Remote Sensing User Demand Mining Method Based on Directed Item Graph-Based Double-Layer FP-Tree

Zhonggang Zheng[1,2,3,4,5], Jinghui Zhang[5]*, Haibei Yao[5]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;

[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

[3]University of Chinese Academy of Sciences, Beijing 100049, China

[4]Key Laboratory of Network Information System Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100094, China

[5]Beijing Institute of Remote Sensing Information, Beijing 100089, China

**Corresponding Author.*

Abstract

With the development of remote sensing technology and the commercialization of aerospace industry, the requirements of remote sensing observation will grow rapidly in the future, and the shortage of remote sensing resources and the conflict between observation tasks will become increasingly prominent. For the lack of mature data mining means for remote sensing user demand analysis and demand pre-processing, this paper proposes a novel association analysis method based on directed term graph-based double-layer FP-tree to extract users' remote sensing demands and observation target characteristics. Firstly, by analyzing the dependencies between the demand items, the directed term graph about user, target, payload and observation date is designed. Secondly, a double-layer FP-tree based on the directed term graph is presented. Such a tree can effectively express the logical relationship between demand data and realize the data compression of common input parts, which features small data amount and fast operation. And then, an association rule analysis method based on FP-growth algorithm for remote sensing user demand mining is proposed. In the process of association analysis, the double-layer FP-tree is decomposed into four sub-FP-trees, and for each sub-FP-tree, FP-growth algorithm is adopted to generate frequent itemset. This frequent itemset generation process has some superior properties, i.e., the data set is scanned only twice, the search path is short, search range is small, and there is no need to generate candidate itemset, and no repeated frequent itemset is generated. Finally, a transaction set of remote sensing demands is analyzed to verify the effectiveness of the proposed method. Some conclusions are drawn in the end of this paper.

Keywords: *Remote sensing demand mining, user, association rule analysis, directed item graph, double-layer FP-tree, FP-growth algorithm*

I. Introduction

In recent years, with the development of space launch technology, satellite application technology, payload technology and other space support technologies, as well as the improvement of supporting software and hardware, aerospace science and technology has played an increasingly prominent role in national economy and national defense construction of all countries in the world. Due to the vigorous promotion of relevant policies of various countries and emerging space enterprises, the aerospace industry has entered the commercialization process in many fields, with the types and number of satellites in orbit increasing day by day. Remote sensing satellite, as an extremely important strategic resource, has achieved a dominant role in many fields such as the land and resources survey, the natural disaster emergency rescue, the urban construction and management, the earth's environment monitoring and so on [1,2]. And a long-term and stable operation of earth observation satellite system will be established, then there

will be a large number of remote sensing satellite providing data services for users [3-6]. It can be imagined that the commercialization of remote sensing satellite application will cause an explosive growth trend of demand for remote sensing data from users such as government management departments, scientific research institutes and local enterprises, with various forms, contents and requirements [7-11].

However, resource shortage and demand conflict are urgent practical problems faced by remote sensing satellite management, and especially in the future, these problems will become more prominent. Therefore, mining the hidden information of users, observation targets, observation time intervals, sensor types and other demands from a large number of historical remote sensing demands is an important means to realize demand arrangement in advance, reasonable allocation of resources and guarantee the implementation of users' remote sensing demands. In fact, some simple, intuitive rules can be artificially judged. For example, China's aid agencies to Sudan usually submit remote sensing demands for damaged roads in rainy season in August every year. According to this rule, observation mission can be generated and submitted to remote sensing satellite data exchange center as soon as possible in order to obtain a better resource arrangement for observation.

The difficulty of remote sensing user demand analysis is that the relations and rules among the parameters representing demands are implicit rather than explicit, which requires complex data analysis or operation to find out, and especially in the condition of a large amount of data, such a problem is particularly intractable [12,13]. In order to mine meaningful and potential demand rules from a large amount of remote sensing data, it is necessary to adopt some special data analysis methods [14-16]. Methods based on statistics or rule summaries can partly reflect the regularity of demands, but it is difficult to explore the relationship between different data and its impact of regularity [17]. Data mining is a new discipline that can find potential and valuable rules and knowledge from large quantities of data [18,19]. And it has been an effective tool to analyze huge amounts of data and may become an important technical means to solve the mining and analysis of remote sensing user demand. At present, data mining has been applied in many fields such as bioinformatics, medical diagnosis, and scientific data analysis. However, using data mining to analyze remote sensing user demand is rarely studied.

