APPLICATIONS OF STOCHASTIC MODELS IN WEB PAGE RANKING

E. Preethi¹, Dr. R. Arumugam²
¹Department of Mathematics, Periyar Maniammai University, Thanjavur
²Assistant Professor, Department of Mathematics, Periyar Maniammai University, Thanjavur

Abstract: This paper focuses on the applications of Markov chain on PageRank and discussed a few methods to handle the dangling node problem. The PageRank is modeled as the behavior of a randomized Web surfer; this model can be seen as Markov chain to forecast the behavior of a system that travels from one state to another state considering only the current condition.

Keywords: Markov Chain, Web surfer, PageRank, Transition Probability, Dangling node.

I. INTRODUCTION

Ranking or link analysis determine the success of the Web search engines as they compute the importance and relevance of individual page on the World Wide Web. Examples of link analysis algorithms are HITS (Hyperlink Induced Topic Search), PageRank and SALSA (Stochastic Approach for Link Structure Analysis). Google is intended to crawl and index the Web efficiently and create much more satisfying search results than existing systems [1]. These algorithms rely on the relation structure of the Web pages. HITS developed by Jon Kleinberg, is a query depend algorithm, which analyze the authorities and hubs value of a page while SALSA combines the random walk feature in PageRank and the center authority idea from HITS. Information retrieval methods use eigenvector calculations based on the popular methods of HITS and PageRank [2].

PageRank is a query and content self-governing. Query independent means that the PageRank ranks of all the pages offline after the crawler download and index pages and the rank remains constant for all the pages. PageRank model also estimates recommended solution methods, storage issues, existence and uniqueness [4]. Content independent means the PageRank does not include the contents of a Web page for ranking rather it uses the link structure of the Web to calculate the rank. Markov chain uses only a matrix and a vector to model and predict it. Markov chains are used in places where there is a transition of states.

II. PRELIMINARIES

2.1 Definition: (Markov chain)

Markov Chain is a random process used by a system that at any given time \( t=1, 2, \ldots, n \) occupies one of a finite number of states. At each time \( t \) the system moves from state \( v \) to \( u \) with probability \( P_{uv} \) that does not depends on \( t \) and \( P_{uv} \) is called as transition probability which is an important feature of Markov chain and it decides the next state of the object by considering only the current state and not any previous states.

2.2 Definition: (Transition Matrix)

Transition Matrix \( T \) is an \( n \times n \) matrix formed from the transition probability of the Markov process, where \( n \) represents the number of states. Each entry in the transition matrix \( t_{uv} \) is equal to the probability of moving from state \( v \) to state \( u \) in one time slot. So, \( 0 \leq t_{uv} \leq 1 \) must be true for all \( u,v=1,2,\ldots,n \).
Example:

Three state transition matrices is

\[
t_{uv} = \begin{bmatrix}
1 & 1 & 1 \\
4 & 2 & 4 \\
1 & 0 & 1 \\
2 & 1 & 2 \\
1 & 1 & 1 \\
2 & 4 & 4
\end{bmatrix}
\]

2.3 Definition: (PageRank)

PageRank (PR) is a calculation, famously invented by Google founders Larry Page and Sergey Brin, which evaluates the quality and quantity of links to a webpage to determine a relative score of that page’s importance and authority.

The equation of the PageRank is,

\[
PR(p) = d \sum_{q \in \text{Pr}_p} \frac{PR(q)}{O_q} + (1-d)
\]

Here, \(d\) is a damping factor such that \(0 < d < 1\) and \(O_q\) is the number of out-going links of page \(q\).

2.4 Definition: (Dangling node)

In the transition matrix, if sum of any rows is zero that indicates that there is a page with no forward links. This type of page is called as dangling node or hanging node.

2.5 Definition: (Web Graph)

PageRank algorithm treats the Web as a directed labeled graph whose nodes are the pages and the edges are the hyperlinks between them. This directed graph structure in the Web is called as Web Graph.

Example:

A graph \(G\) consists of two sets \(V\) and \(E\). The set \(V\) is a finite, nonempty set of vertices. The set \(E\) is a set of pairs of vertices; these pairs are called edges. The notation \(V(G)\) and \(E(G)\) represent the sets of vertices and edges, respectively of graph \(G\). It can also be expressed \(G = (V, E)\) to represent a graph. The graph below is a directed graph with 3 vertices and 3 edges.

