Indicators Recommendation Algorithm based on Information Entropy and FP Growth for Power Report System

Yang Yifan¹ *, He Jin², Zhang Yumeng³, Zheng Jian⁴ and Liu Ning⁵

1,⁎Assistant Engineer, State Grid Tianjin Information & Telecommunication Company, Tianjin, China, E-mail: lg08133274@163.com.
2 Engineer, State Grid Tianjin Information & Telecommunication Company, Tianjin, China, E-mail: jin.he@tj.sgcc.com
3 Engineer, State Grid Tianjin Information & Telecommunication Company, Tianjin, China, E-mail: yumeng.zhang@tj.sgcc.com.cn
4 Senior Engineer, State Grid Tianjin Electric Power Company, Tianjin, China, E-mail: alex_lopez_cn@aliyun.com
5 Senior Engineer, State Grid Tianjin Electric Power Company, Henan, China, E-mail: hdln512@126.com

Abstract. For the Indicators recommendation problem of a power report system without user grade but a large amount of browsing records data, a novel algorithm of mining association rules based on information entropy and Fp-growth is proposed. Firstly, though normalizing the browsing frequency, the information entropy is used to determine the weights of items so as to calculate the interesting measurement value for users. Besides, the forgetting function is introduced to reflect the interest transfer. Then, the FP-growth algorithm is improved by dynamically adjusting support and activity. Finally, the browsing records data of the power report system were used as the test dataset for experiments. The results showed that the algorithm can effectively recommend interest items for users, and the recommendation performance is significantly improved compared with other algorithms.

1. Introduction
With China’s power construction stepping to a new stage, the demand for information extraction is constantly increasing due to the huge amount of power data documents [1]. As a common way of data information exchanging between different management departments, reports model is faced with how to deal with more complex and massive data, so as to provide diversified information in form and content according to the interest of user. The traditional report model can't thoroughly and automatically get the association characteristics of overall power business, and the ability of analysing flexibly and intelligently is also very limited. Therefore, it is necessary to combine the big data technology and artificial intelligence to carry out the research on the intelligent report model [2].

In the research of intelligent report model, it is one of the key issues to select power data analysis indicators according to user’s interests. For the indicators recommendation problem of power report system, this paper focuses on the mining algorithm of user interests for system without user grade but a large amount of browsing records data. Common mining algorithms include Apriori [3], FP growth [4] and matrix based collaborative filtering (CF) [5]. The Apriori algorithm scan the whole data set in each iteration, which result in problems of high I / O ratio and more memory [6]. The CF algorithm has the problem of matrix sparsity. The improved matrix factorization CF algorithm [7] [8] alleviated
the computational cost caused by matrix sparsity, and the processing ability of non interest items is not ideal.

Tradition FP growth algorithm plays an important role in the mining of association rules because of its simplicity, low memory requirements and high efficiency. Considering the advantages of FP growth algorithm in mining efficiency and insensitivity to matrix sparsity, in this paper, the FP-growth algorithm is improved by using information entropy to measure the user's interest, and introducing the forgetting function to track the user interest, so that it can reflect the transfer of user interest. Then, dynamic support and activity are introduced into the FP-tree to achieve accurate recommendation. Through the proposed algorithm, it can automatically and intelligently meet the demand of report content diversity for power system, effectively reduce the workload of report analysis, and provide the basis for the construction and application of intelligent power data report model.

2. Preliminary

2.1. Information entropy

The information entropy [9] derived from the thermodynamics, which is proposed by C.E. Shannon, the father of information theory. In thermodynamics, thermal entropy is a physical quantity that represents the confusion degree of molecular state, while Shannon uses the concept of information entropy to describe the uncertainty of information sources. Suppose that there are $n$ states in a system $X$, denoted as $X\{x_1, x_2, \cdots, x_n\}$. $p(x_i)$ $(i = 1, 2, \cdots, n)$ denotes the appearing probability of $x_i$ in system $X$. The Shannon entropy $H(x)$ of system $X$ is defined as:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log (p(x_i))$$

where $0 \leq p(x_i) \leq 1$ and $\sum_{i=1}^{n} p(x_i) = 1$. When $p(x_i) = 0$, $0 \log(0) = 0$.

According to Shannon’s theory of information entropy, the greater the information entropy is, the higher the disorder degree of information is, and the less information it contains; the smaller the entropy is, the lower the disorder degree of information is, and the more information it contains.

2.2. Forgetting function

The interests and hobbies of people usually change greatly over time. H. Ebbinghaus, a German psychologist, gives a forgetting curve to describe the law of forgetting affairs for human through a systematic study of forgetting phenomenon [10].