In data mining, association rule analysis is a relatively independent algorithm field. It has a sound theoretical foundation, can solve a specific type of problems independently [20,21], and can be used to find meaningful connections hidden in large data sets [22,23]. Association rule analysis, also known as Market Basket Analysis, was originally proposed to discover interesting relationships in huge amounts of customer purchase data [24]. If retailers want to know the customer's purchase habits and what other items the customer will purchase during a shopping trip, a basket analysis may be required. Association analysis is to analyze customers' purchase habits by finding the relationship between the items they put into the "market basket". Some relationship between items is called association. The finding of such associations could help retailers know which items are frequently purchased at the same time, thus helping them improve marketing strategies.

Compared with the conventional Market Basket Analysis, remote sensing user demand analysis has many new features. First of all, remote sensing user demand mining is mainly user-oriented, focusing on the analysis of user preferences, to develop evidence-based, feasible and satisfactory observation plans for users. That is to say, the process of remote sensing user demand mining has a certain tendency, and the importance of each data is not equal. Secondly, there exist some certain logical relationships between different items. For example, to observe which target is determined by the user, and which payload is used is determined by both the user and the target; conversely, these logical relationships are not valid. Moreover, the value range of some items is limited. For example, the payload can be usually divided into optical payload, SAR payload and electronic reconnaissance payload (referred to as Elec. payload in the following paper), and the observation time interval usually can be several days, i.e., a week or a month.

Apriori algorithm is the first association rule analysis algorithm and is one of the most classic algorithms for mining association rules [25,26]. This algorithm innovatively uses support measure-based pruning technique to restrain the exponential growth of the number of candidate itemsets, uses the strategy of level-wise traversal and "generate-test"

to generate frequent itemsets, and uses the Apriori principle to compress the search space. But its disadvantages include too many database scans and the generation of a large number of unnecessary itemsets [27]. Unlike Apriori algorithm, FP-growth algorithm uses FP-tree to generate frequent itemsets [28]. For some typical transaction sets, FP-growth algorithm is several orders of magnitude faster than the standard Apriori algorithm [29,30]. FP-tree is a compressed representation of input data that is constructed by reading transactions one by one and mapping each transaction to a path in the FP-tree. Because different transactions may have several items in common, their paths may partially overlap. The more paths overlap each other, the better the compression with FP-tree [31]. For remote sensing user demand mining, the good applicability of FP-growth algorithm is mainly reflected in the compression of data with limited range of observation time, payload type and other items. The larger the amount of data, the higher the degree of compression. If the FP-tree is small enough to be stored in memory, one can find frequent itemsets directly from the memory, rather than having to repeatedly scan the data stored on the hard disk.

In this paper, a novel association rule mining algorithm based on directed term graph-based FP-tree is proposed to mine for the association rules among users, observation targets, payloads and observation dates from large quantities of historical remote sensing demand data. The remainder of this paper is organized as follows: In Sect.1, the issue of remote sensing user demand mining is described and its characteristics are analyzed. In Sect.2, related concepts and theorems involved in this paper are introduced. In Sect.3, according to the logical relationships between remote sensing demand items, a directed item graph is designed, a double-layer FP-tree based on the directed item graph is presented, and an association analysis method of remote sensing user demand based on FP-growth algorithm is proposed. In Sect.4, the proposed method is verified, and the conclusions are drawn in Sect. 5.

II. The Description of Remote Sensing User Demand Mining Problem

For a remote sensing demand, what the user cares about mainly include observation targets, payload types and observation time, but the status of these three demand items is not unequal. The user is most concerned about the observation target, followed by what payload and when to observe. Therefore, in a remote sensing demand transaction, there are four items: user, target, payload and observation time. Remote sensing user demand mining is different from conventional Market Basket Analysis, which is mainly reflected in the following five points:

1. Remote sensing serves the user; thus, the user is the most important.

2. For users, they are most interested in the observation target and will decide the type of payload and observation time.

3. The target, the payload and the observation time are not completely independent. The type of payload depends on weather and what kind of target properties the user is interested in. The observation time may be related to the season, the satellite orbit motion and the periodic change of the target.

4. There may be some correlations among observation targets, that is, users' demand may need observations of many different targets to be satisfied.

Here, some examples of remote sensing user demand data are given in Table 1.

Table 1 Remote sensing user demand data

No.	User	Observation target	Payload type	Observation date
1	User A	Target 1	CCD	Aug. 3, 2019
2	User A	Target 1	CCD, SAR	Aug. 5, 2020
3	User B	Target 2	SAR	Sept. 27, 2020
4	User A	Target 2	SAR	Sept. 16, 2020

According to the data in the above table, we can intuitively find out some simple results or rules: (a) User A has the most demands for remote sensing data and needs to be serviced preferentially; (b) User A observed Target 1 for two consecutive years, and the observation time was in early August, and the optical payload was adopted for both of them; (c) Target 2's observations were usually made in the second half of the year with a SAR payload. (d) Both User A and User B have observation demands for Target 2 and we can recommend remote sensing data services of Target 1 to User B.

