The Google PageRank Algorithm

Jiyu He(✉)

The University of Manchester, Oxford Road, Manchester M13 9PL, UK
aallisonhjyy@gmail.com

Abstract. The original purpose of Google's PageRank algorithm is to assess the significance of web pages based on their link structure. However, based on the principle of PageRank, it could be used to any network or graph in any domain, also over time it has been developed for social network analysis, link recommendation, and prediction. In this article we will discuss the math behind PageRank and its basic implementation, examine how these various applications are related mathematically and conceptually, and some lacks this classical web computing algorithm has.

Keywords: PageRank · Markov chain · Random walk

1 Introduction

In most region of mathematics, data could be seen as graphs, such as Internet. To analysis the Internet graph, PageRank algorithm was proposed. The PageRank algorithm was mentioned by Page and Brin in 1996 and then be used in Google engine, to determine the importance of web pages and sorting.

The basic idea of PageRank algorithm is to define a random walk model on a directed graph, which is a first-order Markov chain, to describe the behavior of a random walker visits each node along the directed graph. Under certain conditions, the probability of visiting each node converges to a stationary distribution under the limits. Currently, the stationary probability value of each node is its PageRank value, illustrating the importance of the node. The higher the PageRank value, the more important a page is and the higher it is likely to be in the ranking of Internet searches. Assuming that the Internet is a directed graph, then the visitor jumps to the next page with equal probability according to the connected hyperlink on each web page and continues to make such random jumps on the web, forming a first-order Markov chain.

Also, we explore the "potential organization" in the network through community discovery, that is, divide the nodes in the graph into different communities, so that the connections within the community are close and the connections between the communities are sparse. It is not a simple thing to discover the community of an extreme large network. Therefore, we can start from the local network when discovering the community of the large network. How to start from a node and finding the community of the node in the local network let us have a new acknowledge which called *personalized PageRank* [1]. For instance, in a shopping website, we need to recommend different

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 417–422, 2023.
https://doi.org/10.2991/978-94-6463-198-2_43

products according to users' preferences, so we need to calculate the importance of certain items to a user, in other words, the importance of other nodes to a node in the network. Personalized PageRank provides searchers a graph where they could always find what they are interested in.

In addition, PageRank algorithm could be defined on any directional graph, and has been applied to social impact analysis, text summarization and many other problems. Also, it triggered a lot of other deeper algorithms, like choosing the expert searchers looking for in a professional forum through Expertise Rank.

In this review we will introduce the basic frame of PageRank and how people improved it.

2 Digraph and Random Walk Model

The digraph is denoted as $G = V + E$, where V and E represent the set of nodes and directed edges respectively. For example, the Internet can be viewed as a digraph, where each web page is a node of the digraph and each hyperlink between web pages is an edge of the digraph. A sequence of edges from one node to another is called a path, and the number of edges is called the length of the path. A digraph is strongly connected graph if it starts from any node and can reach any other node. The digraph in Fig. 1 is a strongly connected graph.

Given that k is a natural number greater than 1, a node is called periodic if the length of the path from a node of the digraph back to that node is a multiple of k. A digraph is called aperiodic graph if it contains no periodic nodes; otherwise, it is periodic.

Figure 2 is an example of a periodic digraph. From the node A and return to A, you must go through the path A-B-C-A, all possible paths are multiples of the length of 3, so the node A is a periodic node.

Given a digraph containing n nodes, the directed graph is defined on a random walk model, which called the first order Markov chain. Assume that the probability of transition from one node to all nodes connected by directed edges is equal, to be specific, let a transition matrix be a n order matrix named M

$$M = \left[m_{ij}\right]_{n*n} \tag{1}$$

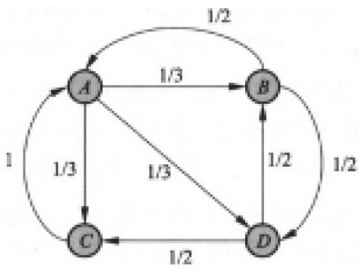

Fig. 1. Digraph

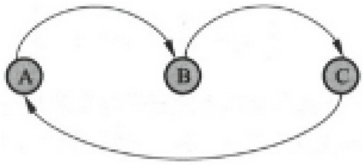

Fig. 2. Periodic digraph

Set i be the row and j be the column, element mij value as: If the node j has directed edges connected by k, and the node i is one of its connected nodes, then

$$m_{ij} = 1/k \tag{2}$$

Otherwise

$$m_{ij} = 0, i, j = 1, 2, \cdots\cdots, n \tag{3}$$

Notice that the transition matrix has the following properties

$$\begin{matrix} m_{ij} \geq 0 \\ \sum_{i=1}^{n} m_{ij} = 1 \end{matrix} \tag{4}$$

That is, each element is non-negative, and the sum of the elements in each column is 1, i.e. the matrix M is a random matrix. Random walks on digraphs form Markov chains. The random walk by a unit time teleports one state, this is to say, if the current time the random walker in the node j, then if the next moment the probability it in node i is mij. Then we could say the probability has the Markov property which it only depends on the current state, has nothing to do with the past.

