Mining Frequent Itemsets of Novel Characters Based on Association Rules

Hui Cao1*, Jiao Li1

1Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu, China

*Corresponding author’s e-mail: caohui321@sina.com

Abstract. This article belongs to the field of Data Mining. It selected the novel "White Deer Plain" as the research object, using the association rules mining method to research the book characters and show the character of interpersonal relationship. In this paper, weka was used as an auxiliary tool, and Apriori algorithm was used in association rules. The whole work text was input into the database, and each chapter of the article was separated. Through co-occurrence analysis, the name nodes in the text were extracted as keywords, and weight was assigned. Keywords appearing in a chapter were treated as a record, and a list of keywords from all chapters was combined to form a data set. Multiple scans of the database are made, and frequent item sets were found from the constructed data set, and association rules between people were found. The method assists the reader to clarify and grasp the relationship between the characters before reading the full text. It greatly saves the time of reading.

1. Introduction

1.1 Overseas Research Status

Association rules include two basic algorithms: Apriori algorithm and FP-growth algorithm. In 1993, association rules were first proposed by R. Agrawal, and then Apriori algorithm. Apriori algorithm was first proposed to solve the shopping basket problem, also known as apriori algorithm. In 1994, they established the theory of Item Set Lattice Space and put forward the famous Apriori algorithm based on the above two theorems. Up to now, Apriori is still widely discussed as a classical algorithm for mining association rules. Later, many researchers have done a lot of research on the mining of association rules1. The main research directions of Apriori algorithm up to now include: multi-cycle mining algorithm (hierarchical mining algorithm), incremental updating algorithm, distribution, parallel mining algorithm, multi-level association rules mining algorithm, multi-value association rules mining algorithm, Association Rules Mining Algorithm Based on concept lattice, etc.

1.2 Domestic Research Status

The research of association rules in China is mainly concentrated on three aspects: algorithm, theory and practical application. There are fewer research areas and examples than abroad. The most classical Apriori algorithm has been improved by Wang Zhigang and other scholars, and its efficiency has been greatly improved. Combining the advantages of FP-tree2 and Apriori algorithm, the distributed parallel algorithm is studied.
2. Basic Principles of Association Rules

Association rules belong to one of the data mining methods. Give a set \( I = \{i_1, i_2, \ldots, i_m\} \), Transaction Data Set \( T = \{t_1, t_2, \ldots, t_n\} \). Each item \( t_n \) contains a subset of \( I \). In Association analysis, a set containing 0 or more items is called item set. If item set \( L \) contains \( k \) items, then \( L \) is called \( k \)-item set \(^3\). Association rules are determined by support and confidence. Suppose there is a database with \( D \), which includes the percentage of transactions \( A \) and \( B \) at the same time, that is, support \( (A \Rightarrow B) = P(A \_B) \); Confidence is the percentage of transaction \( D \) containing \( A \), the same conditional probability of transaction \( D \) containing \( B \). The threshold of minimum support and minimum confidence are set artificially. If the minimum support and minimum confidence are satisfied, the association rules are interesting. The process of mining association rules is divided into two steps: mining frequent itemsets in the first step and finding association rules in the second step. The process of mining association rules is shown in Figure 1 below.

3. Apriori algorithm

This paper uses Apriori algorithm in association rules to mine. The basic idea of the algorithm \(^4\) is to find all the frequency sets first, and the frequency of these itemsets is at least the same as the predefined minimum support. Then strong association rules are generated from frequency sets, which must satisfy minimum support and minimum confidence. Then use the frequency set found in step 1 to generate the desired rules, generating all the rules that contain only the items of the set, with only one item on the right of each rule. Once these rules are generated, only those rules that are greater than the minimum credibility given by the user are left behind. The Apriori algorithm first scans the data set to determine the support degree of each item, and then gets the set \( L_1 \) of all frequent itemsets. Then it generates candidate binomial sets by using the generated frequent itemsets, and scans the database to count the support degree of all candidate binomial sets, and gets the set \( L_2 \) of frequent binomial sets. Repeat the above procedure until no new frequent \( K \) itemsets are generated.

\begin{itemize}
  \item \( L_1 = \text{find\_frequent\_1\_itemsets}(D) \);
  \item for \( (k=2;L_{k-1} \neq \Phi; k++) \) {
    \item \( C_k = \text{apriori\_gen}(L_{k-1}, \text{min\_sup}) \);
    \item for each transaction \( t \in D \) //scan \( D \) for counts
    \item \( C_t = \text{subset}(C_k, t) \); //get the subsets of \( t \) that are candidates
    \item for each candidate \( c \in C_t \)
      \item \( c.\text{count}++ \);
  \item }
  \item \( L_k = \{c \in C_k | c.\text{count} \geq \text{min\_sup} \} \)
  \item }
  \item return \( L = \bigcup_k L_k \).
\end{itemize}

The flow chart of Apriori algorithm as shown in Figure 3-1 below.
4. Mining results
The research object of this paper is the novel White Deer Plain, which uses data mining tool Weka to mine association rules. This paper chooses the first 10 chapters of Bailuyuan as corpus research. The first 10 chapters are extracted from each chapter, each chapter is treated as a data item, and finally the data set is synthesized.

