IJAMML 7:1-2 (2017) 23-38 September and December 2017
ISSN: 2394-2258
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/ijamml_7100121910

THE RANKING OF CHINESE AIRPORTS

Zhiwei Dou, Peichang Guo, Zipeng Zhou and Sixue Zhang

School of Science, China University of Geosciences, Beijing, 100083, P. R. China

Abstract

Google created PageRank to determine which pages are important. The mathematics of PageRank, however, are general and apply to any network. In this paper, we use Google's PageRank method to evaluate the importance of Chinese airports via the airlines' link structure. The idea in this paper is that, a delay occurring in an airport will have impacts on other airports. The airports, whose delays have the bigger impacts on other airports, are considered to be more important airports in our model.

Keywords: PageRank, ranking, Chinese airports.

1. Introduction

Skytrax created the Air Travel review website (www.airlinequality.com) as an independent customer forum, which has become the leading review site for airline, airport and associated air travel reviews. They classify the airports according to the quality of airport facilities and service standards. This system is a good choice for customers to evaluate

*Corresponding author.
E-mail address: peichang@cugb.edu.cn (Peichang Guo).

2010 Mathematics Subject Classification: 15A18.
Submitted by Los Nguyen.
Received December 18, 2017; Revised December 31, 2017

whether they can get a comfortable experience during their travel. But, there is another critical factor for choosing an airport, which is the possibility of their flight to delay. Some details of the algorithms and data currently used by Skytrax are not open and customers' reviews (part of the data used in assessing) may not be objective enough, so it is not always true that the World's 5-Star airports have less possibility to delay than 4-Star airports. Although there have been statistics and rankings for delay rate, it is still confusing what actually make some airports have more chance to delay and whether their conditions are related. It is believed that an airport that has more airlines and throughput can have more chance to be late but we have to check this idea before admit it. On the one hand, it is beneficial if we can find a subjective model to evaluate the detention of airports. On the other hand, although the condition of delay is based on many unpredictable reasons, such as weather, air control and so on, we wonder whether there is a easy and evenly-matched way that can help us find out the relations between probabilities of airports' delay. When we learned about the Google PageRank method, we find out it can be well applied in this problem since airports are exactly linked by airlines in a complex net structure. In fact, when an airport is well operated, a delay occurs in an airport is largely because one of the planes arrive late, which disturbs the airport's original plan. The original queue in the airport to take off and landing will be disturbed, which leads to a delay.

In this paper, we use Google's PageRank method to evaluate the importance of Chinese airports via the airlines' link structure. We consider a random surfer model with a set of nodes (airports), and construct transition probability matrix based on the practical airline data. The teleporting step is designed to model an external factor that influences the importance of each node. It makes the scores of the airports unique and easy to compute. The idea in this paper is that, a delay occurring in an airport will have impacts on other airports. If there is a delay in an important airport, then the number of airports affected

by the delay will be large. We use Google's PageRank method to score the impacts of a delay in different airports. The airport, whose delay has the biggest impact on other airports, is considered to be the most important airport in our model.

The rest of this paper is organized as follows: In Section 2, we introduce the Google PageRank methodology briefly and formulate the ranking problem of the airports. In Section 3, we do numerical experiments to see how the values of parameter α affects the ranking result. We use our method to examine all airports (dominant) in China by recent airline record and give some conclusions in Section 4. Finally, we discuss some open problems, which is left for our future work.

2. Airport Ranking Based on Google PageRank Method

The Google PageRank methodology was used to determine the importance of web pages at first. In that model, a random surfer, with probability α, randomly transitions according to the link structure of the web, and with probability $1 - \alpha$ teleports according to a teleportation distribution vector ν, where ν is usually a uniform distribution over all airports. In this paper, we see an airport as a random surfer and we replace the notion of "transitioning according to the link structure of the web" with "transitioning according to a stochastic matrix P". This simple change abstracts the mathematics of PageRank from the web and forms the basis of our airport ranking based on Google PageRank method. Furthermore, the vector ν is a critical modelling tool for centrality use that will resemble a uniform distribution over all possibilities [1].

