Research on Association of 95598 Customer Service Work Orders Based on Sequential Frequent Pattern Mining

Xu Xin\textsuperscript{1,2}, Fu Jun\textsuperscript{1,2}, Sun Zhijie\textsuperscript{1,2}, Wang Li\textsuperscript{1,2}, Liu Xiaowei\textsuperscript{1,2}, Li Zhiguo\textsuperscript{3}

\textsuperscript{1}Institute of Electric Power Science of State Grid Jibei Electric Power Co., Ltd., Beijing 100045;
\textsuperscript{2}Huadian Electric Power Research Institute Co., Ltd., Beijing 100045;
\textsuperscript{3}Qinhuangdao Power Supply Company of State Grid Jibei Electric Power Co., Ltd., Qinhuangdao, Hebei Province 066000

Abstract. With the continuous development of social economy and people's living standard, people's demand for electricity is also increasing day by day, as well as the demand for power supply capability of power supply enterprises and service level of staff. This has brought greater pressure and challenges to power supply enterprises, and the most prominent challenge is how to deal with customer complaints. In order to reduce customer complaints, it is necessary to predict whether customers will complain in the future according to the call track of customers who have already reflected their demands. Therefore, in complaint prediction, frequent pattern mining based on customer call track is particularly important.

1. Introduction
This article uses the algorithms of Apriori frequent pattern mining and sequence frequent pattern mining Prefixspan to generate features about sequence tracks. By comparing the feature results generated by the two algorithms, we can find out which frequent pattern mining algorithm is more conducive to pattern construction in complaint prediction.

2. Algorithm introduction

2.1 Three evaluation criteria for frequent item sets
Support is the proportion of the number of times of associated data appearing in the total data set.

\[
\text{Support}(x,y) = \frac{\text{NUMBER}(xy)}{\text{NUM}(All\ Samples)} \tag{1}
\]

Confidence reflects the probability that another data will appear after one data appears.

\[
\text{Confidence}(x|y) = \frac{P(xy)}{P(y)} \tag{2}
\]

Lift reflects the correlation between X and Y.

\[
L(x \leq y) = \frac{P(x|y)}{P(x)} \tag{3}
\]

2.2 Apriori frequent pattern mining
Apriori algorithm uses support as the criterion for judging frequent item sets. Its purpose is to find the largest K frequent item sets, find the frequent item sets that meet the support criterion, and find the largest number of frequent item sets.

Apriori algorithm flow is as follows:
Input: Data set D, Support degree α
Output: Maximum Frequent K – item set
(1) Scan the entire data set to obtain all the data that have appeared as candidate frequent item sets. K=1, frequent 0 item sets are empty sets.
(2) Mining frequent K-item set.
(3) Scan data to calculate the support of candidate frequent K-item set.
(4) Remove the data set with the support degree lower than the threshold in the candidate frequent K-item set to obtain the frequent K-item set; if the obtained frequent K-item set are empty, directly returning the set of frequent K-1 item set as the algorithm result, and ending the algorithm. If there is only one frequent K-item set, the set of frequent K-item set is directly returned as the result of the algorithm, and the algorithm ends.
(5) Generating candidate frequent K+1 item set based on frequent K item set.
(6) Make K = K+1 and proceed to Step 2.

2.3 prefixispan sequence pattern mining algorithm
PrefixSpan sequence pattern mining algorithm is a kind of association analysis, which can be used to mine other work order types associated with complaint strong correlation work order.
A sequence is a complete information flow. An item is a set of the smallest constituent units in the sequence, and the item is composed of events. The important step of the algorithm is sequence pattern discovery:
The algorithm flow is as follows:
Input: Sequence Data Set S and Support Threshold
Output: all frequent sequence sets meeting the support requirements
(1) Find out all prefixes with a length of 1 and corresponding projection databases .
(2) Count prefixes with length of 1, deleting items corresponding to prefixes with support lower than the threshold from data set S, and obtaining all frequent 1-item sequences at the same time, i=1.
(3) Recursive mining is carried out for each prefix with length i satisfying the support requirement.
(4) Sequence pattern discovery .
Given the data set D and the minimum support threshold minsup specified by the user, the task of sequence pattern discovery is to find all sequences whose support is greater than or equal to minsup. A common method to generate sequence patterns is to enumerate all possible sequences and count their respective support.
(5) Candidate generation
A pair of frequent (k-1)-sequences are merged to generate candidate k-sequences. In order to avoid repeated generation of candidates, the traditional Apriori algorithm merges a pair of frequent k-item set only when the current k-1 item are the same. Similar methods can also be used for sequences.
(6) Candidate pruning
If at least one candidate k-sequence is infrequent, it will be subtracted.
(7) Support count
During the support count, the algorithm enumerates all candidate k-sequence belonging to a specific data sequence, and the support of these candidates will increase. After counting, the algorithm identifies frequent k-sequence and discards candidates whose support is less than the threshold.