Figure keywords are shown in Table 1 below.
Write the script according to the. ARFF format. Mining frequent itemsets according to Apriori algorithm, the parameters set are shown in Figure 3 below.
Mining frequent itemsets includes 14 frequent itemsets, as shown in Table 2 below. There are 9 frequent binomial sets, as shown in Table 3 below. There are four frequent trinomial sets, as shown in Table 4 below. Frequent Quadrinomial sets, as shown in Table 5 below and five best Association rules, are shown in Table 6 below.
Table 1. List of keywords

| ID | List of keywords                  |
|----|-----------------------------------|
| 1  | Bai Jiaxuan, Lusan, Hushi        |
| 2  | Bai Jiaxuan, Mr. Zhu             |
| 3  | Bai Jiaxuan, Lao Xiuci, Lu Zilin, Wu Changgui |
| 4  | Bai Jiaxuan, Lu san, Mr. Zhu, Luzilin |
| 5  | Bai Jiaxuan, Lu Zilin, Heiwa, Mr. Xu, Lu Zhaopeng |
| 6  | Bai Jiaxuan, Lu san, Mr. Zhu, Luzilin, Governor Fang |
| 7  | Bai Jiaxuan, Lu san, Luzilin, Mr. Xu, Tian Fuxian |
| 8  | Bai Jiaxuan, Lu san, Luzilin, Heiwa, Mr. Xu, Bai Xinger, Bailing |
| 9  | Hei Wa, Li Xiang, Wang Xiang, Huang Laowu |
| 10 | Bai Jiaxuan, Lu san, Luzilin, Heiwa, Bai Zhao, Lutaiheng |

Table 2. Frequent item set

| ID | Frequent Item Set|
|----|-------------------|
| 1  | Bai Jiaxuan       |
| 2  | Hu shi            |
| 3  | Lao Xiucai        |
| 4  | Wu Changgui       |
| 5  | Lu Zhaopeng       |
| 6  | Governors Fang    |
| 7  | Tian Fuxian       |
| 8  | Bai Xinger        |
| 9  | Bai ling          |
| 10 | Li Xiang          |
| 11 | Wang Xiang        |
| 12 | Huang Laowu       |
| 13 | Bai Zhaooshi      |
| 14 | Lu Taiheng        |

Table 3. Frequent binomial sets

| ID | Frequent Binomial Set|
|----|-----------------------|
| 1  | {Bai Jiaxuan, Li Xiang} |
| 2  | {Bai Jiaxuan, Wang Xiang} |
| 3  | {Bai Jiaxuan, Huang Laowu} |

Table 4. Frequent trinomial sets

| ID | Frequent Trinomial Set|
|----|------------------------|
| 1  | {Bai Jiaxuan, Li Xiang, Wang Xiang} |
| 2  | {Bai Jiaxuan, Li Xiang, Huang Laowu} |
| 3  | {Bai Jiaxuan, Wang Xiang, Huang Laowu} |
| 4  | {Li Xiang, Wang Xiang, Huang Laowu} |

Table 5. Frequent Quadrinomial sets

| ID | Frequent Quadrinomial Set|
|----|---------------------------|
| 1  | {Bai Jiaxuan, Li Xiang, Wang Xiang, Huang Laowu} |

Figure 3. Parameter settings

```java
// Code snippet goes here...
```
Table 6. Best rules

| ID | Best rules               |
|----|-------------------------|
| 1. | {Li Xiang} ==> {Bai Jiaxuan} |
| 2. | {Bai Jiaxuan} ==> {Li Xiang} |
| 3. | {Wang Xiang} ===> {Bai Jiaxuan} |
| 4. | {Bai Jiaxuan} ===> {Wang Xiang} |
| 5. | {Huang Lao Wu} ===> {Bai Jiaxuan} |

5. Summary
Mining frequent itemsets is an important and popular research topic in association rules. This paper briefly introduces the principle of association rules and the process of mining frequent itemsets using Apriori algorithm, as well as the results of mining association rules using data mining tool Weka. In this paper, there are still many shortcomings in using keyword extraction method. We hope to make some improvements in frequent itemsets and association rules mining of keywords, change the parameters and compare them. It is hoped that another method of association rules, FP-growth algorithm, can be used to mine frequent itemsets next time. At the same time, we hope that association rules can be applied in more fields and even in life.

Acknowledgments
This paper was supported by the national science and technology support program "big data-driven machine translation between Chinese and English and Chinese minority languages "(project no. 2017YFB1002103).

Reference
[1] jiaojiao Chao. Overview of association rule mining. Baidu wenku.2017.
[2] Sun Jinxin. Research on Association Rules in Data Mining [J].208,6,8(3). Intelligent Computers and Applications
[3] Guangpu Li, Miaohua Huang. Research progress and mainstream methods of frequent itemset mining [J]. Computer Science. 2018, 11, 45 (11A).1-11.
[4] Lina Lu, Yaping Chen, Hengyi Wei. Research on the algorithm Apriori of mining association rules. Mini-Micro System,2002,21(9):940-943.
[5] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concept and Technology [D]. Machinery Press, 2015.