A Friend Recognition Method Based on Campus Card Data

Yinghao Chen¹, Boyong Wu² and Jie Zhang³

¹Northeastern University, Shenyang City, Liaoning Province, China
²Northeastern University, Shenyang City, Liaoning Province, China
³Northeastern University, Shenyang City, Liaoning Province, China

¹1481626303@qq.com, ²474157449@qq.com, ³762468972@qq.com

Key words: big data, the identification of friendship, friend network, college education

Abstract: With the increasingly rapid development of information age, we ushered in the era of big data[1,2]. It is attractive to many experts and scholars to explore the mystery of human life and society with the big data. Aiming at the identification of friend relationships, a new method based on campus cards and big data has been proposed which includes four steps, namely data preprocessing, generation of similar behavior matrix, determination of similar behavior threshold and identification of friend relationships. And compared to the traditional questionnaires, this method is more practical and instructive. Furthermore, it is found that 45703 pairs of students have been identified as the friends to each other through the campus card data in library. Besides, the friend network diagram has been built, based on which education administrators could better guide students’ campus lives and some businesses could advertise more precisely and efficiently.

1. Introduction

“That friends should come to one from afar, is this not after all delightful?” At all times and in all over the world, friends play an extremely important role in the growth and development of a person. In the information age of today, a person's circle of friends not only has an impact on its own development, but also has a great influence on many things such as the expansion of some businesses and the spread of public opinion. Now assume you are the manager of a fitness business, and you need to attract new users for the company by old customers, then which old customers will you focus on for the advertising? Besides, if you want to develop student users, then how do you expand the consumer group, quickly and effectively, through the relationship between these students? What is exciting is that if we know these people's circle of friends, we will know the amount of their influence. Then the formulation of these programs[3,4] will become much easier. Therefore, it is necessary to establish a stable, efficient and large-scale friend identification method. A large number of existing articles are based on surveys and online registration information to identify friendships[5,6,7,8,9]. Although these methods can have guiding significance, to some extent, for social research and enterprise development, there are three obvious deficiencies through specific analysis. Firstly, there are always some false elements of the registration information filled by users so that it can hardly reflect their own real information. Secondly, a user's online behavior does not really reflect his real-world friend relationships. However, it is only the real-world friend relationships that can help them expand the customer source effectively, such as banking, supermarkets, gyms and other service places. Finally, it is universally acknowledged that some of the friendships obtained on the online virtual platform are weak or geographically far apart. Obviously, these so-called friendships do not exist or are not in a particular group. Therefore, there is little practical significance for our research.

Thanks to the emergence of smart devices and the generation of high-performance computers, we have access to a large number of student behavior data and have the ability to analyze and process these data. In China's campus, almost all of the students have a campus card that can be used for identification and consumption. The data recorded by the smart cards can almost reflect a student’s behaviors in all aspects of their school lives. For example, based on the massive data, we can easily get the number of times of a student’s visits to the library during a week, and the specific
and exact time at which a student leaves or returns to dormitory. In addition, we can also know whether a student eats breakfast or not even including what he or she has eaten within a week.

Therefore, according to the real reflection of each student's daily behavior, we can dig deeply into the behavioral characteristics of college students. And a person's network of relationships is usually reflected in a person's behavior. Those close friends often produce similar behavior, such as access to the dormitory, access to the library, having dinner and so on. Therefore, accordingly, this article will use the massive data of students' smart cards to identify the friendships between college students. Obviously, this method based on the massive data will be more accurate and instructive compared to the traditional questionnaires.

2. Materials and Methods

2.1 Data Description

The campus card data collected from the school's academic office is shown in FIG. 1. Each row represents a campus card record, from left to right showing the student's id, the category of student's operations, the time of students swiping card. And each piece of data, not only records the exact time of students swiping card, but also records the time of the cardholder entering or leaving the library and other detailed information. Therefore, based on the information provided by these data, we can construct a network of relationships among these students to identify friendships.

![Figure 1](image1.png)

2.2 Data Preprocessing

2.2.1 Data cleaning

We use 10 data, for example, to clean up the data. Data cleaning includes deleting incomplete data, deleting the quotation marks in the record of student's operations in each piece of data, quotation marks and slashes in the record of the time of students swiping card. After the above processing, the result is shown in FIG. 2, and the processed data is saved to the data warehouse.

![Figure 2](image2.png)

2.2.2 Data Classification

Because in each piece of data, not only the time of students swiping card is recorded, but also the time of the cardholder who is entering or leaving the library. And if two users have friendships, they usually go to the library at the same time or leave at the same time. Therefore, we assume that only two users entering or leaving a library at the same time can be called a similar behavior. Based on this assumption, we classify the data into incoming data and outgoing data, and start two text documents named "library_in" and "library_out". The data in "library_in" and "Library_out" are shown in FIG. 3 and FIG. 4, respectively.

![Figure 3](image3.png)

![Figure 4](image4.png)
2.2.3 Data sorting
Taking "library_in" as an example, the data in the text is sorted by time using the c++ sorting method, and the result is shown in FIG. 5.

<table>
<thead>
<tr>
<th>id</th>
<th>year</th>
<th>mon</th>
<th>day</th>
<th>hour</th>
<th>min</th>
<th>sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>9773</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>09</td>
<td>01</td>
<td>24</td>
</tr>
<tr>
<td>9773</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>09</td>
<td>01</td>
<td>25</td>
</tr>
<tr>
<td>29851</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>09</td>
<td>01</td>
<td>37</td>
</tr>
<tr>
<td>7203</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>09</td>
<td>01</td>
<td>45</td>
</tr>
<tr>
<td>69083</td>
<td>2014</td>
<td>11</td>
<td>16</td>
<td>20</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>19151</td>
<td>2014</td>
<td>11</td>
<td>16</td>
<td>20</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>9773</td>
<td>2014</td>
<td>11</td>
<td>16</td>
<td>20</td>
<td>18</td>
<td>52</td>
</tr>
<tr>
<td>5732</td>
<td>2014</td>
<td>11</td>
<td>16</td>
<td>20</td>
<td>18</td>
<td>31</td>
</tr>
</tbody>
</table>

Figure 5.

