The Best Coach Selection Based on Cosine Similarity and Improved PageRank Algorithm

WeiHua Wang[1, a *]

[1]North China Electric Power University Baoding 071000, China

aemail:15175289765@163.com

Keywords: Entropy Method; Cosine Similarity; Improved PageRank Algorithm; Time Line Horizon; The Best Coach.

Abstract. In order to find the best college coaches throughout the 20th century, this paper constructs the selection model based on the data of the American college football coaches, and screens the data fundamentally to select the candidates. Then the preliminary ranking is gained via Entropy Method. Considering the effect of the time line horizon, a comprehensive selection model based on Cosine Similarity Method and Improved PageRank Algorithm is built to eliminate the time factor. Compared the preliminary ranking with the integrated one, the result shows that the time line horizon will affect the evaluation and the comprehensive model is effective.

Introduction

The best coach refers to the head coach who commands the athletes won the champion in the considerable level of international competitions[1]. The coach has huge impact on the professional level of athletes. Therefore, the establishment of a scientific model to choose the best coach is essential.

For the sake of constructing a reasonable model to choose the best coach, a logical index set is built and the large amount of data are preprocessed to select the candidates. Entropy Method[2] is an approach to determine the weight of each index. It can eliminate the interference of human factors and make the evaluation results more scientific. Accordingly, it is used to rank the 10 candidates preliminarily by comparing the coaches' score obtained with entropy.

To eliminate the effects caused by the time line horizon, the annual evaluation model is built. The parameter values of the ideal coach annually are defined firstly. Then every coach's index values are mapped to a high-dimensional vector space and the Cosine Similarity method is used to compare the similarity of coaches and the ideal coach to get the annual rankings.

The PageRank Algorithm was invented originally by Google to rank the web pages. This paper offers an improved PageRank Algorithm based on the annual ranking results above to get the comprehensive ranking. After establishing the topological map between each coach , their PR values are calculated through the PageRank Algorithm. Comparing them, the ultimate ranking can be get.

The Best Coach Selection Model

To choose the best college coach more sientifically, a sequence of indexes are needed to be selected to evaluate the skill of a coach. Then the data of coaches are collected and preprocessed. The preliminary ranking is get via Entropy Method. And the comprehensive selection model combining Cosine Similarity and Improved PageRank Algorithm is built to eliminate the effect of time factor.

The Evaluation Index Set.

Table 1 The Evaluation Index Set[3]

Symbol	Definition
SOS	a rating of strength of schedule.
Champion index	the total number of champions that the teams obtained during the coaching period of a coach.
W-L	the ratio of the games that the teams won and the games that the team lost during the coaching period of a coach .
SRS	a rating that takes into account average point deferential and strength of schedule
W-G	the percentage of games that the teams coached by a coach won during the coaches coaching period.
Games	the total number of games that a coach coached.
Yrs	the total number of years that the teams coached by a coach.

Data preprocessing and Preliminary Ranking . The data of American college football coaches are collected from sports reference[4]. Screen them to select 361 coaches whose coaching years are more than 10 years. For a coach, the Cames, W-L and Champion index are the vital evaluation standards. So the three indexes are added the same weight and divided into five classes as below.

Table 2 The Number of Games Directed

Games	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Score	1	2	3	4	5

Table 3 Win-Loss Percentage

W-L	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Score	1	2	3	4	5

Table 4 The Championship Index

Champion index	1-5	6-10	11-15	16-20	21-25
Score	1	2	3	4	5

Puting up the score of these classes, the weighted average of each coach is calculated. And the ten candidates of the best coach are chosen.

The Entropy Method[5] is used to determine the weights of indexes based on the degree of information ordering in them. The bigger the Entropy is, the greater the uncertainty of the variable is and vice versa. This paper uses the Entropy Method to rank the coaches preliminarily.

1）The definition of standardizing

$$Y_{ij} = \frac{y_{ij}}{\sum_{i=1}^{m} y_{ij}} \tag{1}$$

2）The Information entropy of index j

$$e_j = -\frac{1}{\ln m}\sum_{i=1}^{m} Y_{ij} \ln Y_{ij} \tag{2}$$

The Information utility value

$$g_j = 1 - e_j \tag{3}$$

3）The weight of the evaluation index W_j

$$W_j = \frac{g_j}{\sum_{j=1}^{n} g_j} \tag{4}$$

The weight of these seven indexes are obtained.

Table 5 The weights of seven indexes

Yrs	G	W-G	W-L	SRS	SOS	Champion ship index
0.21	0.23	0.011	0.0117	0.0964	0.1393	0.3016

4) Evaluate comprehensively and calculate the weighted value of each coach.

$$F = \sum W_j y_{ij} \tag{5}$$

By comparing the weighted values, the preliminary ranking of the ten candidates is get.

