Implementation of School-Enterprise Two-Sided Matching Based on Pearson and Top Trading Cycle

Meng Zhu, Ben Wang and Quanyin Zhu*
Faculty of Computer & Software Engineering, Huaiyin Institute of Technology, Huaian, China
*Corresponding author’s email: hyitzqy@126.com

Abstract. Due to a large number of graduates and employment information, two-sided matching is increasingly used in employment. However, the general recruitment websites are not rich in student data, so their recommendation will not be very accurate. In order to improve the efficiency of recommendation, a recommendation system based on Pearson and Top Trading Cycle is designed by using the real graduate data of our school. Through the application of Pearson and TTC (Top Trading Cycle), the stable match between the students and the enterprise is realized. We use Pearson similarity algorithm to calculate the similarity between students and sort them from large to small, then we arrange the similarity of order as user preferences in TTC algorithm. Take each of the students as the starting point and the end point, the similarity as the evaluation criteria, and the rest of the students to form a ring. We believe that all the enterprises of the past students in the ring area are the recommended enterprises to be allocated to the students in the ring area, and the students who are to be assigned are recommended students recommended by them.

1. Introduction
In the recent years, the number of graduates and the number of recruitment information are more and more. As a new medium of information dissemination, the Internet can provide a completely new and efficient model for graduate employment. In 1990s, as a result of the “information overload” problem, the recommendation system was born, such as Google email filtering [1]. Currently, the mainstream recommendation methods include collaborative filtering-based recommendation [2], content-based recommendation [3], utility-based recommendation [4], association rule-based recommendation [5], combined recommendation [6] and knowledge-based recommendation [7].

The research of two-sided matching decision making problem originates from university admission and stable marriage assignment problem [8]. Alvin E. Roth et al. conducted an in-depth study on the game analysis of two-sided matching [9].

However, all the approaches above are not applicable to the specific data of the school. Because the school’s students are rich in data, and the company’s data is only the name of the enterprise. According to the above problems, we propose a new algorithm combined Pearson correlation coefficient with TTC (Top trading cycle) algorithm. From the experiments, we find the algorithm we propose in this paper can solve the problem effectively.

The structure of this paper is as follows: section 1 introduces the recent situation about recommendation and the situation about two-sided matching decision making problem; section 2 introduce the calculation method of similarity; section 3 introduces TTC algorithm; section 4 states our experiments and the relative evaluation results; section 5 draws conclusions and future work.
2. Similarity Algorithm

2.1. The Concept and Principle of Similarity Algorithm

In data analysis, data mining and search engines, we often need to know the size of the differences between individuals, and then evaluate the individual’s similarities and categories. Common data analysis of the relevant analysis, data mining classification clustering (e.g. K-Means) algorithm, search engine article recommendation.

Similarity is the comparison of the similarities between two things. Generally by calculating the distance between the features of things, if the distance is small, then the similarity is large; if the distance is large, then the similarity is small. For example, two kinds of fruit, will be from the color, size, vitamin content and other features to compare the similarities.

Problem definition: There are two objects, X and Y, which all contain N-dimensional features, \( X = \{x_1, x_2, \ldots, x_n\}, \ Y = \{y_1, y_2, \ldots, y_n\} \), calculate the similarity between X and Y. There are five commonly used methods, as follows:

2.1.1. Euclidean Distance. The formula is

\[
u_i = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

where \( x, y \) are the features of X, Y. Euclidean distance is the most commonly used distance calculation formula, which measures the absolute distance between various points in multi-dimensional space, which is a good method when the data is very dense and continues. Because the calculation is based on the absolute values of the features of each dimension, the Euclidean metric needs to ensure that each dimension index is at the same scale level. For example, the use of the Euclidean distance for two different units of height (cm) and body weight (kg) may invalidate the result.

