemented our algorithms on the DBLP dataset and discovered various interesting results. The DBLP dataset contains 10,00,000 research papers from the various fields of research and published in various conferences all over the world. The time dependent algorithm converges faster than the time independent algorithm.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No of Nodes | 10000 | 100000 | 200000 | 300000 |
| Time Dependent(no of iterations to converge) | 1 | 31 | 30 | 30 |
| Time Independent(no of iterations to converge) | 1 | 36 | 46 | 106 |

**5.CONCLUSION AND FUTURE WORK**

Thus the modified iterative page rank algorithm does a helpful job of ranking the paper, authors, journals which will be really helpful for the researchers . We saw the difference the time independent concept has created. Further we would like to implement theme mining on the keywords based on the ranking of the research papers. This can be extended to keywords in conferences, which could help new researchers to identify the conferences of his/her interest. Similarly we can also extend this work to keywords by authors . We can implement the algorithm on more variety of datasets and also based on different factors to rank the papers if found beneficial.

**6.REFERENCES**

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7):107–117, 1998.

[2] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web’s link structure. Computer, 32(8):60–67, 1999.

[3] Y. Ding, E. Yan, A. Frazho, J. Caverlee, “PageRank for Ranking Authors in Co-citation Networks,” Journal of the American Society for Information Science and Technology, 2009.

[4] E. Garfield, “The History and Meaning of the Journal Impact Factor,” Journal of the American Medical Association, 2006.

[5] J. E. Hirsch, “An Index to Quantify an Individual’s Scientific Research Output,” Proceedings of National Academy of Sciences, 2005.

[6] J. Klienberg, “Authoritative sources in a hyperlinked environment,” Proceeding of the 9th Annual ACM- SIAM Symposium on Discrete Algorithms, 1998.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web,” Technical Report, Stanford InfoLab, 1999.

[8] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, “Arnetminer: Extraction and Mining of Academic Social Networks,” Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[9] Q. Zhao, S. S. Bhowmick, X. Zheng, “Characterizing and Predicting Community Members from Evolutionary and Heterogeneous Networks,” Proceedings of the 17th Conference on Knowledge Management, 2008.

[10] Aditya Pratap Singh ,Kumar Shubhankar ,Vikram Pudi “An Efficient Algorithm for Ranking Research Papers based on Citation Network” .