11/01/18

APPM 2360 project 2:
Network Markov Chains

Recitation Section: 242 Wanqi Yu
Recitation Section: 281 Zongyi Huang
Recitation Section: 253 Lingyin Lu

I. Introduction

The “Markov Chain” is the common model that can describe the multiple moving states. It means that the following states will be dependent on previous states. The diagram can show the process and the arrows show the process probability direction. And if we want to visualize the probability of transition, the matrix can represent the state condition of each probability. In the following report parts, we will discuss the operation of the search engine with the “Markov Chain” model and analyze the stationary distributions inside the model, then, we will discuss the probability page rank eigenvectors with the eigenvalues of λ, which maximum is 1.

II. Modelling the Internet as a Markov Chain

1. The Markov Chain is shown as figure 1.

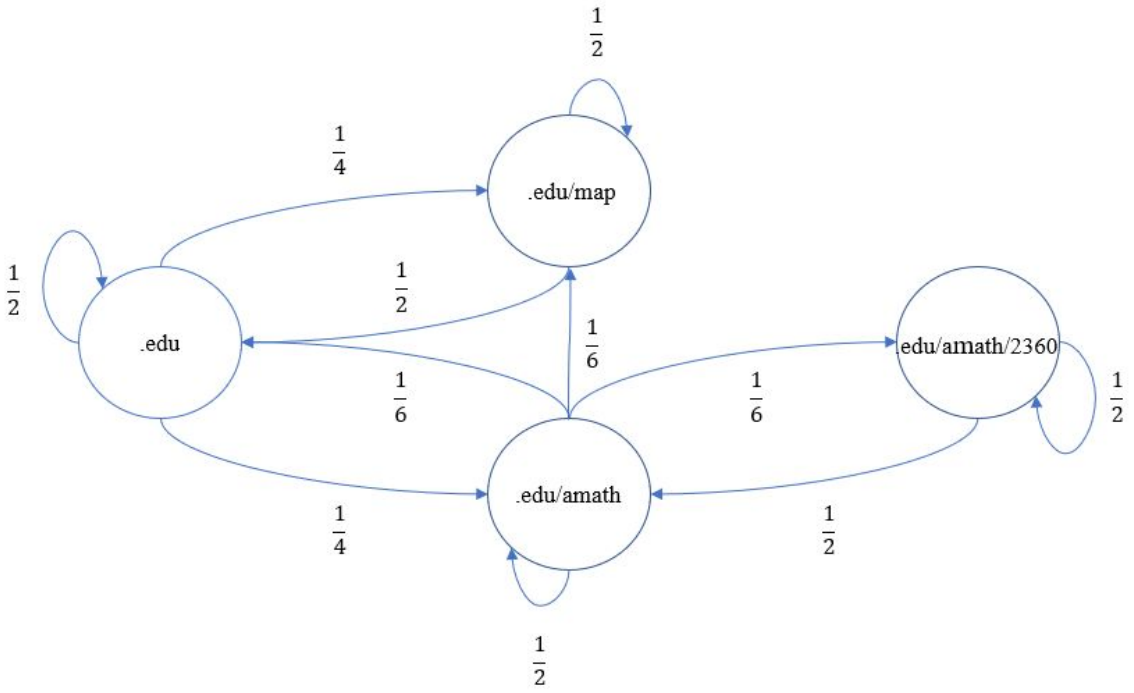


figure 1. Markov Chain of web hyperlink

As shown in this figure, each page has ½ probability link to itself.
Colorado.edu links to .edu/map and .edu/amath, and there are ¼ probability link to either .edu/map or .edu/amath.
Colorado.edu/map only links to colorado.edu, so the probability to .edu page is ½ .
Colorado.edu/amath links to all other pages, and they share ½ probabilities, so the probabilities to each page is ⅙ .
Colorado.edu/amath/2360 only links to .edu/amath, so the probabiliy is ½ .

2. As shown in the Markov Chain, the users have 50% probabilities stay as the homepage, 25% probabilities go to colorado.edu/map and 25% go to colorado.edu/amath.