Before introducing the definition formally, let us first introduce some notation: The matrix I is the identity matrix. The vector e is the column vector of all ones, and all vectors mentioned in this paper are column vectors.

The importance of web pages could be determined through the answer of this linear problem [1]:

Definition 2.1 (The PageRank problem)**.** Let P be a column-stochastic matrix where all entries are nonnegative and the sum of entries in each column is 1. Let ν be a column-stochastic vector $(e^T\nu = 1)$, and let $0 < \alpha < 1$ be the teleportation parameter. Then the PageRank problem is:

$$(I - \alpha P)x = (1 - \alpha)\nu, \tag{2.1}$$

where the solution x is called the PageRank vector.

Thus x is the PageRank of airports we want and we use them for ranking all airports. Then the question is how to get the matrix P for all airports**.** We define an adjacency matrix (A) of the net graph where the vertex set is $V = \{1, \ldots, n\}$ first:

Definition 2.2. A is an $n \times n$ matrix where $A_{i,j}$ is 1 if there is an edge from node j and i and zero otherwise. The graph is directed since airlines are different between two airports, in which case A is non-symmetric. The graph is also weighted due to the number of airlines between two airports, in which case $A_{i,j}$ gives the positive weight of edge (j, i). Nodes with zero weight are assumed to be irrelevant and equivalent to nodes that are not present.

We use the standard construction of PageRank matrix, so the matrix P represents a uniform random walk operation on the graph with adjacency matrix A. When the graph is weighted, the simple generalization is to model a non-uniform walk that chooses subsequent nodes with probability proportional to the connecting node's weight. The elements of matrix P are the probabilities of taking the transition from i to j via a random walk step, which is:

$$P_{(i,\ j)} = \frac{A_{i,j}}{\sum_k A_{k,j}}. \tag{2.2}$$

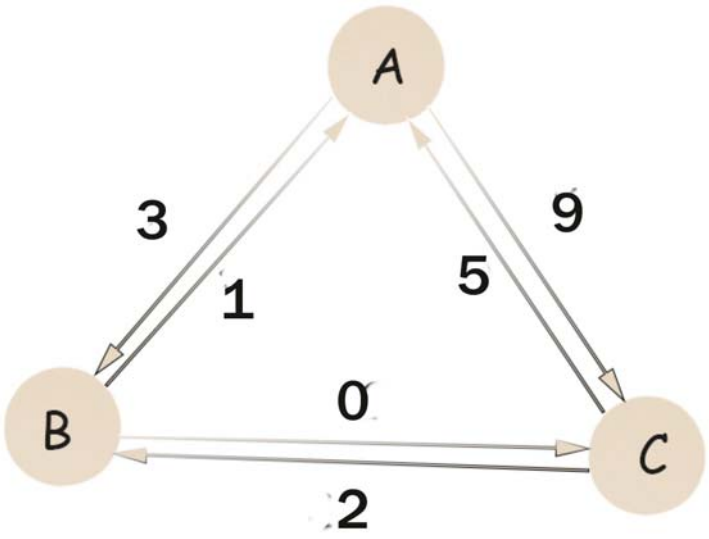

Figure 1. The graph present an example of 3 nodes (airports) with number of airlines between them.

$$A = \begin{bmatrix} 0 & 1 & 5 \\ 3 & 0 & 2 \\ 9 & 0 & 0 \end{bmatrix}$$

$$P = \begin{bmatrix} 0 & 1 & \frac{5}{7} \\ \frac{1}{4} & 0 & \frac{2}{7} \\ \frac{3}{4} & 0 & 0 \end{bmatrix}$$

Figure 2. The two matrix are derived from the link shown in Figure 1.

In our model, $A_{i,j}$ represents the airline number from airport j to airport i. Back to our question, we list all the codes of the 219 airports in China first and sort them from A to Z. This queue is the order (i) to construct the matrix A. For each airport i, we use $A_{i,j}$ to denote the

number of airlines from airport j to airport i during a week. Then we use A to calculate the transition matrix P through the simple formula (2.2). The two figures above (Figure 1 and Figure 2) show a simple example of the construction procedure.