Given the small amount of data in Table 1, not all of the above analysis results may be necessarily meaningful. However, so many rules can be drawn artificially from the only four remote sensing demands. It is conceivable that if large quantities of historical remote sensing user demand data are mined, a large number of meaningful information will be obtained.

III. Related Concepts of Remote Sensing User Demand Mining

3.1. Itemset

Let $I=\{i_1,i_2,\cdots,i_m\}$ be the set of all items. An itemset containing *k* items is called a *k*-itemset. Let $T=\{t_1,t_2,\cdots,t_N\}$ be the set of all transactions, each of which contains an itemset that is a nonempty subset and corresponds to a unique identifier labeled TID (Transaction ID). The width of a transaction represents the number of items in the transaction. If the itemset *X* is a subset of transaction t_j, then transaction t_j is said to contain the itemset *X*. An important indicator of an itemset is its support count, that is, the number of transactions containing a particular itemset, expressed as:

$$count(X)=\left|\{t_i \mid X\subseteq t_i, t_i\in T\}\right| \tag{1}$$

Where $|\cdot|$ represents the number of elements in the set.

For example, in Table 1, there are four transactions, the set of all items *I* = {User A, User B, Target 1, Target 2, CCD, SAR, 2020, 2019, Aug. Sept., Oct.}, the " No." is TID, and all these transaction are 5-itemsets. If itemset *X* = {User A, Target 1}, then $count(X)=2$.

3.2. Support and confidence

An association rule is an implication expression of the form $X\to Y$, where symbol $\to$ represents association operation, and *X* and *Y* are disjoint itemsets, i.e., $X\cap Y=\varnothing$. *X* and *Y* are called Left-Hand-Side (LHS) and Right-Hand-Side (RHS) of association rules respectively. The strength of an association rule can be measured by its support and confidence. Support determines how often a rule is applicable to a given data set, that is, probability, while confidence determines how frequently itemsets in *Y* appear in transactions that contain *X*, that is, conditional probability. Support and confidence can be calculated by the following two expressions:

$$sup(X\to Y)=\frac{count(X\cup Y)}{N} \tag{2}$$

$$conf(X\to Y)=\frac{count(X\cup Y)}{count(X)} \tag{3}$$

For example, in Table 1, if X = {User A} and Y = {Target 1}, then the support and confidence of the association rule are respectively:

$$sup\left(\{\text{User A}\} \rightarrow \{\text{Target 1}\}\right) = \frac{count\left(\{\text{User A, Target 1}\}\right)}{N} = \frac{2}{4} = 0.5 \tag{4}$$

$$conf\left(\{\text{User A}\} \rightarrow \{\text{Target 1}\}\right) = \frac{count\left(\{\text{User A, Target 1}\}\right)}{count\left(\{\text{User A}\}\right)} = \frac{2}{3} = 0.67 \tag{5}$$

Support is an important measure and has desirable property which can be used to eliminate meaningless rules and realize effective discovery of association rules. Confidence measures the reliability of the inference made by an association rule. For a given association rule $X \rightarrow Y$, the higher the confidence, the more likely it is that Y is contained in the transaction that contains X.

It can be seen from the above definitions that association rule analysis method does not need data processing in advance. The analysis result completely based on the original data can truly and objectively reflect the essential relationship between data with a strong persuasive. It is worth mentioning that the inference derived from an association rule does not necessarily imply causality, but only suggests the simultaneous occurrence of the LHS and RHS. The association rule mining problem can be stated as followed:

Definition 1. *Association rule discovery refers to finding all the rules which satisfy support* $\geq sup_{min}$ *and confidence* $\geq conf_{min}$ *in a given set of transactions T, where* sup_{min} *and* $conf_{min}$ *represent the minimum support threshold and the minimum confidence threshold, respectively.*

In the definition above, sup_{min} and $conf_{min}$ are usually given by the user. It should be noted that the setting of supmin should be reasonable. Too small sup_{min} will generate a lot of accidental association rules and increase the computational requirements, while too large sup_{min} will easily lose meaningful association rules. The minimum support threshold sup_{min} reflects the lowest statistical importance of the itemset, while the minimum confidence threshold $conf_{min}$ reflects the lowest reliability of the association rule.

In order to improve the performance of association analysis algorithm, the mining process is usually divided into two parts according to support and confidence, namely frequent itemset generation and association rule generation.

3.3. Frequent itemset generation

The objective of frequent itemset generation is to find all the itemsets that satisfy the minimum support threshold, which are referred to as frequent itemsets. The itemset which is used to generate frequent itemset is called candidate itemset. In general, the computational requirements for frequent itemset generation are more expensive than those of rule generation. Therefore, the computation for generating frequent itemset can be reduced by reducing the number of candidate itemsets and the number of comparisons between candidate itemsets and transactions. To better understand the process of frequent itemset generation, two significant theorems are given as follows.