3. The matrix that describes the Markov Chain is shown as figure 2.

	.edu	/map	/amath	/2360
Go to.edu	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	0
Go to.edu/map	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{6}$	0
Go to.edu/amath	$\frac{1}{4}$	0	$\frac{1}{2}$	$\frac{1}{2}$
Go to amath/2360	0	0	$\frac{1}{6}$	$\frac{1}{2}$

figure 2. The matrix of Probabilities go to each page

III. Stationary Probability Distributions

1. In this case, we describe the probabilities link to other pages as the matrix shown in figure 2. And the probabilities on each of four pages after 100 steps is described as:

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{6} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{6} & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{6} & \frac{1}{2} \end{pmatrix}^{100}$$

By calculating, we can get the result equal to:

$$\begin{pmatrix} 0.363636 & 0.363636 & 0.363636 & 0.363636 \\ 0.272727 & 0.272727 & 0.272727 & 0.272727 \\ 0.272727 & 0.272727 & 0.272727 & 0.272727 \\ 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 \end{pmatrix}$$

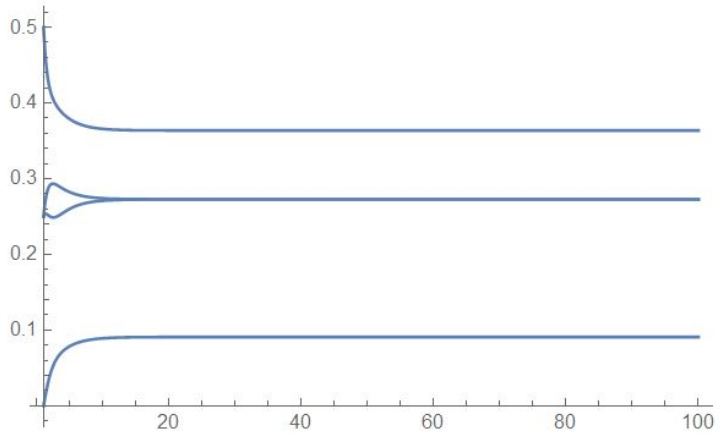
We know that we are going to start from the homepage, so we can express that as a vector:

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

By multiplying two matrices above, we can get the probabilities on each of four pages after performing 100 steps from the homepage.

$$\begin{pmatrix} 0.363636 & 0.363636 & 0.363636 & 0.363636 \\ 0.272727 & 0.272727 & 0.272727 & 0.272727 \\ 0.272727 & 0.272727 & 0.272727 & 0.272727 \\ 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.363636 \\ 0.272727 \\ 0.272727 \\ 0.0909091 \end{pmatrix}$$

The graph shows the probabilities that we stay on each of the four pages tend to close to the answer we get above as the perform steps increasing.



2. The stationary distribution error may be created by the calculated times. We can get each column matrix sum with infinite precision. However, in the real situation, the computer cannot be infinitely precise. For the calculation, we will get more accurate results, because the equation $a_1 + a_2 = 1$ is always true, and the vector is the fixed point. So, when we get the final results, the answers will have round errors of finitely precise.

3. For the characteristic equation, $|A - \lambda I| = 0$, we can get the eigenvectors of a matrix with different eigenvalues. And for each eigenvalue A_i, it is easy to find the eigenvector(s) V_i by solving the algebraic system: $(A_i - \lambda_i I)\, V_i$ **= 0**. And the n-dimensional vector x can be described as $x = c_1 v_1 + \ldots + c_n v_n$. So, the stationary distribution of a Markov chain may be found from the vector x that can be indeed different eigenvectors from each eigenvalue until the max of λ is equal to 1. And the stationary distribution may become a basis of the corresponding matrix.