2.1.2. Manhattan Distance. The formula is

\[u_i = \sum_{i=1}^{n} |x_i - y_i|
\]

where \( x, y \) are the features of X, Y. The Manhattan distance, a term developed by Herman Minkowski in the 19th century, is a geometric term used in geometric metrology spaces to indicate the sum of the absolute wheelbase of two points on a standard coordinate system. Suppose, consider the two-dimensional situation first, only two bands \( x \) and \( y \), user A’s rating is \( (x_1, y_1) \) and user B’s rating is \( (x_2, y_2) \), then the Manhattan distance between them is \( |x_1 - x_2| + |y_1 - y_2| \).

2.1.3. Cosine Similarity. The formula is

\[ \text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \]

where A, B are two vectors, \( \theta \) is the angle. Cosine similarity uses the cosine of the angle between two vectors in vector space as a measure of the difference between two individuals. Compared to distance metrics, the cosine similarity focuses more on the difference in direction of the two vectors than the distance or length. The value of the cosine similarity is in the range [-1, 1], and the larger the value, the more similar it is. Cosine similarity is widely used in collaborative filtering algorithms, especially Itembase collaborative filtering.
2.1.4. Jaccard Similarity. The formula is

\[ Jaccard(X,Y) = \frac{X \cap Y}{X \cup Y} \]  

(4)

Jaccard coefficient is mainly used to calculate the similarity between individual measures of symbol metrics or Boolean metrics, because individual feature attributes are identified by symbol metrics or Boolean values, so it is impossible to measure the size of specific values of differences and only obtain “whether the same”. This result, so the Jaccard coefficient only concerns the problem of whether the features shared by individuals are consistent. The larger the Jaccard coefficient value is, the higher the sample similarity is.

2.1.5. Pearson Correlation Coefficient. The formula is

\[ r(X,Y) = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}} \]  

(5)

where x, y are the features of X and Y and n is the sample size. The Pearson correlation coefficient is a linear correlation coefficient. The Pearson correlation coefficient is a statistic used to reflect the degree of linear correlation between the two variables. \( r \) describes the degree of linear correlation between two variables and the value of \( r \) is in the range \([-1, 1]\). If \( r > 0 \), it indicates that the variables are positively correlated. That is, the larger the value of one variable is, the larger the value of the other variable is. If \( r < 0 \), it indicates that the two variables are negatively correlated, that is, the greater the value of one variable, the smaller the value of the other variable. The larger the absolute value of \( r \) is, the stronger the correlation is. It is important to note that there is no causal relationship. If \( r = 0 \), it means that there is no linear correlation between the two variables, but there may be other ways of correlation (e.g. the curve).

2.2. Pearson Similarity Algorithm

We adopt pearson correlation coefficient algorithm in the paper, which is one of the most popular similarity algorithms. The specific ideas as the following:

(a) Quantify each student’s information and store the information as a collection.

(b) The student ID is used as a key, and the corresponding student information set is taken as a value, so that each student information set is stored as an element in the set A.

(c) Calculate the similarity of any two students in set A and store the two students’ school numbers and the similarity coefficient as a set.

2.3. Students Similarity

In this paper, students are divided into two categories. One is fresh graduates, and the other is employed former students.

Usually, students in a school are divided into different professions. Students in the same major will have different learning orientations, and there will also be differences in student professional competence in the same learning direction. There is no doubt that academic performance and awards can be used to assess a student’s professional ability. Based on this principle, we ensure that we can correctly calculate the degree of similarity and persuasiveness of professional abilities of recent graduates and former students.

Considering the discrimination and easily acquired of the attributes, we draw a conclusion that there are eleven attributes (android, asp, cpp, java, linux, database, os, data structure, .net, math, others) playing a decisive role to choose the items. We put the three attributes as student characteristics to calculate similarities among students. The specific calculation formula defined as:
\[ r(X,Y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}} \]  

where \( x, y \) are the features of \( X \) and \( Y \) and \( n \) is the sample size and \( r(X,Y) \) represents the similarity coefficient between \( X \) and \( Y \). See Figure 1.