Now we attain the Equation (2.1) we need to solve. Although there are various algorithms to solve this question, we choose the GMRES method, which is a fast and accurate method to deal with such a sparse matrix. It is a Krylov subspace method that rely on the Arnoldi process. Basically, we construct an $m \times m$ upper Hessenberg matrix H_m with the entries $h_{i,j}$ and an orthonormal basis $V_m = \{\nu_1, \ldots, \nu_m\}$. Since the largest eigenvalue of a Pagerank matrix is 1, the Arnoldi-type algorithm seeks an unit norm vector $x^A \in K_m(A, \nu_1)$ satisfying [2]

$$\begin{aligned} \left\|(A - I)x^A\right\|_2 &= \min_{u \in K_m(A, \nu_1),\, \|u\|_2 = 1} \|(A - I)u\|_2 \qquad (2.3) \\ &= \min_{y \in C_m, \|y\|_2 = 1} \|(A - I)V_m y\|_2 \\ &= \min_{y \in C_m, \|y\|_2 = 1} \|(H_m - I_m)y\|_2 \qquad (2.4) \\ &= \sigma_{\min}(H_m - I_m), \end{aligned}$$

and that is the answer of the equation. The detailed algorithm for computing PageRank vector is given in the Appendix.

3. Asymptotics at Large and Small α

The expected result of every airport represent its impacts on other airports when the airlines of the airport is late. Ideally, we use this value as the importance of an airport and rank all airports together.

Though there are widely used regularizers with a variety of established optimality models and results, PageRank is then a strategy of regularizing the importance of nodes. We view (2.1) through the perspective [1].

$$\text{PageRank} = \alpha(\text{the graph} \cdot \text{PageRank}) + (1 - \alpha)(\text{the regularizer}).$$

If α is small, then we depend almost entirely on the regularizer to determine the solution, whereas as α becomes larger the effect is decreased. However, it is only in the limit as α draws asymptotically close to 1 that the effect of regularization goes away. Most of the studies use values of α in the range 0.5 to 0.99 that incorporate a great deal of the regularized effect into the solution. This occurs in most of the uses of PageRank: it protects the ordering against strange outliers in the graph. This regularization view then leads to one of the persistent questions about regularization: how much should we regularize? In terms of our problem, the question is: what should α be? We have to clarify the effects of selecting a given α.

The PageRank vector is a rational function of α and its sensitivity becomes extreme as $\alpha \to 1$ [4]. Furthermore, regularization would argue that α should lie away from 1. Thus, there are reasons that α should not be too big. A simple analysis of the PageRank equation shows that if α is too close to 0, the vector will contain little information beyond the regularization term. Thus, α should not be too small either. The values $\alpha = 0.85$ and $\alpha = 0.5$ may be suitable in most conditions. They are compromises that reflect reasonable choices in order to observe the beneficial regularization effects. Although the conditions of delay can always be changed, a more regular α is preferred to study an objective rule. To compare, we list the on-time performance rank of twenty airports whose throughput is over twenty million in October, 2017 first (Table 1) [3].

(The rankings are based on real statistics and may vary from time to time. Although this is the latest information we have, we have to admit that there are some minor deviations.)

Table 1.

Rank	Airport	On time performance (%)
1	XMN	59.39
2	NNG	63.74
3	SHE	66.63
4	FOC	68.37
5	HRB	69.27
6	PEK	70.35
7	KWE	70.45
8	HGH	71.48
9	WUH	72.65
10	NKG	73.11
11	SYX	74.55
12	PVG	74.78
13	CAN	75.4
14	CGO	75.75
15	TAO	75.96
16	HAK	76.7
17	TSN	76.89
18	XIY	77.7
19	CSX	78.35
20	SHA	79.18

The following figure shows the distribution of these data (Figure 3) (the horizontal axis present the throughput of the airport).