3. Feature construction

3.1 Data description
This project is based on the historical work order data of 95598 seats in electric power(95598 is the national grid company power service hotline). The data are used in 2017 and January to June 2018; Nine types of work orders are involved 'report','complaint', 'business consultation', 'service application', 'suggestion', 'opinion', 'customer reminder', 'praise' and 'business supervision'; The total amount of data is over 2.4 million. Fields involved: 98. There are 9 fields finally selected. The details are as follows:
Table 1. Feature fields

| Field name     | Field meaning                  |
|----------------|--------------------------------|
| ACCEPTCONTENT  | Accepted content               |
| HANDLESITUATION| Handling situation             |
| ORDERNO        | National Network Work Order No.|
| CALLNO         | Calling No.                    |
| TEL            | Contact number:                |
| TEL2           | Contact number 2               |
| REQSTARTDATE   | Request start time             |
| REQFINISHDATE  | Request end time               |
| TYPECODE       | Business type                  |

Apriori algorithm and Prefixspan algorithm are used to mine frequent patterns.

3.2 Apriori Frequent pattern mining research

![Flow chart of correlation analysis based on ACCEPTCONTENT](image)

(1) All incoming call information of the user is found according to the calling number of the work order, the time information extracted according to the national network work order number is limited to 5 days, the associated work order of the user is searched forward (if a complaint work order is encountered within 5 days, stop), and sequence length characteristics are constructed.

(2) Based on the generated sequence length feature, the problem type is extracted by using the acceptance content in the original field, and then it is digitized, and the Apriori algorithm is used for frequent pattern mining. Only two item sets are considered here.

Result analysis:

(1) It can be seen that many types have high confidence in "frequent power outages", but since the support for "frequent power outages" is very high, reaching more than 96%, the value of this information is limited. The support of "rush repair service" is also very high, accounting for more than 82%, and similar to "frequent power outages". However, it can still be seen that "business expansion of reporting equipment" and "power outage due to overdue bills" have obvious positive correlation with it.

(2) The confidence of "construction personnel service standard", "business expansion report", "rush repair time limit", "payment reminder", "personnel attitude" and "personnel violation" in "overdue payment and power resumption" is higher than 80%.

(3) The confidence level of "business expansion reporting", "overdue payment stopping and resuming electricity" and "rush repair time limit" to "rush payment" is higher than 80%.

(4) The confidence level of "construction personnel service standard", "industry expansion reporting", "rush repair time limit", "rush payment", "personnel attitude" and "personnel violation" to ("overdue power outage", "frequent power outage") is higher than 80%.
To sum up, "frequent power outages" are the complaints with the highest frequency, appearing almost every day.

![Diagram](image)

**Figure 2. Apriori frequent pattern mining flow chart based on type code**

1. All incoming call information of the user is found according to the calling number of the work order, and the time information extracted according to the national network work order number is limited to 15 days. The associated work order of the user is searched forward (if a complaint work order is encountered within 15 days, it will be stopped) to construct sequence length characteristics.

2. Based on the generated sequence length feature, frequent pattern mining is carried out using Apriori algorithm by using the type code in the original field with the minimum support set to 0.01.

3. According to the comparison between the sequence formed by the current work order type to be predicted and the reserved sequence, a binary label is constructed.

| Apriori promotion sequence | Support | Lift |
|----------------------------|---------|------|
| [1, 5]                     | 0.029   | 1.37 |
| [3, 5]                     | 0.021   | 0.99 |

(1 for fault repair, 3 for consulting work orders, and 5 for complaints)

1. Generally speaking, there are differences between results and assumptions: the work order closely related to the complaint work order is not a reminder work order, but a failure repair work order and a consultation work order. This shows that the probability of complaint work orders generated by these two types of work orders is not very high during the process of reminders, or if the problem is solved after reminders, the probability of complaints is also very low.