4. The Markov Chain with five pages is shown as figure 3.

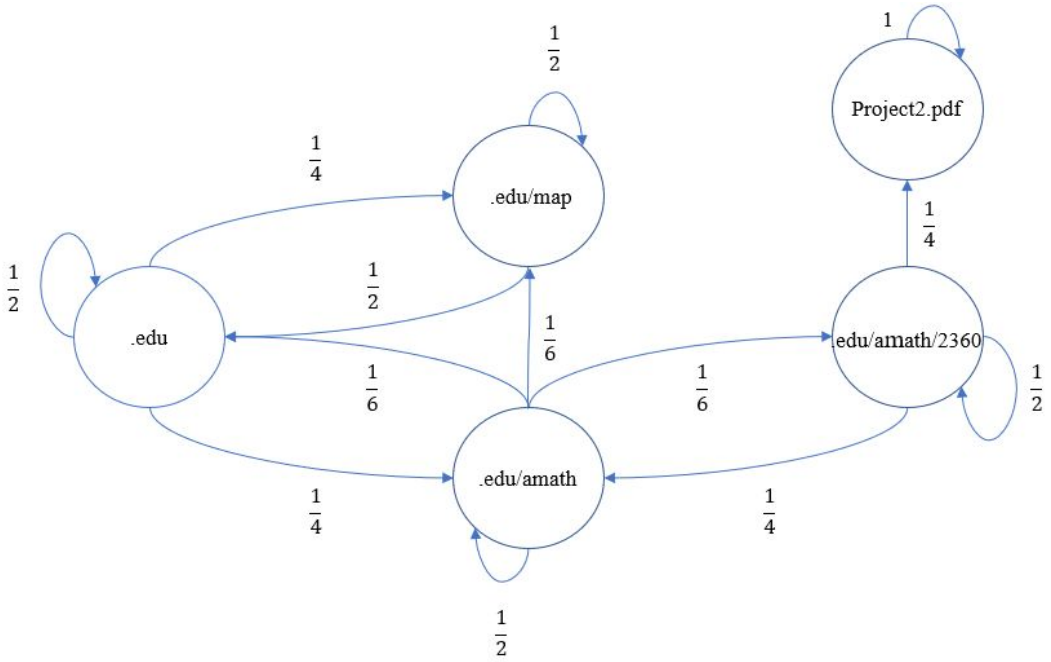


figure 3. Markov Chain with five pages

The new Matrix that describes all five pages and the probabilities to each page is shown as figure 4.

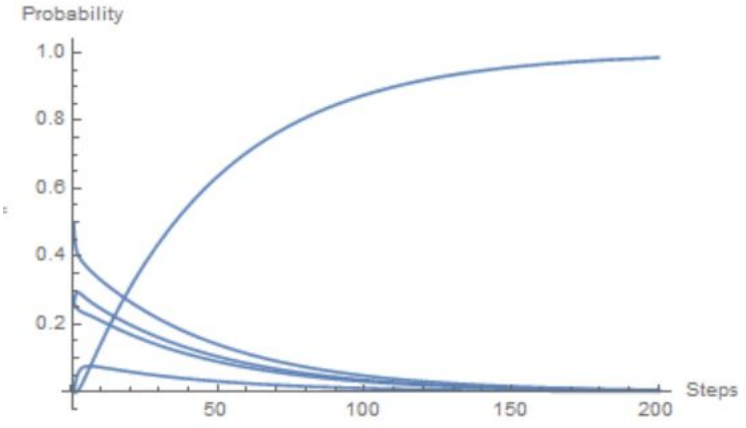
	.edu	/map	/amath	/2360	Project2
Go to.edu	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	0	0
Go to.edu/map	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{6}$	0	0
Go to.edu/amath	$\frac{1}{4}$	0	$\frac{1}{2}$	$\frac{1}{4}$	0
Go to amath/2360	0	0	$\frac{1}{6}$	$\frac{1}{2}$	0
Go to project2.pdf	0	0	0	$\frac{1}{4}$	1

figure 4. The matrix of five pages

By using the same strategy as part 1, we get the probabilities on each of five pages after performing 200 steps.