![Figure 1. Data format of student similarity](image)

### 3. TTC-based Tow-Sided Matching Algorithm

#### 3.1. Top Trading Cycle (TTC)

The TTC algorithm was first proposed by Shapley & Scarf in 1974. But the two authors discussed the housing problem in this article, and they argued that there is always a core in the housing problem. David Gale proposed the TTC algorithm, and this algorithm belongs to the core, so the TTC algorithm is also called Gale’s Top Trading Cycle algorithm.

We use an example of housing allocation to find out how the TTC algorithm works. Assume that there are 16 people, each holding a house. The set of 16 people is:

\[ A = \{a_1, a_2, a_3, \ldots, a_{16}\} \]  

and corresponding house set for each person is:

\[ H = \{h_1, h_2, h_3, \ldots, h_{16}\} \]  

Everyone has their own preference for all houses (including their own house). Step 1: Everyone points to their favorite house with an arrow, and the house points to the owner of the house. In this cycle, which person points to which house, the house is assigned to the person pointing to the house, and the people and houses in the ring are removed from the sets \( A \) and \( H \). Step 2: Repeat Step 1 until all people and houses are assigned. Everyone has their own preferences, see Figure 2. The final result is shown in Figure 3. They achieved a stable match and were more efficient.

![Figure 2. User's preference for the house](image)
3.2. TTC-based Two-sided Matching Algorithm

Usually, a university will store information on the employment of previous students, but often do not have the recruitment information and business information at that time. There are two main ways of two-sided matching between students and enterprises. One is the method that both parties actively choose, and the other is to have the detailed information of students and enterprises and use matching algorithms such as clustering, association rules, and collaborative filtering. However, the first method is applicable to the distribution of work, which is not applicable to the situation where students are free to find work. The second method has the problem of data fraud. Like most recruitment websites, their information is filled in by job applicants and companies themselves and cannot guarantee the truth of the information.

For the case of missing enterprise data, we propose a method based on the TTC algorithm to solve this problem. This algorithm is based on student’s objective data, such as school grades, awards, and so on, to calculate the similarity between fresh students and previous students. Instead of subjective selection, similarity is applied to the TTC algorithm.

Algorithm:

Input: Original student information matrix R (R is a m×n two-dimensional matrix, m presents for the number of students, n presents for the number of items and the unrated items are filled with 0). Original student set S (S stores students’ unique identification tags).

Output: The matrix PR which is full filled with the fresh graduates and recommended enterprises for them.

(d) Calculate the similarity $r_{ij}$ between each student and store in the set OR (OR is a set and the first item is $S_i$, the second item is $S_j$, the third item is $r_{ij}$).

(e) Create the new sets $R_i$ and SR (SR is a set that contains a student’s all OR sets. $R_i$ is a set of key-value pairs, key is the student’s unique identity, and value is the set SR of a student).

(f) Traverse $R_i$ and sort the set SR in each element of $R_i$ in descending order, and get new sets $R_i'$ and $SR'$, and create a new set C.

(g) Traverse S and find $SR'$ based on $S_i$, and add $SR'$ to set C.

(h) When the second item of $C_i$ is not $S_j$, perform step (f); otherwise perform step (j).

(i) When the number of elements in the set C is less than 30 and the second item in $C_i$ is not $S_j$, perform step (g); otherwise perform step (h).

(j) According to the second item in $C_i$, the $SR'$ is found, and i is increased by 1 and $SR'_{jm}$ is added to the set C. Perform step (f).

(k) When the second item of $C_i$ is not $S_j$, perform step (i); otherwise perform step (j).

(l) For each element in the set C, find their $SR'$ set according to the first item of the element, and remove $SR'_{jm}$ from $SR'$ and add the set U. The element $U_k$ with the largest similarity coefficient in the set U is taken out, $U_k'$ is inserted at the k position in C, and all the elements in C after k are removed. Perform step (e).

