On PageRank Algorithm and Markov Chain Reduction

Wenxing Ye

Abstract

The PageRank is used by search engines to reflect the popularity and importance of a page based on its reference ranking. Since the web changes very fast, the PageRank has to be regularly updated. Such updates is an challenging task due to the huge size of the World Wide Web. Consequently, the analysis of the PageRank has become a hot topic with vast literature ranging from the original paper by Brin and Page, to the works by specialists in Markov chains, linear algebra, numerical methods, information retrieval and other fields.

In this work, we shall concentrate on the Markov chain formulation of the PageRank problem. And analyze the implications of aggregation method (Deng. K & P., 2009) in PageRank computation. Such method exploit the block structure of the web and takes the dynamics into consideration. The analysis will be based on the information theory and Perron-Frobenius theory.

1. Introduction

The World-Wide Web may be viewed an ultra-complex graph through which sharing and dissemination of information are happening on an unprecedented scale. The pages and hyperlinks of the World-Wide Web are nodes and arcs in this directed graph. The size of world-wide-web makes it a very challenging problem to locate specific information in it. Formally known as information retrieval, this problem has become very hot research topic.

PageRank is one of the earliest techniques to increase the efficiency of Information Retrieval on the Web. It is essentially the method of finding page authorities produced by the Web graph structure. First introduced in 1998 by Google search engine (Brin & Page., 1998), the PageRank of a page \( i \) reflects the importance of this page basing on: 1) how many pages link to \( i \), and 2) how important are the pages that link to \( i \). In a search, a document is retrieved with regard to a particular query either if it contains all the words that are in the query or if these words are in the text of a link which is pointing to the document itself. After retrieving the documents related to a particular query, PageRank orders them and display the result to the end user. Such a step is necessary because a search can return much more pages than we expect. In the spring of 2005, it was estimated that Google made searches among 8.1 billion web pages.

Markov chain is named after Andrey Markov, is a mathematical system model that describes transitions from one state to another on a state space. It used to model random processes that is characterized by memoryless property: the next state depends only on the current state and not on the sequence of events that preceded it. This is also known as the Markov property. There are many applications of Markov chains, including physics, biology, economics, engineering, and other fields. Perhaps one of the most recent examples of the use of Markov chains is the web search engine. Modeling the World wide web as a huge Markov chain implies treating the pages as nodes and the links between them as transitions.

In this paper we focus on the Markov chain interpretation of PageRank algorithm. Specifically, we use the Markov chain aggregation as a tool to analyze the properties of page rank, based on the aggregation result, we also proposed an alternative formulation for the PageRank algorithm. The rest part of this paper is organized as follows: Section 2 review related work in the literature. Section 3 introduce the Markov chain interpretation of PageRank algorithm, Section 4 and 5 elaborate on the aggregation of Markov chain and in particular, the aggregation of PageRank Markov Chain. Future work is discussed in Section 6.
2. Related Work

The definition of PageRank is given in (Brin & Page, 1998), and further details about PageRank is presented in (L. Page & Winograd.). Other work on the implementation and computation of PageRank can be found in (Kim & Lee., 2002). An information theoretic aggregation framework for reducing the size of Markov chains is given in (Deng. K & P., 2009). Interpretation of PageRank algorithm using Markov chains can be found by doing a Google search with these two key words “Markov Chain” and “PageRank”. Stoachstic stability of Markov chains is discussed in (Meyn & Tweedie, 2009). The convergence speed of Markov chains are studied in (Levin & Wilmer, 2006) and (Ravi Montenegro, 2006).

3. PageRank and Markov Chain Interpretation

The Google PageRank is one method that the search engine Google uses to determine the importance of a page. It uses a special Markov chain to compute the rank of web pages and this rank determines in which order the pages should be listed in search results given by Google.