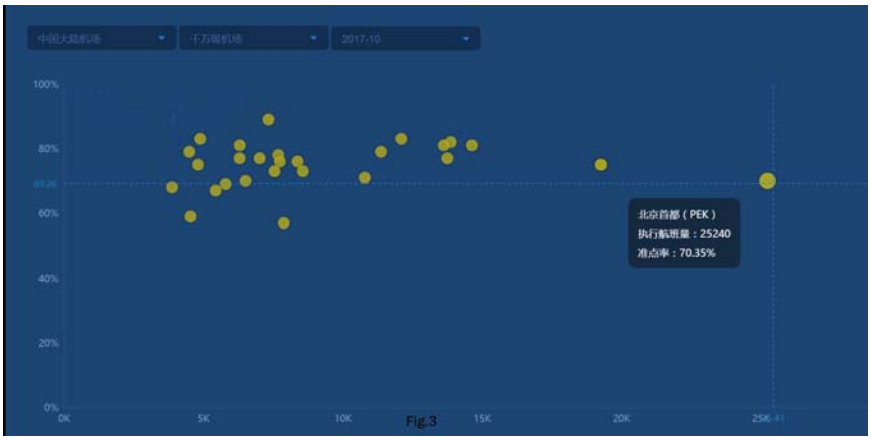

Figure 3.

6 different α is tested: 0.15, 0.5, 0.75, 0.8, 0.85, and 0.9.

The following table shows the ranking of 20 most important airports calculated based on these α (Table 2).

Table 2.

Rank\α	0.9	0.85	0.8	0.75	0.5	0.15
1	PEK	PEK	PEK	PEK	PEK	CAN
2	CAN	CAN	CAN	CAN	CAN	URC
3	CKG	CKG	CKG	CKG	XIY	PEK
4	XIY	XIY	XIY	XIY	URC	KMG
5	CTU	CTU	CTU	CTU	CKG	PVG
6	SZX	SZX	SZX	KMG	KMG	XIY
7	KMG	KMG	KMG	SZX	CTU	XMN
8	PVG	PVG	PVG	PVG	PVG	HGH
9	SHA	URC	URC	URC	SZX	CGO
10	URC	SHA	SHA	SHA	HGH	CTU
11	HGH	HGH	HGH	HGH	SHA	HAK
12	XMN	XMN	XMN	XMN	XMN	NNG
13	DLC	DLC	DLC	DLC	TSN	CKG
14	HRB	HRB	TSN	TSN	HET	INC
15	KWE	TSN	HRB	HRB	DLC	AXF
16	TSN	KWE	KWE	KWE	CGO	HET
17	TAO	TAO	TAO	TAO	HRB	LZY
18	NKG	NKG	NKG	NKG	KWE	YBP
19	CSX	CGO	CGO	CGO	NKG	JGN
20	CGO	CSX	CSX	HET	HAK	KCA

When compared with Table 1, we could not find a perfect match at first. However, when we focus on the scatter diagram (Figure 3), we find that the most important airport is always in the center of the chart. Besides, higher ranked airports always show similar delays. Thus we can tell that the delay of high-importance airports influence other airports. This phenomenon is obvious, especially when we compare the ranking with the following data in September, 2017 [3] (Figure 4 (the horizontal axis present the throughput of airports)). We will discuss our findings in the following section.