(m) For each element in the set C, it is divided into set A and set B according to the first item of the element (The set A is a set of fresh students, and the set B is a set of previous students). For each element in set B, we get the previous student employment enterprises according to the first item of the element, and recommend the enterprises to the fresh students represented by the first item of the element in set A.

(n) According to the recommendation of the new students, the corresponding students are recommended to the companies.
4. Experiments and Analysis
In all the evaluation criteria, one of the important evaluation criteria for two-sided matching is the success rate of matching. In this paper, we believe that all companies recommended for student are successful if the student joins the company.

We adopt the dataset from the HuaiYin Institute of Technology, including students’ school grades, awards, and entry companies. In the dataset, there are three data sheet including student grade sheet, student award sheet and student employment sheet.

After many experiments, it shows that all students will get recommended companies. Considering comprehensively, the total number of companies we recommend will not exceed five, but there is no upper limit in the experiment, show in Figure 4.

![Figure 4. Statistics of the number of people who are recommended](image)

Because of the special nature of the data, we have no extra plans for comparing experimental results. However, we have divided the existing data into two parts, one part accounts for 80% of the total data, and used it as experimental basic data. The rest is used as test data to verify the matching accuracy of this solution.

According to the previous method of calculating the correct rate, experiments show that the success rate of matching is above 22.4%. The results are as we expected, because most students choose different companies. For the specificity of the data, we propose that as long as there is a recommendation company, the recommendation is considered successful, and everyone has at least one recommendation company. This scheme has a good practical significance.

5. Conclusion
With more and more graduates and a large number of recruitment information, it has become an urgent problem for graduates to find the right job quickly and accurately. In the absence of enterprise data in the school, we improve the TTC algorithm according to the actual situation, so that each student has his own personalized recommendation. First, we quantified the student data and calculated similarities among students through Pearson's algorithm because students' professional abilities are closely related to their academic performance and awards. Secondly, in order to solve the problem of missing enterprise data, we adopted the idea of the TTC algorithm and effectively solved the problem that the traditional bilateral matching system requires both parties' detailed data. Finally, through the data in the experiment, it is proved that the bilateral matching system can have good recommendation quality for school-specific data.
6. References

[1] Schafer J B, Konstan J, Riedl J. Recommender systems in e-commerce [C]//Proceedings of the 1st ACM conference on Electronic commerce. ACM, 1999: 158 - 166.

[2] Yang B, Lei Y, Liu J, et al. Social collaborative filtering by trust [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39 (8): 1633 -1647.

[3] Musto C, Semeraro G, de Gemmis M, et al. Learning word embeddings from wikipedia for content-based recommender systems [C]// European Conference on Information Retrieval. Springer, Cham, 2016: 729 -734.

[4] Li Z, Fang X, Bai X, et al. Utility-based link recommendation for online social networks[J]. Management Science, 2016, 63 (6): 1938 -1952.

[5] Moonen L, Di Alesio S, Binkley D, et al. Practical guidelines for change recommendation using association rule mining [C]// Automated Software Engineering (ASE), 2016 31st IEEE/ACM International Conference on. IEEE, 2016: 732 -743.

[6] Kong X, Xia F, Wang J, et al. Time-location-relationship combined service recommendation based on taxi trajectory data[J]. IEEE Transactions on Industrial Informatics, 2017, 13 (3): 1202-1212.

[7] Jung H, Chung K. Knowledge-based dietary nutrition recommendation for obese management [J]. Information Technology and Management, 2016, 17(1): 29 - 42.

[8] Gale D, Shapley L S. College admissions and the stability of marriage [J]. The American Mathematical Monthly, 1962, 69 (1): 9 - 15.

[9] Roth A E, Sotomayor M. Two-sided matching [J]. Handbook of game theory with economic applications, 1992, 1: 485 - 541.