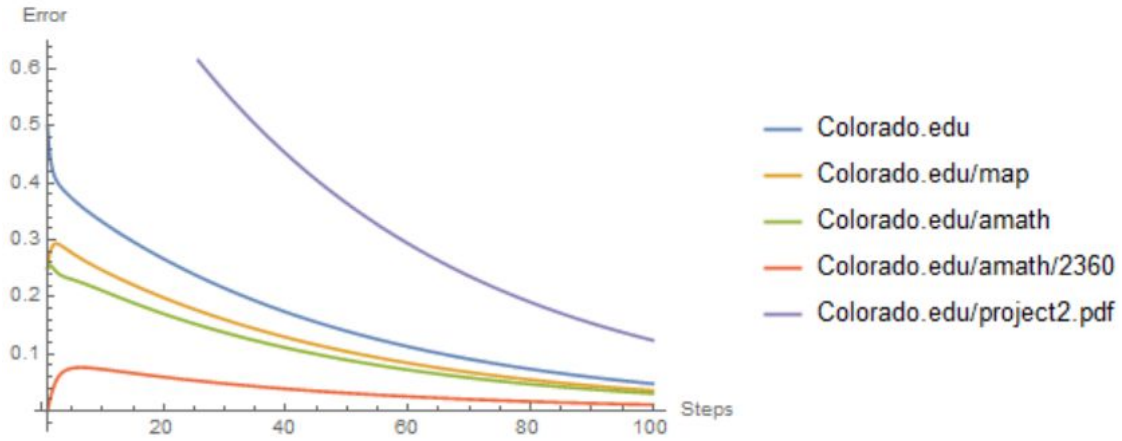
$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{6} & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{6} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{6} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 1 \end{pmatrix}^{200} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.00550719 \\ 0.00410039 \\ 0.00351561 \\ 0.00122409 \\ 0.985653 \end{pmatrix}$$

If we plot the graph of probabilities on each of five pages changing as performing steps, we see that the probability on colorado.edu/amath/2360/project2.pdf tends to be 1, and probabilities on other pages tend to be 0.



The reason that the stationary distributions will be like this graph is the project2.pdf page does not link to any other pages, and this became a closed loop. As we get into this page, there is no way to link to other pages. Therefore, the probability tends to be 1 as the preformed steps increasing.

5. To find the error, we assume $Error = \left|V_1 - X_n\right|$, V_1 is the final probabilities for 5 pages(According to Question 3.1.4), X_n is the current probability by changing of steps n. We assume 100 steps and create the graph of error:



From the graph, we can see that the error caused on initial value X_0 is huge but it tend to be zero by increasing steps.

IV. Application: Page Ranking

1. In fact, we will assign the rank for each web page, which consists of the ranks of them. The first page links to all the other pages. And during the page rank, each page will transfer their relevance to other pages with specific numbers. Therefore, we can create a transition matrix of these numbers. In this case, we can assume the initial rank vector as V. The following pages will increase the relevance of a web page. So the following pages can be expressed as $AV, A^2V, A^3V, ..., A^nV$. And the results of these equations will tend to get an equilibrium value, which is the page rank vector. This vector can show the probability of each page, the page ranking will depend on these probabilities to present the importance of each page which is based on stationary distribution. And the probability of the first page(colorado.edu) is the largest one in all pages.

2. From section 3.4, the final probability will gradually approach the project.pdf. However, not all of the users will use the project2.pdf, in this case, the probability of pages will have round errors with finite calculating steps. So, it is not a reasonable ranking because the probability of project2.pdf is the final result that is not suitable for each user. Therefore, this kind of rank is not reasonable because of the limited range.

V. Conclusion

In this project, we try to use the “Markov Chain” model to analyze the execution of the search engine. This model can tell us each state with the relevance of each other. And the probability of each state can be calculated by the transition matrix. At first, the search engine can create the matrix of all web pages resources. And when we type in the search, the search engine will count the order of the keyword in each web pages. And then, the stationary-distribution-based algorithm will rank the importance or relevance of each of the four pages in the first page from page rank algorithm. It means that if we search for a web page with another web page’s hyperlink, we consider that we can get the single stable fixed point by stationary probability distributions with an infinite number of times for each page. In this case, we can easily get the stationary probability distributions from our transition matrix structure. Therefore, the stationary-distribution-based algorithm will rank the importance of each relevant pages from the first page, then, users can get the results that they need after ranking.