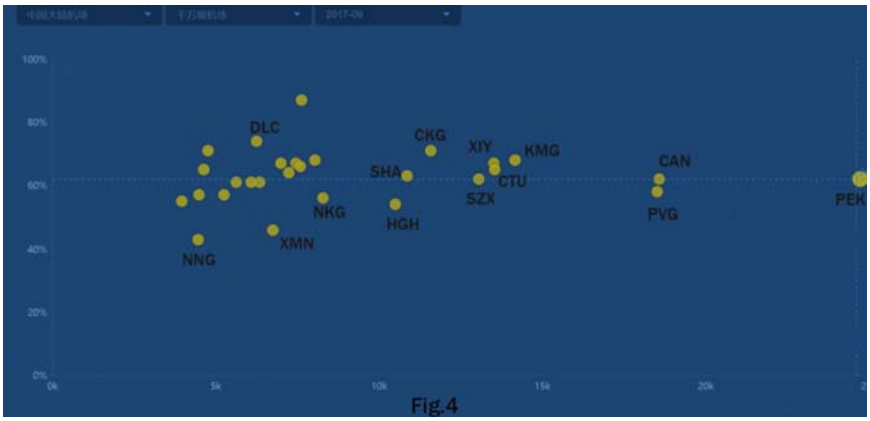

Figure 4.

Back to the question of how to choose α, because it is neither too big nor too small, we prefer to think $\alpha = 0.75$ as a suitable value, either because it matches well with the throughput ranking.

4. Numerical Experiment and Discussion

We have developed algorithms to assess the importance of airports. We investigate all the airports which have run airlines (219) in China. All the airport flights are recorded on a weekly basis. We calculate the value of all airports as follows (Figure 5). The result agree with the throughput of airports. Thus we divide all the airports into 3 groups: throughput over 200 million, from 20 million to 200 million and less than 20 million.

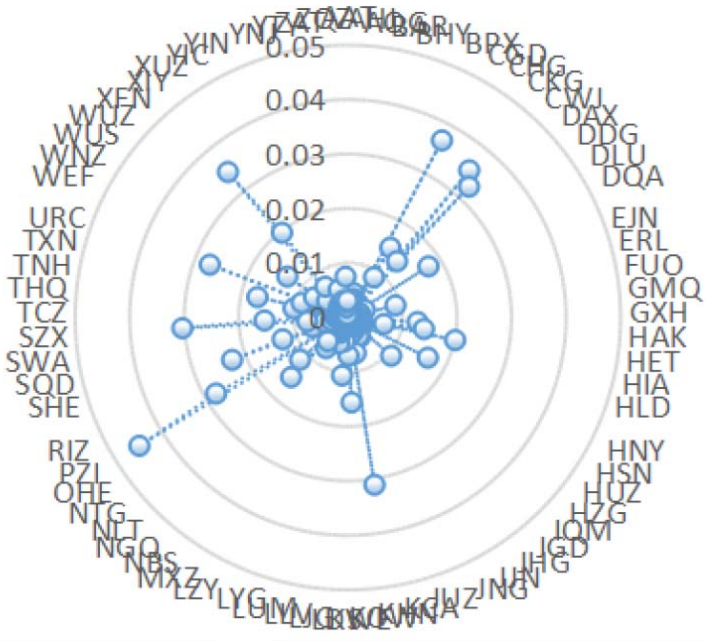

Figure 5.

Then we focus on the distribution of small (< 2 million) and middle-size (2 million ~ 20 million) airports (Figure 6). We notice that the delays at these airports were scattered around the most important airport as we expected. However, some less important airports have a higher probability of being on-time than some important airports, possibly because smaller airports have fewer flights and are less affected by other airports. Therefore, these figures are reasonable and match our rankings.

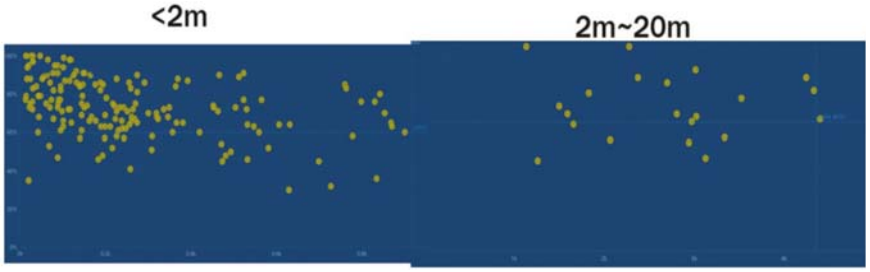

Figure 6.