Theorem 1. *The necessary condition for the k-itemset to be frequent is that all its subsets are frequent.*

Theorem 2. *If any subset of the k-itemset is not frequent, then the k-itemset is not frequent.*

3.4. Association rule generation

Finding association rules from frequent itemsets is association rule generation. The objective of association rule generation is to mine for all rules with high confidence from the frequent itemsets generated in the previous step. These rules are called strong rules. For frequent *k*-itemset, a total of 2^k-2 association rules can be generated.

For example, frequent itemset *Y* is divided into two nonempty itemsets *X* and $Y-X$, and $X \to Y-X$ satisfies the minimum confidence threshold, then *X* and $Y-X$ is a strong rule. Since rules are generated by frequent itemsets, they have already satisfied the support threshold. In addition, the support counts for these two itemsets have already obtained when the frequent itemsets are generated, so there is no need to scan the entire transaction sets again. The theorem holding for the confidence measure is as follows [32].

Theorem 3. *If the rule $X \to Y-X$ does not satisfy the confidence threshold, then the rule $X' \to Y-X'$ must not satisfy the confidence threshold as well, where X' is a subset of X.*

Although association rules determined only on the basis of support and confidence are persuasive to a certain extent, the limitation of confidence measure lies in that it ignores the support of itemset in the rule RHS, and the rule with high confidence may lead to misdirection [33]. In addition, the data's dimension and volume in the actual database are always huge, which makes it easy to generate a large number of meaningless or uninterested association rules. Therefore, some other objective measures are usually needed to further filter the association rules generated based on support and confidence. In this paper, a measure called lift is adopted to further evaluate the generated association rules and defined as follows:

$$lift\left(X \to Y\right) = \frac{conf\left(X \to Y\right)}{sup\left(Y\right)} = \frac{N \cdot count\left(X \cup Y\right)}{count\left(X\right) \cdot count\left(Y\right)} \tag{6}$$

The above expression represents the ratio of the probability of the occurrence of *Y* on the basis of the occurrence of *X* to the probability of the occurrence of *Y* alone.

IV. Mining and Analysis of Remote Sensing User Demand

Considering many repeated items in the historical remote sensing user demand data, this paper adopts FP-growth algorithm to analyze the remote sensing user demand. FP-growth algorithm uses a compact data structure called FP-tree to organize data and can extract frequent itemsets directly from this structure. This algorithm is completely different from the frequent itemset generation method of Apriori algorithm.

4.1. Directed item graph-based double-layer fp-tree

Different from the conventional Market Basket Analysis, not all the items in the remote sensing demand data are related. According to the prior information of remote sensing demand data characteristics, unassociated items can be pruned in advance to reduce the amount of calculation and eliminate meaningless association rules. In this paper, a directed item graph is designed, which reflects the directed dependency relationships between each item in the remote sensing demand transaction.

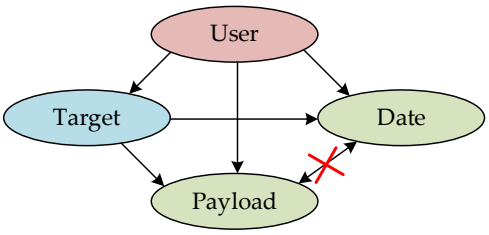

Fig 1: Directed item graph of remote sensing demand items.

In this graph, the observation target is determined by the user, the payload is determined by both the user and the target, and the observation date is determined by both the user and the target too. There is no necessary connection between the payload and the observation date. For example, the user determines when and use which payload to observe which target, and the target characteristics determine when and use which payload can make the observation effect best. However, such logical relations are not valid conversely, thus the graph in Figure 1 is directed.

Based on the above analysis, a double-layer FP-tree based on directed item graph is presented here. The purpose of constructing a FP-tree with double layers is to cut off the relationship between the payload and the observation date. The conventional generation method of FP-tree is based on the support count of frequent 1-itemsets, but taking the uniqueness of remote sensing demand transactions into account, the generation of FP-tree in this paper is based on both the logical relationship of items in Figure 1 and the frequent 1-itemsets, that is, only the items with connections may have associations. In the process of remote sensing demand mining, the demand of users is the most important, followed by target properties that can be mined from the historical remote sensing user demand data.

Next, take the demand data in Table 1 as an example to generate a double-layer FP-tree proposed in this paper. The generation process of this double-layer FP-tree is shown in Figure 2, where Target 1 and Target 2 are abbreviated to "T1" and "T2", respectively.