2. At the same time, according to statistics, 72.7% of the work orders that generated complaints contained repair work orders \( P(\text{failure repair} | \text{complaint}) = 72.7\% \); 68.7% contained consulting work orders \( P(\text{consulting} | \text{complaint}) = 68.7\% \).

3. Calculate the imbalance ratio \( IR(\text{failure repair, complaint}) = 1.37 \) for failure repair work orders and complaint work orders, and the imbalance ratio \( IR(\text{failure repair, complaint}) = 1.00 \) for consultation work orders and complaint work orders, both values are very close to 1, and the correlation between failure repair work orders and consultation work orders and complaint work orders is not very strong.

### 3.3 Feature construction of sequential frequent pattern mining
(1) All incoming call information of the user is found according to the calling number of the work order, the time information extracted according to the national network work order number is limited to 30 days, the associated work order of the user is searched forward (if a complaint work order is encountered within 30 days, stop), and sequence length characteristics are constructed.

(2) Based on the generated sequence length feature, the TYPECODE in the original field is used for sequence frequent pattern mining.

(3) According to the comparison between the sequence formed by the current work order type to be predicted and the reserved sequence, a binary label is constructed.

Result analysis:

(1) Generally speaking, the number of users who have complained accounts for 1.31% of all work order users; The sequence containing complaints accounted for 0.93% of all work order sequences; Direct complaint sequence after incoming calls accounted for 45.38% of all complaint sequences; Among the users who complained, 47.45% complained directly; Users with direct complaint experience accounted for 48.87% of the total complaints (indicating that there is no obvious trend of multiple complaints for such users);

(2) Judging from the work order type of "failure to report for repair", the proportion of complaints (complaints caused by "failure to report for repair") after one feedback or more is 80.51%; After no less than two feedbacks, 41.49% complained. After no less than three times of feedback, the proportion of complaints was 20.67%.

(3) From the work order type of "business consultation", the proportion of complaints (complaints caused by "business consultation") after >=1 feedback is 47.48%; After no less than two feedbacks, the proportion of complaints was 20.42%. After no less than three times of feedback, the proportion of complaints was 9.66%.

(4) Based on the analysis of the two types of work orders, namely "failure to report for repair" and "business consultation", the proportion of complaints after "failure to report for repair" and "business consultation" is 23.68% (including "business consultation", "failure to report for repair" and "business consultation": 6.12% ); The proportion of complaints after "business consultation" and "fault repair report" is 18.34% (including "fault repair report" - "business consultation" - "fault repair report": 9.64%);

(5) Judging from the work order type of "user reminders", the probability of complaints after reminders is 6.70%; The probability of complaints after repair and consultation, and then reminders are 4.13% and 3.05 % respectively. The probability of submitting a complaint again after the reminder or consultation is 2.63% and 2.56% respectively.

3.4 Prefix span sequence frequent pattern mining
In the sequence pattern mining of 3.3, only the support is considered, and the lift is not considered, so there is a Prefixspan algorithm about the lift.
Figure 4. Prefix Span Sequence Frequent Pattern Mining Flowchart

Analysis of specific process

(1) All incoming call information of the user is found according to the calling number of the work order, the time information extracted according to the national network work order number is limited to 30 days, the associated work order of the user is searched forward (if a complaint work order is encountered within 30 days, stop), and sequence length characteristics are constructed.

(2) Based on the generated sequence length feature, the TYPECODE in the original field and Prefixspan algorithm are used to mine the sequence length that is equal to the lifting sequence length.

(3) According to the comparison between the sequence formed by the current work order type to be predicted and the reserved sequence, a binary label is constructed.

Table 3. Prefix Span Sequence Frequent Pattern Mining Results

| Lifting sequence | Lifting |
|------------------|---------|
| [1, 3, 3, 15,5]  | 14.13   |
| [1, 1, 1, 15, 1,5] | 13.99   |
| [15, 1, 1, 1,5]  | 13.22   |
| [1, 15, 1, 1,5]  | 13.21   |
| [1, 1, 1, 1, 1, 1, 1,1,5] | 12.67   |
| [1, 1, 1, 1, 15,5] | 12.55   |
| [1, 1, 3, 15,5]  | 12.45   |
| [1, 1, 15, 1,5]  | 12.16   |
| [15, 15, 1,5]    | 11.79   |
| [1, 1, 1, 1, 1, 1, 1, 1,1,5] | 11.52   |

(1) 3.4 Prefixspan sequence frequent pattern mining can be seen as an improvement of 3.3 sequence pattern mining. Prefixspan takes into account the influence of lifting degree on the final result. From the result, the support degree of sequence [1,3,3,15] is 0.00039, which is not high in the sequence, but its importance will be greatly improved considering the lifting degree of 14.13.