2.3 Generation of Similar Behavior Matrix
We assume that when two users swipe campus card in the same location, and the interval between two users swiping campus card is less than 30 seconds, there is a similar behavior. The following describes in detail the generation of similar behavior matrix:

First, we create an empty queue and then traverse the text by line. If the queue is empty, the time information read in the current row will be queued; if there is information in the queue, the time information read in the row will be compared to the time information of the head of the queue. If the time interval is less than 30 seconds, it is regarded as a similarity and output to the "friend" text, and continue to traverse the queue downwards; if the time intervals compared to all the time information in the queue are less than 30 seconds, the time information in the current row will be added to the tail of the queue; if the time interval is greater than 30 seconds, replace the message at that point in the queue with the current line in the text, and continue traversing the text down. Find the time interval between two students swiping their campus card that is not more than 30 seconds to get a pair of friends, and output its two id in ascending order keeping the time information of the friendship. Generate the matrix by the time unit of half month, and the result is shown in FIG. 6. Then the data in the "friend" document is divided into several small texts according to a half-month period, and finally, in an ascending order of id, generate the similar behavior matrix as shown in FIG. 7.

<table>
<thead>
<tr>
<th>id1</th>
<th>id2</th>
<th>year</th>
<th>mon</th>
<th>day</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>9773</td>
<td>9773</td>
<td>2014</td>
<td>11</td>
<td>21</td>
<td>0800</td>
</tr>
<tr>
<td>9773</td>
<td>9773</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>0800</td>
</tr>
<tr>
<td>29851</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>0800</td>
<td></td>
</tr>
<tr>
<td>9773</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>0800</td>
<td></td>
</tr>
<tr>
<td>7203</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>0800</td>
<td></td>
</tr>
<tr>
<td>7203</td>
<td>9773</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>0800</td>
</tr>
<tr>
<td>7203</td>
<td>9773</td>
<td>2014</td>
<td>12</td>
<td>21</td>
<td>0800</td>
</tr>
</tbody>
</table>

Figure 6.

2.4 Determination of Similar Behavior Threshold
Based on the generated similarity behavior matrix, we obtain the data above the diagonal of the similar behavior matrix and generate Pareto charts, as shown in FIG. 8. According to the generated Pareto chart, we take the corresponding number of similar behavior, whose cumulative percentage exceeds $\alpha$ for the first time, as the similar behavior threshold. Therefore, it can be seen from FIG. 8 that the cumulative percentage is 85.04% when the number of similar behavior is one. And the cumulative percentage, when the number of similar behavior is 2, is 96.38%. At this point, the threshold value set of 90% has been exceeded, so that 2 is the similar behavior threshold.

2.5 Identification of Friendship
We identify the pair of students whose number of similar behaviors with the other exceeds 2 as friends to each other, accordingly, for example, students with user id 9733 and 9773 are identified as friends.
3. Results and Discussion

This article uses the massive data of students’ campus cards to identify the friendships between college students. It is found that 45703 pairs of students in the college have been identified as the friends to each other through the campus card data in library. In addition, according to the data obtained from the research, it is easy to build a friend network of college students. And what is more crucial is that it's feasible to find some core nodes whose node degree is bigger compared to other nodes as the heart of his or her circle of friends from the friend network map. Owing to their great impact on other people around them, they can be regarded as key objects for some survey carried out by school and advertising served by some companies.

Compared to traditional research methods, this method based on campus card in real campus life has fewer false elements, stronger friendships, more pertinence and reliability. What’s more, because these data really depict the daily lives of college students, it is more practically significant for our research and formulation of some programs. However, it is undeniable that this data-based study has some contingencies. For example, in real life, two people who are strange to each other may often appear together because of their similar daily habits so that they may be mistakenly identified as a pair of friends. In order to reduce the chance of data, it is necessary to use a large number of different occasions’ data, and even the behavior data of these different scenes can be combined to identify the relationship between students.

College education is an important issue in the field of education and management. How to develop correct education policy to improve students' comprehensive abilities is a question that every college administrator concern about. The usual educational policies are universal. They can achieve some good results for some students, but cannot solve personal problems. Personal problems like poverty, psychological problems and emotional problems cannot be solved just by universal policies. At the same time, due to differences in region, family education and religious beliefs, different individuals also need specific and targeted solution even if they face the same problem. Owning to these massive data and some effective methods, each student's daily behavior could be vividly presented, based on which it is more accurate and effective to conduct a precise investigation and the counseling of study and daily life for some specific students in need. For example, it is highly instructive to find those isolated nodes from a friend network as an important target for psychological counseling, but also can study the relationship between college students eating breakfast and academic performance.
4. Conclusion

Aiming at the identification of friend relationships, a new method based on massive data has been proposed. This method mainly includes four steps, namely data preprocessing, generation of similar behavior matrix, determination of similar behavior threshold and identification of friend relationships. Besides, thanks to the massive data of campus card, the friend network diagram has been built successfully. And compared to the traditional questionnaires, this method based on the massive data is more accurate and instructive.

Acknowledgements

The authors would like to acknowledge Jimei Shen in Northeastern University for valuable discussions. This work is partially supported by National Undergraduate Training Program for Innovation and Entrepreneurship.

References