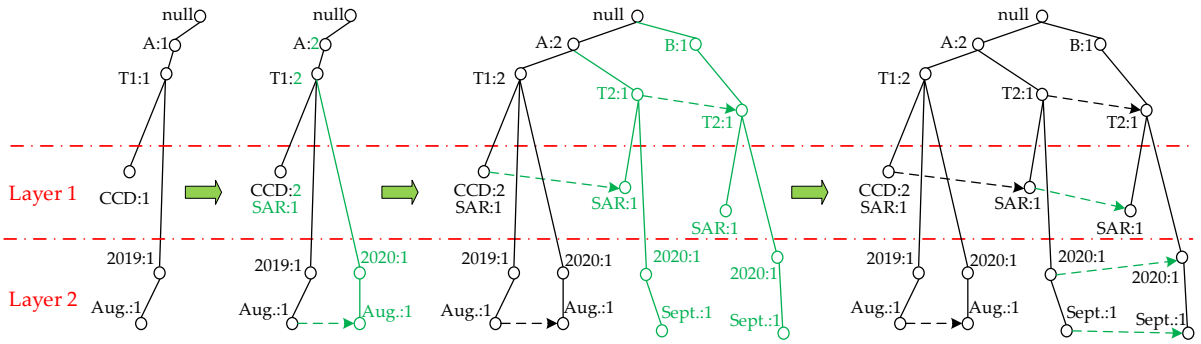

Fig 2: The generation process of the directed item graph-based double-layer FP-tree.

Initially, the FP-tree contains only one root node, represented by" null". Then, according to each transaction in Table 1, the FP-tree is expanded as follows:

1. Scan the data set, determine the support count for each 1-itemset, and discard the infrequent itemset according to the minimum support threshold. Due to the small amount of data in the figure above, the support threshold was not considered here in the process of generating FP-tree.

2. Scan each transaction in the data set again and build the FP-tree. Read the first transaction {User A, Target 1, CCD, 2019, Aug.} and create node "User A" and "T1". Then create node "CCD" in the first layer and create node "2019" and "Aug." in the second layer, forming path null → User A → Target 1 → CCD and path null → User A → Target 1 → 2019 → Aug. All nodes on this path are counted as 1.

3. Read the second transaction {User A, Target 1, CCD, SAR, 2020, Aug.}. Since this transaction shares nodes "User A" and "T1" with the first transaction, it is only necessary to create nodes "2020" and "Aug." It can be seen that the path "null → User A → Target 1" overlaps between the first transaction and the second transaction, which reflects the compressibility of FP-tree for data.

4. Each transaction is read repeatedly and the corresponding path is generated according to the above process. Finally, in order to mine the properties of the observation target, the same items on the same target path need to be connected, and finally the FP-tree on the far right in Figure 2 is formed.

It can be seen from Figure 2 that, firstly, the first layer of the generated double-layer FP-tree is the payload type and the second layer is the observation date. These two layers are connected with users and observation targets, but they are not connected with each other. Secondly, the observation targets required by different users may be the same. For example, both User A and User B have observation demands for Target 2. In order to mine for the properties of Target 2, the same items associated with Target 2 need to be connected regardless of the demands of users, as shown by the green dashed arrow in the sub-figure on the far right.

4.2. Frequent itemset generation based on fp-growth algorithm

FP-growth is an algorithm that generates frequent itemsets from FP-tree by exploring the tree in a bottom-up fashion. FP-growth algorithm uses the divide-and-conquer strategy to decompose a problem into smaller sub-problems. Since the sub-problems are disjoint, FP-growth algorithm will not generate any repeated itemset.

In this paper, based on the double-layer FP-tree generated in the previous section, FP-growth algorithm is adopted to find frequent itemset. Due to the directed dependency relationship among items in remote sensing user demand transactions (as shown in Figure 1), the generated double-layer FP-tree needs to be decomposed. The decomposition of double-layer FP-tree is mainly divided into two steps. The first step is user-based and target-based decomposition, and the second step is layer-level decomposition. The decomposition process is shown in Figure 3.

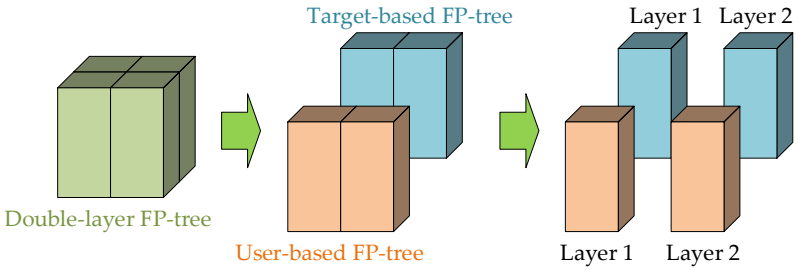

Fig 3: The decomposition process of double-layer FP-tree.