Table 6 The preliminary ranking

Rank	Coach	Weighted Value
1	Joe Paterno	0.1534
2	Bobby Bowden	0.139
3	Bear Bryant	0.1234
4	Lou Holtz	0.1002
5	Tom Osborne	0.099
6	Mack Brown	0.0916
7	Bud Wilkinson	0.0672
8	Barry Switzer	0.0787
9	Mack Brown	0.0916
10	Bob Devaney	0.0576

The Annual Ranking Based on Cosine Similarity Measure. To eliminate the effect of the time line horizon, the annual rankings of the coaches is calculated at first. The Cosine Similarity measure is applied to define the distance between two vectors in cluster analysis originally[6]. Through measuring the angle between two vectors, the similarity between them is defined. The cosine value is between -1 and 1. It is 1 when the two vectors point to the same direction. And the exactly opposite direction is -1. This method can be used in any dimension, and it owns the obvious advantages compared with the Mahalanobis distance especially in the high-dimensional space.

The similarity between the two vectors could be expressed by the cosine value of their angle:

$$similarity = \cos(\theta) = \frac{\mathrm{A} \bullet \mathrm{B}}{\|\mathrm{A}\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{6}$$

The greater the cosine value is, the smaller the angle is. And the similarity is greater.

Vector Space Model (VSM) [7,8]is an algebraic model used to characterize a text file. The documents and queries are both represented by vectors. If the term appears in a document, its value called phrases weights in a vector is non-zero. Each text is mapped into the weight vector of text feature. And every characteristic value is assigned a weight to express the importance of it in this text.

The text can be expressed as

$$Z_i = (t_1 \times W_1,\ t_2 \times W_2, \cdots, t_n \times W_n) \tag{7}$$

Therefore the cosine similarity between texts A_1, A_2 is defined as the VSM. The cosine angle is

$$S(V_{Z_1}, V_{Z_2}) = \frac{V_{Z_1} \bullet V_{Z_2}}{\|V_{Z_1}\| \bullet \|V_{Z_2}\|} \tag{8}$$

The VSM is established, and researchers collect the seven indexes value of the 10 candidates in the 20th century. Coach A is regarded as an example to illustrate the method.

1) Normalize the seven indexes and multiply the weights. Assign them to the seven- dimensions.

$$A_i=(t_1\times W_1,\ t_2\times W_2,\cdots,t_7\times W_7) \tag{9}$$

2)Set the index values of the ideal coach which are the maximum ones of the candidates every year.

$$L_{iM}=(t_{1iM}\times W_1,\ t_{2iM}\times W_2,\cdots,t_{7iM}\times W_7) \tag{10}$$

Based on the Cosine Similarity Measure formula, the cosine value of the angle between the seven dimensional vectors corresponding to the 10 coaches and the one to the ideal coach can be calculated.

$$S(A_i,L_{iM})=\frac{A_i\bullet L_{iM}}{\|A_i\|\bullet\|L_{iM}\|} \tag{11}$$

By comparing the size of the angles, the annual rankings are obtained.

The Comprehensive Ranking Based on PageRank Algorithm. For the coaching year of the coaches is different, it is essential to find the contact among them. Using the Graph Theory, each coach can be regarded as a node in a graph. And the numbers marked next to the edges represent the number of years the coach is better than another one.

PageRank Algorithm is designed by Google used to determine the rank of a specific web page via a vast hyperlink network of relationships[9]. The basic idea of it mainly comes from the analysis of the citation in traditional bibliometrical method. And the more a literature is cited by other literatures, the higher the quality of this literature is. The basic definition of PR value is defined as.

$$PR(P)=\frac{1\text{-}d}{m}+d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \tag{12}$$

PR(P) is the PR value of webpage P. m is the totality of the webpage's nodes. d is the attenuation coefficient on the (0,1) interval(usually taken as 0.85). T_i(i=1,2,L,n) is the webpage pointing into Webpage P and chains into it. $C(T_i)$ is the number of links webpage P outward pointing out. $PR(T_i)$ stands for the PR value of Webpage T_i.

The formula is iterated repeatedly until the calculated PR value converges to a fixed number.

Suppose a small group composed by four pages: A, B, C, D. If all the web pages link to A, then the PR value of A is the sum of the B, C, D's PR values.

$$PR(A)=PR(B)+PR(C)+PR(\mathrm{D}) \tag{13}$$

And the PR value of a web page is allocated according to the totality of chaining out.

$$PR(A)=\frac{PR(B)}{L(B)}+\frac{PR(C)}{L(C)}+\frac{PR(\mathrm{D})}{\mathrm{L(D)}} \tag{14}$$

Finally, all of these are converted to a form that the percentage is multiplied by a coefficient d .If a web page dosen't link outward, the PR value that it transfers out is 0. Therefore, Google gives each page a minimum value $(1\text{-}d)/N$ through mathematical system.

$$PR(A)=(\frac{PR(B)}{L(B)}+\frac{PR(C)}{L(C)}+\frac{PR(\mathrm{D})}{\mathrm{L(D)}})+\frac{1\text{-}d}{N} \tag{15}$$

With reference to the relationship among the web pages, the topological graph among the ten coaches can be built. The specific relationship is indicated in Fig. 1. And the value of the connection on the line represents the weight between the two coaches.