![Ebbinghaus curve](image)

The curve shows that the process of forgetting is non-linear. At beginning, people forget fast, then gradually slow. Then people almost no longer forgetting after a certain period of time. According to the characteristics of Ebbinghaus forgetting description, an exponential decay function is usually uses as forgetting function.

$$f(t) = ae^{-bt} + c$$

Where, $a = 0.6487$, $b = 0.7962$, $c = 0.3513$. $f(t)$ is the proportion of remaining memory. The larger the value is, the user remember the specified item more deeply. $t$ is the time interval from the beginning, with unit of seconds.
Considering that the accessing records of users at different moment have different impact on the description of the user's interest at a certain moment. Because the latest browsing records can better reflect the current interest of user. The forgetting function [11] used is as follow.

\[ f(t) = ae^{-b(t_{ui} - t_{0})^2} + c \]  

where \( t_{ui} \) is the time when user \( u \) browse the item \( i \), \( t_0 \) is the earliest time when user browse the item, \( t_{max} \) is the latest time when user browse the item.

3. Proposed approach

Traditional FP-growth algorithm mine frequent itemset from the user's data, and then establish the relationship between users and items through the frequent itemset, so as to achieve association rules mining, and applied to the recommendation system. In this section, the traditional FP growth algorithm is improved to achieve accurate interest recommendation of user. The user data studied in this paper an extremely sparse, with the statistical sparsity of 96%. \( I = \{i_1, i_2, \cdots , i_m\} \) is the set of items, \( \{u_1, u_2, \cdots , u_n\} \) is the set of users, where \( n \) is the total number of items and \( m \) is the number of user. Table 1 lists the browsing records of a user from 2020.01 to 2020.05. From table 1, we can see the characteristics of user browsing records and data sparsity.

|       | a | b | c | d | e | f | g | h |
|-------|---|---|---|---|---|---|---|---|
| January | 924 | 35 | 9 |
| February | 215 | 44 | 15 | 11 | 2 |
| March | 36 | 378 | 3 | 5 |
| April | 47 | 73 | 1089 |
| May | 11 | 7219 |

3.1. Measurement of interest

For the recommendation system without rating, the data is the frequency record for user of browsing different items. Traditional FP-growth algorithm can only find out whether the item is browsed by users, but cannot record the number of times the item is browsed by users. Therefore, the information entropy matrix is used to measure the user's interest in the item, and the forgetting function is introduced to track the user's interest transfer, so as to reflect the change of user's interest with time. The steps are as follows:

**Step1:** An interest measurement matrix \( r(n \times m) \) is build for users, where \( n \) is the number of users, \( m \) is the number of items.

\[
    r = \begin{bmatrix}
        r_{11} & r_{12} & \cdots & r_{1m} \\
        \vdots & \vdots & \ddots & \vdots \\
        r_{n1} & r_{n2} & \cdots & r_{nm}
    \end{bmatrix}
\]  

(4)

where \( r_{ui} \) in the matrix is the number of times for user \( u \) accesses the ith item.

**Step2:** The elements in matrix \( r \) are transformed, as shown in the formula (5).

\[
    r_{ui}^* = \frac{r_{ui}}{\sum_{u=1}^{n} r_{ui}}
\]  

(5)

Then

\[
    r^* = \begin{bmatrix}
        r_{11}^* & r_{12}^* & \cdots & r_{1m}^* \\
        \vdots & \vdots & \ddots & \vdots \\
        r_{n1}^* & r_{n2}^* & \cdots & r_{nm}^*
    \end{bmatrix}
\]  

(6)

**Step3:** According to the information entropy formula (1), the information entropy of each item is calculated

\[
    E_i = -k \sum_{u=1}^{n} r_{ui}^* \ln(r_{ui}^*)
\]  

(7)
where \( k = \frac{1}{\ln(n)} \), \( j \) is the number of attributes, \( j = 1,2,\cdots,b \). When \( r_{ui} = 0 \), \( E_i = 0 \).

**Step4:** Calculate of the item weights

\[
w_i = \frac{1-E_i}{m-\sum_{j=1}^{m}E_j} \tag{8}
\]

**Step5:** Obtain the interesting measurement value of users

\[
R = \begin{bmatrix}
w_{11}r_{11}^* & w_{21}r_{12}^* & \cdots & w_{m1}r_{1m}^* \\
\vdots & \vdots & \ddots & \vdots \\
w_{1n}r_{n1}^* & w_{2n}r_{n2}^* & \cdots & w_{mn}r_{nm}^*
\end{bmatrix} \tag{9}
\]

where the higher the value of \( R_{ui} \), the more interesting in the item \( i \) for user \( u \).