User-based decomposition is mainly used to mine for the remote sensing demands of every user, while target-based decomposition is mainly used to mine for the target properties. Layer-level decomposition is mainly used to separate payloads and observation dates. Taking Figure 2 as an example, the first step is to separate User A from the double-layer FP-tree, as shown in the left figure in Figure 4. In the second step, the double-layer FP-tree of User A is further decomposed into two single-layer FP-trees, as shown in the two figures on the right in Figure 4.

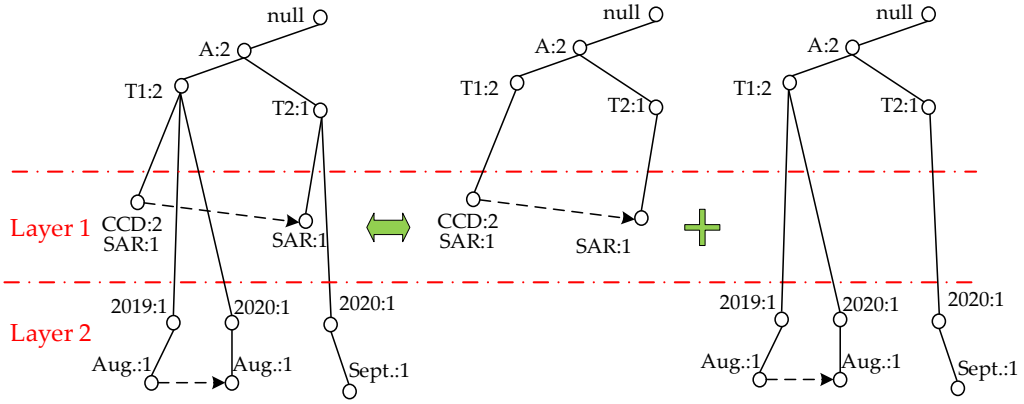

Fig 4: User A-based double-layer FP-tree decomposition.

FP-growth algorithm is used to generate the frequent itemsets of the two FP-trees on the right in Figure 4 respectively, and the combination of the two frequent itemsets generated from these two trees is the frequent itemsets of User A. Two points should be noted in the generation process of frequent itemsets: (a) in the first layer, only the frequent itemsets with payload suffix are generated, and the frequent itemsets in the second layer are generated normally; (b) Since this FP-tree is for User A, each generated frequent itemset needs to contain item "User A". It is not difficult to find that the generation process of frequent itemsets does not produce repeated itemsets. Figure 5 shows the remaining decomposition results of the double-layer FP-tree in Figure 2, and these sub-FP-trees are the decomposition results based on User B, Target 1 and Target 2, respectively, from left to right.

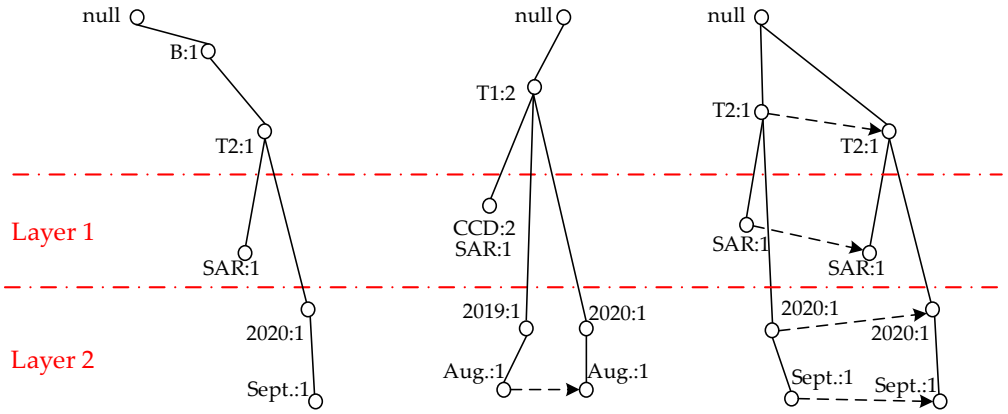

Fig 5: Double-layer FP-tree decomposition results based on User B, Target 1 and Target 2, respectively.

After the frequent itemset generation using the above method, the association rule generation method introduced in Sect. 2 can be used for further remote sensing user demand mining.

V. Transaction Case Analysis

In this section, the proposed data mining method based on the directed item graph-based double-layer FP-tree is used in the association rule analysis of a group of historical remote sensing demand cases. Set the minimum support threshold as sup_{min} = 0.2 and the minimum confidence threshold as $conf_{min}$ = 0.5. The historical remote sensing user demand transactions are listed in Table 2.