The PageRank algorithm combined with the weights is used to calculate the PR value of each coach. Assume that the initial PR value of three coaches is 1, then A coach's PR value is:

$$PR(A) = (\frac{PR(C)}{1} + \frac{PR(B)}{k+n} \times k) \times d + \frac{1-d}{N} \qquad (d=0.5, N=10) \qquad (16)$$

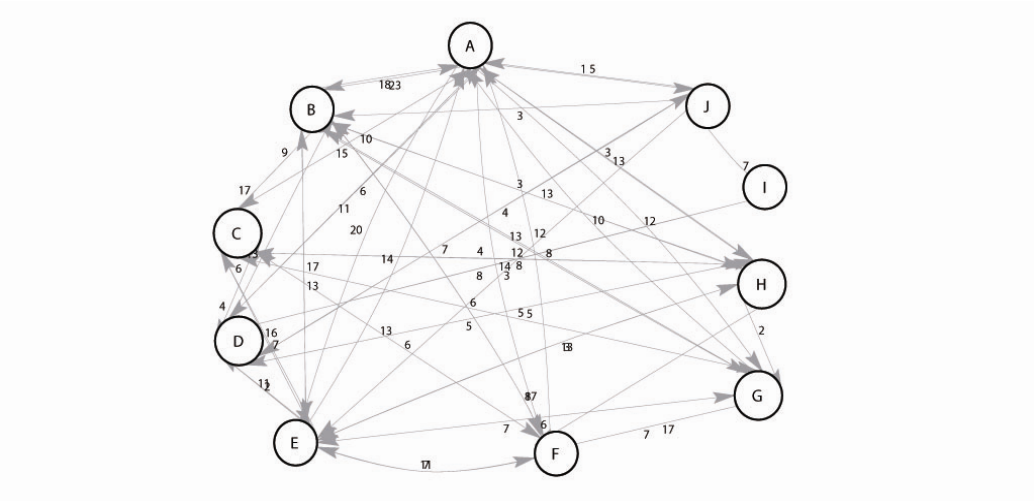

Fig. 1 The Topological Graph among the Ten Coaches

Experimental Analysis

The comprehensive ranking is gained according to the PR values calculated as above:

Table 7 The Comprehensive Ranking

Rank	Coach	PR	Rank	Coach	PR
1	Bear Bryant	1.4585	6	Lou Holtz	0.9832
2	Joe Paterno	1.2596	7	Bud Wilkinson	0.8419
3	Tom Osborne	1.1447	8	Steve Spurrier	0.7815
4	Bobby Bowden	1.0822	9	Mack Brown	0.7813
5	Barry Switzer	1.0084	10	Bob Devaney	0.6587

Comparing the preliminary ranking with the comprehensive one, it is obvious that time line horizon does make an effect to the evaluation[10].

Summary

In the study, Entropy Method is applied firstly to rank the ten candidates which are obtained after preprocessing the data. To eliminate the effect of the time factor, a comprehensive evaluation model is established, which combines the Cosine Similarity method to get the annual rankings and the PageRank Algorithm to get the ultimate ranking. Compared the preliminarily ranking with the comprehensive one, the result shows that the time line horizon indeed has effect to the evaluation result and the comprehensive selection model is effective to choose the best coach throughout the 20th century.

References

[1] Information on http://en.wikipedia.org/wiki/Best_Coach/Manager_ESPY_Award

[2] Information on http://baike.baidu.com/view/2523301.htm

[3] Information on http://www.sports-reference.com/cfb/coaches/ike-armstrong-1.html

[4] Information on http://www.sports-reference.com/cfb/coaches/
[5] Information on http://www.baike.com/wiki/%E7%86%B5%E5%80%BC%E6%B3%95
[6] Information on http://en.wikipedia.org/wiki/Cosine_similarity
[7] Q.L. Guo, Y.M. Li , Q. Tang: Application Research of Computers, Vol. 25 (2008) No.11, P.3256.
[8] Information on http://blog.163.com/wyy_820211/blog/static/22922408200821911503061 1/
[9] S.Y. Wu, T. Xu: Library and Information Service, Vol. 47 (2003) No. 2, P.55.
[10] Information on http://www.sports-reference.com/cfb/coaches/