Besides, we have to mention that the airport URC is top in our ranking and yet away from the central in the diagram. URC has a high on-time ratio. We think it might because that URC locate in a large and remote province in China which is dry all the time. Thus URC rarely delay due to climatic reasons, and the weather is the main reason for the aircraft delays. So we take URC as an exception.

Finally, we can conclude that the ranking reflects whether the delay in the airport will result in another airport detention.

5. Conclusion

The mathematical model is tough now, which needs to be improved in the future work. For example, we need to take the external factors into account more carefully. If some important airports are in monsoon, then the on-time performance of other airports will be influenced. How to model the real world case is a research goal.

Competing Interests. The authors declare that they have no competing interests.

References

[1] D. Gleich, PageRank beyond the web, SIAM Review 57(3) (2015), 321-363.

[2] G. Wu and Y. Wei, Arnoldi versus GMRES for computing PageRank, ACM Trans. Inform. Sys. 28(3) (2010), 1-28.

[3] Data. variflight.com

[online] Available at: https://data.variflight.com/ [Accessed 14 Dec. 2017].

[4] A. N. Langville and C. D. Meyer, Deeper inside PageRank, Internet Math. 1 (2004), 335-380.

[5] A. N. Langville and C. D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, Princeton, NJ, 2006.

[6] D. F. Gleich, P. G. Constantine, A. Flaxman and A. Gunawardana, Tracking the random surfer: Empirically measured teleportation parameters in PageRank, in Proceedings of the 19th International Conference on the World Wide Web, WWW'10, ACM Press, New York (2010a), 381-390.

[7] G. Corso, A. Gulli and F. Romani, Fast PageRank via a sparse linear system, Internet Math. 2 (2006), 251-273.

[8] P. Grindrod and D. J. Higham, A dynamical systems view of network centrality, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. (2014), 470 (2165): 2013083.

[9] G. Jeh and J. Widom, SimRank: A measure of structural-context similarity, in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'02, ACM, New York (2002), 538-543.

[10] D. Koschutzki, K. A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl and O. Zlotowski, Centrality indices in network analysis: Methodological foundations, U. Brandes and T. Erlebach eds., Lecture Notes in Comput. Sci. 3418, Springer, New York (2005), 16-61.

■

Appendix A: Proof of the Problem

Many take the following definition as the definition of PageRank: Let $P_{i,j}$ be the probability of transitioning from page j to page i (or, more generally, from "thing j" to "thing i"). The stationary distribution of the PageRank Markov chain is called the PageRank vector x, which is the solution of the eigenvalue problem:

$$(\alpha P + (1-\alpha)\nu e^T)x = x.$$

The formulation is equivalent to ours if we seek an eigenvector x of (2.1) with $x \geq 0$ and $e^T x = 1,$ in which case

$$x = (\alpha P + (1-\alpha)\nu e^T)x = \alpha P x + (1-\alpha)\nu \Leftrightarrow (I - \alpha P)x = (1-\alpha)\nu.$$

Thus we choose (2.1) which seems easier for calculating.

Appendix B: Algorithm

(1) Start: Given a random vector ν_1 satisfying that $e^T \nu_1 = 1,$ and then $\nu_1 = b$, $\beta = \|\nu_1\|_2$, and $\nu_1 = \nu_1/\beta$;

(2) Iterate: (2.1) for $j = 1, \ldots, m$

(2.2) $C = A\nu_j$;

(2.3) for $i = 1, \ldots, j$

(2.4) $h_{i,j} = \nu_i^T q$;

(2.5) $C = C - h_{i,j}\nu_i$;

(2.6) end for

(2.7) $h_{j+1,j} = \|C\|_2$;

(2.8) if $h_{j+1,j} = 0$

(2.9) break;

(2.10) end if

(2.11) $v_{j+1} = C/h_{j+1,j}$;

(2.12) end for

(3) Compute y_m, the minimizer of $\|\beta e_1 - H_m y\|_2$;

(4) Compute approximate vector: $x = U_m \cdot y_m$.