**Step6:** Calculate the time weight matrix \( F \) according to the forgetting function (3).

\[
F = \begin{bmatrix}
w_{11}^f & w_{12}^f & \cdots & w_{1m}^f \\
\vdots & \vdots & \ddots & \vdots \\
w_{n1}^f & w_{n2}^f & \cdots & w_{nm}^f
\end{bmatrix} \tag{10}
\]

where \( w_{ui}^f \) is the time weight for user \( u \) of browsing the item \( i \).

**Step7:** The time weight matrix \( F \) is used to attenuate the user’s interest properly, then obtain the user’s real-time interest matrix \( R^* \)

\[
R^* = R \ast F \tag{11}
\]

After attenuation, the user’s interest value will change with time, which conform to the law of forgetting, and finally can reflect the user’s real-time interest.

### 3.2. Dynamic support

When the traditional FP-growth algorithm establishes the item header table, it usually sets the support according to the total number of users. The support is calculated with the proportion of 2\% [12].

\[
E_i = \text{Size}(U) \cdot 2\% \tag{12}
\]

According to the above formula, the items are selected with the proportion of the total number of users, which ignore the particularity of the project itself. For example, when an advertisement is pushed to a large number of users, the project will be considered to be the most interested project. It is obviously wrong. Therefore, we hope to retain interesting projects and filter out non interesting projects through dynamic support. Considering the browsing frequency proportion of each item among users is calculated by formula (11), so it can be taken as the user interest factor, the support degree of interest items will be significantly higher than that of non interest items. Therefore, the items are selected with the proportion of the user interest factor in this paper, so as to achieve the purpose of retaining interested projects and filtering out non interested projects.

\[
E_i' = \text{Size}(U) \cdot \sum_{u=1}^{U_i} R_{ui}^* \tag{13}
\]

where \( U_i \) is the user who has browsed the item \( i \). Then the item header table is created. And compare the size of equation (13) and equation (12). When \( E_i' \geq E_i \), it shows that the item meets the threshold requirements, when \( E_i' < E_i \), it show the item is filtered out. In this way, the items interested by most people are more likely to be retained in the header table, while the non interest items are more likely to be filtered out. Moreover, combined with formula (3), the effect of forgetting factor can also be reflected in the item header table. If the items that were liked by most people in the past are rarely used now, it is difficult to appear in the recommendation candidate set. Finally, the accurate recommendation for user interest can be realized.

### 3.3. Discovery of frequent patterns from IFP-tree

For user recommendation system, how to recommend preferred items for new users without history browsing record is a classic problem. In this paper, the item activity is introduced to express the popularity of the item, and the activity is added to the node of FP Tree to realize the user interest item mining. The item activity is defined as follows.

**Definition 1:** the product of the number of users browsed and the interest factor is called item activity, which is calculated as follows.

\[
A_i = \text{Size}(U_i) \cdot \sum_{u=1}^{U_i} R_{ui}^* \tag{14}
\]
The summation factor of the above formula represents a total value of interest of item $i$. The larger the value of $A_i$, the easier the item $i$ is accepted by the public.

### Table 2. Transactional database

| TID | Browsing Items | Frequent Items |
|-----|----------------|----------------|
| 10  | f,a,c,d,g,I,m,p | f,c,a,m,p       |
| 20  | a,b,c,f,I,m,o   | f,c,a,b,m       |
| 30  | b,f,h,j,o       | f,b             |
| 40  | b,c,k,s,p       | c,b,p           |
| 50  | a,f,c,e,I,p,m,n | f,c,a,m,p       |

### Figure 2. Improved Fp-tree.

Table 2 is an example of transactional database, its improved Fp-tree is shown as figure 2. As can be seen from Figure 2, the item with higher active is closer to the root, the total value of interest is higher and at the top of ranking results.

### 4. Experiment and analysis

In order to verify the effectiveness of the algorithm, the browsing records of the power monitoring system were used as the test data set. The data set collected a total of 3470 browsing records from 54 users, including 79 items. These data were stored in the database, and the effectiveness verification and performance tests were carried out respectively.

#### 4.1. User interest computing

This section selects the data records of five users from the dataset to verify the effectiveness of the algorithm. $r$ is the original access frequency matrix of the original user, $R$ is the interest measurement matrix of the original location frequency matrix calculated by formula (9), $F$ is the time weight matrix calculated by formula (10), and $R'$ is the user real-time interest matrix calculated by formula (11). According to the formula (12), the project activity $A = (0.1988, 0.0709, 0.2851, 0.1933, 0.0910, 0.1562, 0.1205)$, of which the activity of project $I_2$ is the largest, $A(3) = 0.2851$. It can be seen that the original frequency of some projects is similar, and the difference of interest is relatively large after the attenuation of forgetting function. Although the total interest value of project $I_5$ is larger, which is affected by forgetting function, the activity is only 0.0910. According to the activity, the public is more interested in $I_2$, $I_1$, $I_4$. 