To understand PageRank from Markov chain perspective, we first denote all the web pages by a state space $W$ with size $n$, meaning $n$ pages. Let $C = \{c_{ij}\}$ denote the adjacency matrix, a $n \times n$ matrix with $c_{ij} = 1$ if there is an hyperlink from page $i$ to page $j$ and $c_{ij} = 0$ otherwise. The outgoing degree of page $i$, meaning the number of pages that can be reached from page $i$, will be the row sums, denoted as $s_i$:

$$s_i = \sum_{j=1}^{n} c_{ij}$$

Now normalizing the adjacency matrix $C$ by its row sums, we get $\bar{W} = \{w_{ij}\}$, defined by

$$w_{ij} = \begin{cases} c_{ij}/s_i, & \text{if } s_i \geq 1. \\ 1/n, & \text{if } s_i = 0. \end{cases}$$

This basically models the behavior of an internet user when he is at page $i$. If there are $s_i$ outgoing links on page $i$, the user will pick one, with equal probability, as its next page to visit. If there is no outgoing link on page $i$, the user will pick a random page according to uniform distribution to jump to. This is a simplified model of the user behavior but the generality is not lost.

By this, $\bar{W}$ can be seen as a transition probability matrix of the Markov chain with state space $W$. Furthermore, to define the Markov chain an additional tuning parameter $\gamma$ should be introduced, which is between 0 and 1 and describe the user specific preference. The transition probability matrix of the Markov chain used in PageRank is then given by:

$$P = \gamma \bar{W} + (1 - \gamma) \frac{1}{n} E$$

where $E$ is the $n \times n$ matrix with all entries being 1. This Markov chain describes the behavior of an internet user who, with probability $\gamma$ follow an outgoing link on the current web page with equal probabilities or, if the page has no outgoing links, jumps to another page randomly. Also, with probability $1 - \gamma$, the user jumps to a page randomly with equal probability. On proper choice of $\gamma$, this Markov chain is finite, irreducible also aperiodic, and there exists a unique stationary distribution $\pi$. This stationary distribution is used to rank all the pages in $W$: the page $i$ with the largest $\pi_i$ will be ranked first, the second largest be ranked second, and so on. In the computation of PageRank, $\gamma$ is usually set to 0.85).

4. Aggregation of Markov Chains

An information theoretic Markov chain reduction is proposed in (Deng. K & P., 2009) to find a reduced order Markov chain closest to the original one, measured by Kullback-Leibler rate divergence(Sun & Mehta, 2010; Cover & Thomas, 1991):

$$R(P||Q) = \sum_{i,j \in N} \pi_i P_{ij} \log \left( \frac{P_{ij}}{Q_{ij}} \right)$$

where $P$ and $Q$ are two Markov chains defined on the same state space $\mathbb{N}$, $P_{ij}$ and $Q_{ij}$ are the entries of their state transition matrices, respectively. $\pi_i$ is the $i$-th entry of the stationary density of $P$. It has been shown in (Deng. K & P., 2009) that finding the optimal bi-partition of a given Markov chain can be reduced to an eigenvalue problem

$$\hat{P} \nu = \lambda \Pi \nu$$

where $\hat{P} = \frac{n P + \nu \rho}{\nu + \rho}$ and $\Pi = \text{diag}\{\pi\}$. Arbitrary partition of the Markov chain can be achieved by recursively applying this bi-partition algorithm.

This information theoretic Markov chain reduction approach is able to aggregate the states in a Markov chain according to their dynamic interaction: those states that interact with each other a lot are aggregated into one single state. This property is useful in our analysis of the PageRank algorithm. When several nodes interact frequently with each other, meaning that according to the internet user model, users
tend to jump back-and-forth between them frequently. This is a clear indication that there exist close relation between these pages. For example, it could imply that the topic of these pages are relevant, or belong to a same big category. With the above said, this aggregation approach can help identify hidden information on customer preference that would not be obvious otherwise.

5. Aggregating the PageRank Markov Chain

5.1. Human user behavior vs graph structure

It should be noted that a human internet user would not necessarily observe the model in 5.2. For example, he would not follow all the outgoing links on a page with equal probability. Instead, his choice is made according to the information contained in that link (whatever the link displays) and his interest in these information. Similarly, jumping to another page is also not a random action but constrained to a much smaller set of pages in \( W \), and also with a different probability distribution that is decided by his preference and the circumstance, instead of being uniform. To this end, the information of users can only be tracked if the Markov chain is built based on the true browsing history of human users, which technically is not a very difficult task but privacy might be a concern.

When the rank of the pages are computed from the web crawler, it is however a different story. A web crawler is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (Wikipedia). Web search engines and some other sites use web crawling or spidering software to update their web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping.