(2) In sequential frequent pattern mining and Apriori frequent pattern mining, [3,5] and [1,5] are all very important association item sets. In sequential frequent pattern mining, the support is 0.0032 and 0.0019 respectively. However, the importance of these two item sets is greatly reduced by using Prefixspan algorithm for sequential frequent pattern mining. From the business logic analysis, although fault repair reporting and business consultation will affect whether users complain or not, because there are many complaint sequences in these two types of work orders, accounting for 80.51% and 47.48% respectively, these two types of work orders are regarded as the pre-work order type of complaint sequence, which is of little significance to the pattern. The conclusion drawn by Prefixspan algorithm conforms to the business logic.
3.5 Comparison between frequent pattern mining and sequential frequent pattern mining: Although Apriori and Prefixspan algorithms are both mining algorithms for frequent patterns, Prefixspan algorithm will take into account the timing problem and will consider that a work order appears many times in a sequence, so it will retain more.

From the analysis of display results, Apriori frequent pattern mining does not consider the timing problem and the same TYPECODE is only considered once, so the results are not good, and the correlation between reminders and complaints is not mined.

In the Prefixspan algorithm, the time sequence problem is considered, and the multiple occurrences of a certain work order type are considered. Therefore, Prefixspan not only excavates the influence of work orders on complaints in reminders 3, but also excavates the different influence degrees on complaints under different sequences.

4. Example verification
Since the Prefixspan algorithm is an improvement of the previous sequential frequent pattern mining, Apriori and Prefixspan algorithms are respectively used to build features based on TYPECODE and random forest pattern for prediction under the condition of ensuring the consistent selection of other features. The following is the prediction result.

Table 4. Forecast Results of Random Forest Pattern

| Algorithm  | Positive sample accuracy | Positive sample recall rate |
|------------|--------------------------|----------------------------|
| Apriori    | 0.53                     | 0.63                       |
| Prefixspan | 0.81                     | 0.55                       |

5. Conclusion
From the analysis of display results, Apriori frequent pattern mining does not consider the timing problem, and the same TYPECODE is only considered once, so the results are not good, and the correlation between reminders and complaints is not mined.

In the Prefixspan algorithm, the time sequence problem is considered, and the multiple occurrences of a certain work order type are considered. Therefore, Prefixspan not only excavates the influence of work orders on complaints in reminders 3, but also excavates the different influence degrees on complaints under different sequences.

From the example verification, it can be seen that when the sequence problem is considered in the job type frequent pattern mining, i.e. the sequence frequent pattern mining is carried out by using the Prefixspan algorithm, the constructed feature is much higher in accuracy than the frequent pattern mining feature based on the Apriori algorithm. This indicates that the customer complaint is not only related to the type of the incoming electrical sheet, but also has a great correlation with the order of the incoming electrical sheet types.

In the construction of the complaint prediction pattern, the accuracy of the pattern can be greatly improved by using the features of frequent pattern mining of Prefixspan sequence.

References
[1] Lin Xiqiao, Yan Xu, Huang Wei. Classification and Optimization of 95598 Customer Service Work Orders Based on Principal Component Analysis [J]. Guangxi Electric Power, 2017 (04): 14.
[2] Hu Jiming. Research and Improvement of Apriori Algorithm in Mining Association Rules [J]. Computer Technology and Development. 2016.16 (4); 99-101.
[3] Guo Jianbing, Song Shunlin. An improved algorithm based on Apriori algorithm [J]. Computer Engineering and Design. 2008,29 (11); 2814-2815.
[4] Wu Qi, Apriori Mining Algorithm Based on Cloud Computing [J]. Computer Measurement and Control. 2012.20 (6): 1653 - 1655.
[5] Chen Shu, et al. Research on Rules Obtained by Pattern Method [J], Computer Science, 2004, 31 (5); 138-141.

[6] Magnus Lie Hetland, Pål Sætrom. Evolutionary Rule Mining in Time Series Databases [J] 2005, Machine Learning(2-3):107~125.