Table 2 Remote sensing user demand transactions

TID	User	Observation target	Payload type	Observation date
1	User A	Target 1	CCD, SAR	Sept. 1, 2020

2	User A	Target 1	CCD, SAR	Aug. 24, 2020
3	User A	Target 2	SAR	Jul. 10, 2019
4	User A	Target 2	SAR	Jul. 2, 2020
5	User B	Target 2	CCD	Jul. 20, 2020
6	User B	Target 2	CCD	Jul. 5, 2020
7	User B	Target 4	CCD	Jul. 8, 2020
8	User C	Target 1	CCD, SAR	Sept. 15, 2019
9	User C	Target 3	Elec.	Aug. 21, 2019
10	User D	Target 3	Elec.	Aug. 21, 2019

First, scan the entire transaction data and record the support count for each item, as shown in Table 3.

Table 3 The support count of each 1-itemset

Item	Count	Item	Count	Item	Count	Item	Count
User A	4	Target 1	3	2020	6	Sept.	2
User B	3	Target 2	4	2019	4	CCD	6
User C	2	Target 3	2	Jul.	5	SAR	5
User D	1	Target 4	1	Aug.	3	Elec.	2

In order to make the full use of data, according to the minimum support threshold, only the items with count of 1 in Table 2 are removed, and the others of the corresponding transaction are retained, as shown in Table 4.

Table 4 Two transactions with infrequent 1-itemsets

TID	User	Observation target	Payload type	Observation date
7	User B	/	CCD	Jul. 8, 2020
10	/	Target 3	Elec.	Aug. 21, 2019

Based on this, the double-layer FP-tree of the historical remote sensing demand transactions is established, as shown in the Figure 6.

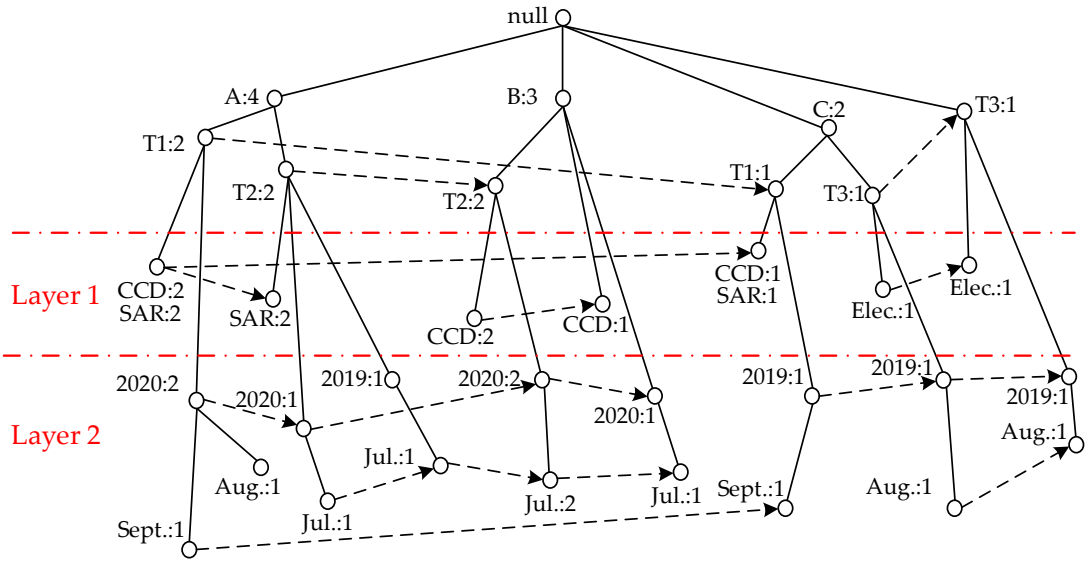

Fig 6: Double-layer FP-tree of historical remote sensing demand transactions.

Next, based on the demand mining method proposed in this paper, we take the user and the target with the most support counts respectively as the examples to conduct association analysis. For users, 1-itemset {User A} has the

most support counts, then decompose the double-layer FP-tree corresponding to User A by layer-level, as shown in the Figure 7.

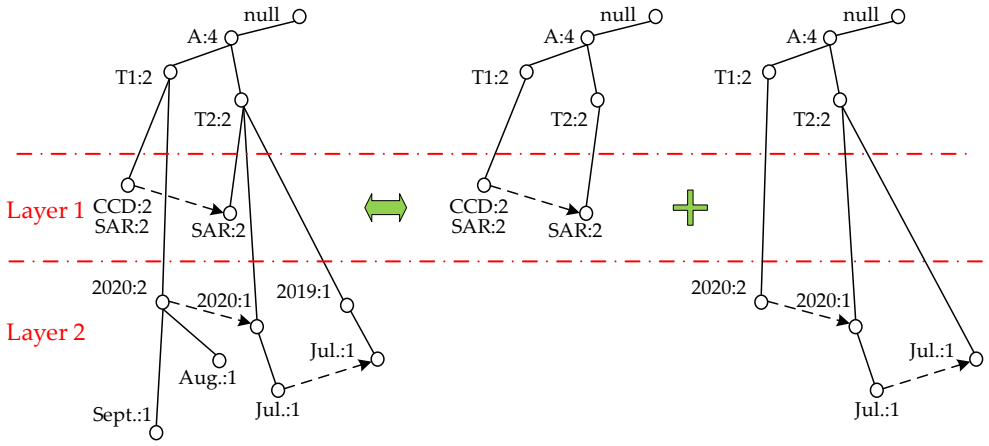

Fig 7: The decomposition of User A-based double-layer FP-tree.