The problem of using we crawlers to compute page rank is that the results only reflect the connectivity of the internet and ensuing importance rank, not necessarily the importance as appreciated by human internet users. For example, there are a number of ways to artificially increase a web site’s rank to make it appear important to web crawlers. On the contrary, if the rank is computed from the browsing history of real human user, the information contained in the jumps can be explored by the algorithm.

5.2. special cases

In the following we discuss some special cases in aggregating the PageRank Markov chain in equation \( \gamma \). Note the two extreme cases of the tunable parameter \( \gamma = 1 \) and \( \gamma = 0 \).

If \( \gamma < 1 \), the Markov chain is aperiodic since all its states have a non-zero probability to transit back to themselves. If \( \gamma = 1 \), we get \( P = W \) and that is the special case when the internet user tends to always follow the link on the current page, and never make a sudden jump away from the current page. However the a-periodic and irreducible property has no guarantee, and all depends on the \( W \). For example, it is possible that there exist two pages which are only linked to each other (each one has a single outgoing link connecting itself to the other page). In this case, this the bigger Markov chain on the whole state space \( W \) is neither aperiodic nor irreducible. Then there is no unique stationary distribution. Unfortunately, in realistic cases considering how the internet is structured, this situation do occur. So a good practice is to keep the parameter \( \gamma < 1 \). It is important to note that if the information theoretic aggregation method is applied to the Markov chain for this \( \gamma = 1 \) case, what it does is to group the pages only according to how they are linked, without consideration of any chance of random jumps.

On the contrary, when \( \gamma = 0 \), we have

\[
P = \frac{1}{n} E
\]

which means that the random walk behavior is totally eliminated, and the internet user will randomly jump to any state of the Markov chain (page) with equal probability. The trajectory will look like pure noise and the rank of pages will not not make any sense. The aggregation algorithm will not work for such cases because obviously all pages are equal and equally connected to each other.

5.3. Higher level information extraction

Different types of information can be drawn from the aggregated Markov chain. Given a Markov chain \( P \) defined on state space \( W \) of cardinality \( n \), suppose these states are aggregated into \( m << n \) groups (we denote the set of these groups \( Z \)), and an aggregated new Markov chain \( Q \) defined on \( Z \) is obtained from \( P \). Then it is interesting to run the page rank algorithm to obtain the ranking of the \( m \) groups. If the original Markov chain \( P \) is based on web crawler data, then this aggregated chain \( Q \) reflects the clusters in the connectivity graph among all pages. And this new
ranking well reflect the importance of these clusters as compared with each other. The new rank

\[ z_1, z_2, \cdots, z_m \]

can be obtained by computing the page rank based on the new Markov chain \( Q \), which is also smaller than \( P \).

If the original Markov chain \( P \) is based on human internet user data, then the aggregated chain can represent pages of different types, for example different topics or different type of content. Then the new Markov chain \( Q \) defined on set \( Z \) reflects the interaction between different types of content. The ranking conducted on new Markov chain \( Q \) can be used to study the importance of these types.

5.4. Alternative version of PageRank

The Markov chain aggregation algorithm suggests that a bi-partition of a given Markov chains should be conducted on a modified version of its state transition matrix \( P \):

\[ \hat{P} = \frac{\Pi P + P' \Pi}{2} \]

where \( \Pi = \text{diag}\{\pi\} \). Only this formula can guarantee the result approximately minimizes the cost function given by Kullback-Leibler divergence rate (Sun & Mehta, 2010; Cover & Thomas, 1991). This also motivates a new type of page rank formulation.

For a Markov transition matrix \( P \) in equation 5.2, instead of directly obtaining page rank from it, a better solution is to compute the page rank based on \( \hat{P} \). This way the states (i.e. pages) in the original Markov chain \( P \) is properly ranked according to their interaction with other pages.

6. Conclusion

In this work we analyzed the page rank algorithm. Based on the Markov chain formulation of this algorithm, we use the Markov chain aggregation as a tool to analyze the properties of page rank. We have shown that useful extra information can be extracted by applying page rank to the aggregated Markov chain. We also proposed an alternative formulation of the page rank algorithm that better rank the internet pages according to the dynamics of Markov chains. The next step is to implement the new algorithm to generate rank score and compare to available page rank algorithm.

References


Wikipedia.