The random walk model can be defined by using the digraph on Fig. 1. The random walker could teleport from node A to node B, C and D with probability 1/3 respectively, and has no possible to back to node A, then we could write the first column of the transition matrix. Also, the random walker could teleport from node B to node A, D respectively with probability 1/2, and no possible to back to node B and C respectively. Then we reach the second column of the matrix, and so on. Lastly, we have the transition matrix

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \tag{5}$$

We use vector Rt represents the probability distribution of random walker visit each node at time t, therefore

$$R_{t+1} = MR_t \tag{6}$$

3 Basic Definition of PageRank

Given a connected and aperiodic digraph with n nodes and define a random walk model on its basic. Assume that the transition matrix is M and the probability distribution of each node accessed at time 0, 1, 2, …,t,… are

$$R_0, MR_0, M^2R_0, \cdots, M^tR_{0,}, \cdots \tag{7}$$

And the limit exists

$$\lim_{t\to\infty} \mathbf{M}^t R_0 = R \tag{8}$$

The limit vector R represents the stationary distribution of Markov chains, satisfying

$$MR = R \tag{9}$$

4 General Definition of PageRank

General transition matrix of random walk model is linear composed by two parts, one is a basic transition matrix M from the directed graph, where the probability from one node to all other nodes it connected are equal; the other one is a completely random transition matrix, any node to any other node has transition probability with 1/n, with linear combination coefficient which is the damped coefficient d. This general random walk markov chain has a stationary distribution, denoted by R

$$R = d\mathbb{M}R + \frac{1-d}{n}e \tag{10}$$

where e is the vector with components of one.

General definition of PageRank means Internet visitors will random walk on the Internet: visitors decided to teleport randomly by hyperlinks from current page with probability d; or a completely random jump with probability (1-d), then it will jump to any web page with equal probability 1/n. The second mechanism which contains damped coefficient ensures that isolated pages (note: a web page that is not referenced by other pages) can be searched. In this way, smooth distribution is guaranteed, so general PageRank is suitable for any structure of the network.

5 Future of PageRank

We also considered with some unanswered problems.

Is it possible to characterize PageRank on an undirected graph in a straightforward way? Consider applying PageRank to an undirected, linked network with uniform teleportation. If $i = 1$, the stationary distribution is proportional to the node degree, and we have a pure random walk. We're not aware of any straightforward descriptions of PageRank's behavior when it gets away from 1. However, we know, the PageRank vector stays significantly associated with the degree, according to empirical data.

The kinds of network data accessible have expanded, and new PageRank-like models have been developed for a range of applications. Dynamical systems provide a logical method to expand PageRank-like principles for time-dependent networks and time-varying teleportation [2]. There is a basic extension of the PageRank notion for higher-order networks, as well as a computationally tractable variant that includes solving a polynomial system of equations [3]. There are novel PageRank constructions for multiplex networks with various sorts of connections among the same set of nodes to provide centrality scores that are dependent on each interaction type [4].

Is it possible the PageRank can calculate and sort huge amounts of data? PageRank was the first distributed computing algorithm to be deployed in a large-scale distributed system with clear logic and straightforward implementation, all graph database and graph computing manufacturers will use the algorithm.

The actual commercial benchmarking often requires two steps:

- Iterating through the whole quantity of data, sorting the findings.
- Returning the top-N results for comparison.

They are both necessary. We discovered that certain systems only do local calculations on specific data, clearly contradicting the algorithmic structure of PageRank's global repetitive computation.

A good example is Neo4j [5]. If the algorithm call and parameter restrict the return to 1000, it only calculates the PageRank value for 1000 vertices. This answer is 100 percent incorrect if the complete data is Twitter, which is comparable to just 1/40000th of the full data computation. Furthermore, if the results permit sorting; if the database does not support sorting, this ability directly reflects the design and implementation of a graph database. Most graph database benchmarks will lose credibility if they do not solve these difficulties.

Although the PageRank algorithm theory is basic, there are many aspects to be aware of, particularly in the benchmark test, such as whether the calculation results can support database writeback, file writeback, flow return, and sorting results.

6 Conclusion

However, we must not disregard the search engine system built up by PageRank, which has revolutionized our understanding of information retrieval. I have to say, it is a great miracle in the history of computer science at the end of the twentieth century; it is not only a ranking algorithm but also a framework on which to make some changes to solve more practical problems; it and its concise logic, invented so far in the search engine field is quite representative of the algorithm, solved hundreds of millions of web page quality assessment problems. In many areas of information retrieval, this data structure-based learning method continues to excite us.

Although nowadays many websites prefer to use hundreds of page quality assessment algorithms to make their multi-layer search engines, in order to push pages from different sections to users in a better and more comprehensive way, and the PageRank algorithm is no longer the only factor that Google uses to determine its rankings, it still plays an active role in determining the importance of websites. It considers the number and quality of

outbound links to your site: if you wish to get a good ranking, you need to get high-quality outbound links to your site. When you're trying to figure out how to rank on Google, consider Google PageRank. It will affect how you optimize your rankings. We predict sustained broad usage of the PageRank concept in new and fascinating applications over the next 20 years, given its universality and intuitive appeal.

References

1. David F Gleich. "PageRank beyond the Web". In: *siam REVIEW* 57.3 (2015), pp. 321– 363.
2. David F Gleich and Ryan A Rossi. "A dynamical system for pagerank with time-dependent teleportation". In: *Internet Mathematics* 10.1–2 (2014), pp. 188–217.
3. DF Gleich, LH Lim and Y Yu. *Multilinear PageRank. arXiv*, cs. 2014.
4. Arda Halu et al. "Multiplex pagerank". In: *PloS one* 8.10 (2013), e78293.
5. Rik Van Bruggen. *Learning Neo4j*. Packt Publishing Ltd, 2014.