![Fig (1) A directed Web Graph](image)

The vertices \(V\) of \(G\), \(V(G) = \{1,2,3\}\). The Edges \(E\) of \(G\), \(E(G) = \{(1, 2), (2, 1), (2, 3), (1,3), (3,1)\}\). In a directed graph with \(n\) vertices, the maximum number of edges is \(n(n-1)\). With 3 vertices, the maximum number of edges can be \(3(3-1) = 6\).

2.6 Definition: (Damping factor)

The damping factor \(d\), which is the click-through probability, is included to prevent sinks (i.e. pages with no outgoing links) from "absorbing" the PageRank of those pages connected to the sinks.
That is why the first term of the PageRank equation, \((1-d)\) is included. It is the chance of being on a random page after restart, while the second term is normalized so that all PageRanks sum to one.

### III. WHY PAGE RANK POPULAR?
- Google PageRank technology makes use of additional structure present in hypertext to provide much higher quality search.
- Google thus crawls the web and indexes the web much more efficiently than existing systems.
- PageRank is calculated mathematically and without human interference.
- Today PageRank has become a popular measurement standard for determining the value of a website.
- Maps user behavior.

### IV. DEMOCRATIC NATURE OF A PAGE RANK

![Diagram of PageRank](https://via.placeholder.com/150)

#### Random Surfer Model
- Assume that there is a “random surfer” who is given a web page at random.
- He keeps clicking on links from one page to another, never hitting “back”.
- But eventually gets bored and starts on another random page.
- The probability that the random surfer visits a new page is the PageRank of that page. Thus it closely models user behavior.

At this stage it is a kind of dangling node. That is there is no other link or source to click on to the other page. So, we can use the markov chain model to the PageRank calculation.

### VI. PAGE RANK CALCULATION

There are a number of link based ranking. Among them PageRank is the most popular link based ranking. PageRank and Google are developed by Brin and Page during their PhD at Stanford University as a research project. The PageRank is the heart of the Google search engine. Google was introduced in the search engine business in 1998. After that, it became one of the most efficient search engines because it is a query independent and content independent search engine. It produces the results faster because it is query independent i.e. the Web pages are downloaded, indexed and ranked offline. When a user types a query on the search engine, the PageRank just finds the pages on the Web that matches the query term and presents those pages to the user in the order of their PageRank.

If an in-coming link comes from a reputed page, then that in-coming link is given a higher weighting than those in-coming links from a non-reputed pages.
PageRank uses only the link structure of the Web to determine the importance of a page rather than going into the contents of a page. PageRank provides a more efficient way to compute the importance of a Web page by counting the number of pages that are linking to it.

**Example:**

A simple Web graph with 3 nodes 1, 2 and 3 as shown below:

![Fig (2)](image)

The PageRank for pages 1, 2 and 3 can be calculated by using the PageRank equation.

To start with, we assume the initial PageRank as 1.0 and do the calculation.

\[
PR(1) = (1 - d) + d \left( \frac{PR(2)}{O_2} + \frac{PR(3)}{O_3} \right)
\]

\[
PR(1) = 0.15 + 0.85 \left( \frac{1}{2} + \frac{1}{1} \right) = 1.425
\]

The damping factor \(d\) is set to 0.85.

Now,

\[
PR(2) = (1 - d) + d \left( \frac{PR(1)}{O_1} \right); \quad PR(2) = 0.15 + 0.85 \left( \frac{1.425}{2} \right) = 0.756
\]

\[
PR(3) = (1 - d) + d \left( \frac{PR(1)}{O_1} + \frac{PR(2)}{O_2} \right); \quad PR(3) = 0.15 + 0.85 \left( \frac{1.425}{2} + \frac{0.756}{2} \right) = 1.077
\]

This PageRank computation continues until PageRank gets converged. PageRank gets converged to a reasonable tolerance.

**VII. APPLICATIONS OF MARKOV CHAIN IN PAGERANK**

To describe the relationship between PageRank and Markov chain, imagine a random surfer surfing the Web, going from one page to another page by randomly choosing an outgoing link from one page to go to the next one. This can sometimes lead to dead ends i.e. pages with no outgoing links.

So, at a certain fraction of the time, the surfer chooses a random page from the Web. This theoretical random selection is known as Markov chain or Markov process. The limiting probability that an infinitely dedicated random surfer visits any particular page is its PageRank.