---

The full text is not fully visible in the image, but the content that is visible appears to be discussing the summation factor in a formula, the importance of interest values in making items accepted by the public, and a table of transactional database entries with corresponding values. There is also a mention of an improved Fp-tree and an experiment and analysis section discussing user interest computing with specific formulas and values.
4.2. Performance testing

In this section, the proposed algorithm is compared with traditional FP-growth algorithm and matrix factorization algorithm. In the recommendation system, three indicators, recall rate, precision rate and $F_1$-measure are commonly used for evaluation. The larger the value of the three indicators, the better the algorithm is. The recall rate indicates the number of behaviors that the user has actually had in the recommended item, which accounts for the proportion of the user’s actual behaviors, and the accuracy rate indicates the proportion of the recommended item that the user has actually had behaviors in the total recommended items, and is calculated as follows:

$$\text{recall} = \frac{\sum_{u \in \mathcal{U}} |R(u) \cap T(u)|}{|T(u)|}$$

$$\text{precision} = \frac{\sum_{u \in \mathcal{U}} |R(u) \cap T(u)|}{|R(u)|}$$

In the formula, $T(u)$ is the recommended item set of user $u$, and $R(u)$ is the set of items actually liked by user $u$ in the test set. The $F_1$-measure considers both the precision rate and the recall rate, which can better reflect the effectiveness of the recommendation algorithm. The calculation is as follows:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

The precision rate, recall rate and $F_1$-measure rate of the proposed algorithm, traditional FP-growth algorithm and matrix factorization algorithm are calculated by test. The results are shown in Figures 3 to 5.
It can be seen from Fig. 3 to Fig. 5 that the performance of the improved FP-growth algorithm is significantly improved compared with other two algorithms. When the number of recommendations is 9, the three indicators are the largest and the recommendation effect is the best. This is mainly due to the improvement of this article not only considers the items that belong to the user's interest characteristics, but also adds a forgetting function to modify the user's interest in time, so that the algorithm can adjust the weight of these items in the Fp-tree.
5. Conclusion

With the continuous improvement of equipment information technology, big data technology and artificial intelligence technology, it is an urgent need to use the achievements of technological development to solve the problems existing in the power report system. In this paper a mining algorithm based on information entropy and FP is proposed to realize user interest mining based on user browsing frequency data. Through the mining algorithm, it can meet the needs of report indicators selection according user interest, and result in content diversity for power system report, so as to solve the problem that data report analysis needs to spend a lot of manpower and low efficiency, which is of great significance to comprehensively improve the management efficiency of power enterprises.

References

[1] Yang Zhi, Liu Zhi, Zhang Fan. Research and application of intelligent power data report model[J]. Electrical Measurement & Instrumentation, 2020, 57(15):95-99.

[2] Jiang Han. Report index model design and flexible customization development [J]. Science & Technology Vision, 2016(26): 165-165.

[3] Dandu S, Deekshatulu B L, Chandra P. Improved algorithm for frequent itemsets mining based on Apriori and FP-tree[J]. Global Journal of Computer Science & Technology, 2013.13(2): 132-137.

[4] Karthikeyan T, Vembandasamy K, Raghavan B. An intelligent type-II diabetes mellitus diagnosis approach using improved FP-Growth with hybrid classifier-based arm[J]. Research Journal of Applied Sciences Engineering & Technology, 2015.11(5): 549-558.

[5] Zhang C, Yang M, Lv J, et al. An improved hybrid collaborative filtering algorithm based on tags and time factor [J]. Big Data Mining and Analytics, 2018, 1(02): 48-56.

[6] Sakhapara M A. Comparative study of Apriori algorithms for parallel mining of frequent itemsets [J]. International Journal of Computer Applications, 2014.90(8): 21-24.

[7] Han Y, Xia K. Data Pre-processing Method Based on User Characteristic of Interests for Web Log Mining[C]// Fourth International Conference on Instrumentation & Measurement. IEEE, 2014: 872-876.

[8] Ye H, Zhang Q, Bai X. A new collaborative filtering algorithm based on modified matrix factorization[C]// 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017: 147-151

[9] SHANNON C E. A mathematical theory of communication [J]. Bell System Technical Journal, 1948.27(4): 379-423.

[10] Xiting Huang. Introduction to psychology [M]. People's education press, 2007: 365-367.

[11] Yanqing Lu. Matrix Decomposition Recommendation Algorithm Research Based on Forgetting Function and Multi-attribute Feature Extraction [D].

[12] Han J, Pei J. Mining frequent patterns without candidate generation[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA. ACM, 2000: 16-18.

Acknowledgements

This work was supported by State Grid Tianjin Information & Telecommunication Company grant KJ20-1-51.