The PageRank iterates as follows:

\[
PR_p^{k+1} = \sum_{q \in \Omega_p} \frac{PR_q^k}{O_q} \quad \text{for } k = 0,1,2,\ldots
\]

Where, \(PR_q^k\) is the PageRank of page \(p\) with \(k\)th iteration, here \(O_q\) is the number of forward links of the page. Let \(T\) be the transition matrix for the Web then \(q^{k+1} = T q^k\). Normally all forward links are chosen equally as per the following...
Problem: 1

Let us considered a sample web graph extracted from a university site contains 7 pages namely Home, Admin, Staff, Student, Library, Department and Alumni. We use this sample Web graph in our Markov analysis and PageRank calculation.

![Diagram of the sample web graph](image-url)

The transition matrix $T$ can be produced by applying the PageRank formula mentioned to the sample web graph. In the transition matrix, if sum of any rows is zero that indicates that there is a page with no forward links. This type of page is called as dangling node or hanging node. There are a couple of methods to eliminate this dangling page problem. They are discussed using the transition matrix below.

![Transition matrix table](image-url)

Dangling pages is handled by replacing all the rows with $\frac{e}{n}$ where, $e$ is a row vector of all ones and $n$ is the order of matrix. For example, the value of $n$ is 7.
The new forward links from the Alumni page is shown using the dotted arrows. This makes the transition matrix $T$ as stochastic as shown below:

![Diagram](image_url)

**Fig (4)**

The new forward links from the Alumni page is shown using the dotted arrows. This makes the transition matrix $T$ as stochastic as shown below.

$$
\begin{bmatrix}
0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\
0 & 1 & 0 & \frac{1}{3} & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 \\
\frac{7}{7} & \frac{7}{7} & \frac{7}{7} & \frac{7}{7} & \frac{7}{7} & \frac{7}{7} & \frac{7}{7} \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 1 & 1 \\
6 & 6 & 6 & 6 & 6 & 6 & 6 \\
0 & 0 & \frac{1}{3} & 0 & \frac{3}{3} & 0 & 1 \\
0 & 0 & 0 & \frac{1}{3} & \frac{3}{3} & 1 & 0 \\
0 & 0 & 0 & 0 & \frac{3}{3} & 0 & 3 \\
\end{bmatrix}
$$

*Table (2)*

The row 3 in the transition matrix, (Alumni page) is connected to all the nodes and also connected back to it (shown in the dotted lines).

Hence, the Alumni page is no more a dangling page.

Let us consider another example of a bus route among cities such as take some named cities like Chennai, Kumbakonam, Madurai, Trichy, Tanjore, Thirunalveli.
Here there is no outgoing links from the city Thanjavur. It means it is a dangling node.

Dangling pages is handled by replacing all the rows (where there is a dangling node ) with $\frac{e}{n}$ , where $e$ is a row vector of all ones and $n$ is the order of matrix.

Therefore, we get the matrix as

$$A = \begin{bmatrix}
0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\
\frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\
\end{bmatrix}$$

There is no dangling node or pages. Every page has an outgoing links.

This stochastic property is not enough to guarantee that Markov model will converge and a steady state vector exists. There is another problem with this transition matrix is that this matrix may not be regular. The general Web’s nature makes the transition matrix not regular. In the graph, every node needs to be connected to every other node. But in the real Web, every page is not connected to every other page i.e. it is not strongly connected. Brin forced all the entries in the transition matrix to satisfy $0 \leq t_{pq} \leq 1$ to make it regular. This ensures convergence of $q_n$ to a unique, positive steady state vector.

VIII. ISSUE WITH PAGERANK

- Prefer old document than new.
- Pages redirect to main page itself raising their rank spoofed PageRank.
- Search optimizer selling high PageRank’s to web masters.

IX. APPLICATIONS BEYOND GOOGLE

- Dynamic price setting
- Programmable networks
- Stock market trading
- Opinion polls
- Web mining
- Theme based ranking
- Reputation system for ecommerce
- Collaborating filtering
- Business intelligence
X. CONCLUSION

This paper brings anonymity about how the PageRank used relevancy set with the Markov chain. This paper highlights the different adjustments done to make the Web graph into a Markov model. In that, the dangling node problem and the methods to handle the dangling nodes are discussed. Mathematical solutions are also provided.

REFERENCES