For the decomposition result of User A-based double-layer FP-tree in Figure 7, it can be seen that the path containing itemset {Sept.} and the path containing itemset {Aug.} has only one respectively, or the support counts of these two 1-itemsets do not satisfy the minimum support threshold. Thus, itemset {Sept.} or {Aug.} are pruned in the process of layer-level decomposition. Then, the FP-growth algorithm is adopted to generate frequent itemsets of these two sub-FP-trees shown in the right two sub-figures. The generated frequent itemsets are shown in Table 5, where the number after the colon represents the support count.

Table 5 Frequent itemsets with regard to User A

Suffix	Frequent itemset
SAR	{User A, SAR}:4, {User A, Target 1, SAR}:2, {User A, Target 2, SAR}:2
CCD	{User A, CCD}:2, {User A, Target 1, CCD}:2
T1	{User A, Target 1}:2
T2	{User A, Target 2}:2
Jul.	{User A, Jul.}:2, {User A, Target 2, Jul.}:2
2020	{User A, 2020}:3, {User A, Target 1, 2020}:2

Since the frequent itemsets in the above table are all for User A, each frequent itemset must contain itemset {User A}. For targets, itemset {Target 2} has the most support counts, then decompose the double-layer FP-tree corresponding to Target 2 by layer-level, as shown in the Figure 8.

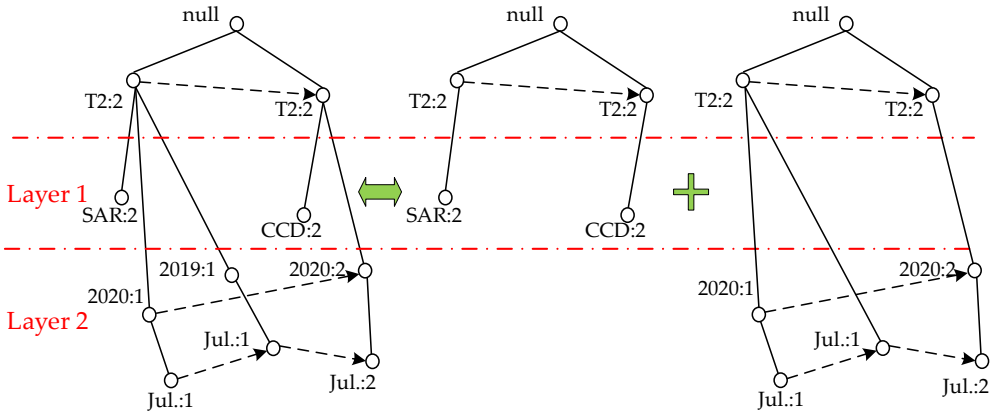

Fig 8: The decomposition of Target 2-based double-layer FP-tree.

Then, the FP-growth algorithm is adopted to generate the frequent itemsets of these two sub-FP-trees. The generated frequent itemsets are shown in Table 6. Since these frequent itemsets are all for Target 2, each frequent itemset must contain the itemset {Target 2}.

Table 6 Frequent itemsets with regard to Target 2

Suffix	Frequent itemset
SAR	{Target 2, SAR}:2
CCD	{Target 2, CCD}:2
Jul.	{Target 2, Jul.}:4, {Target 2, 2020, Jul.}:3
2020	{Target 2, 2020}:3

According to the association rules shown in Table 5 and Table 6, although there is an overlapping itemset {Target 2} between User A-based and Target 2-based double-layer FP-trees, there is no repeated frequent itemset generated, which verifies the rationality and effectiveness of the proposed directed item graph-based double-layer FP-tree. The generation process of frequent itemsets for User B, User C, Target 1 and Target 3 is the same as that for User A and Target 2, thus this paper will not repeat it here and only gives the generation results of these frequent itemsets, as shown in Table 7.

Table 7 Frequent itemsets with regard to User B, User C, Target 1 and Target 3

Suffix	Frequent itemset
SAR	{Target 1, SAR}:3
CCD	{User B, CCD}:3, {User B, Target 2, CCD}:2, {Target 1, CCD}:3
Elec.	{Target 3, Elec.}:2
T2	{User B, T2}:2
Jul.	{User B, Jul.}:3, {User B, 2020, Jul.}:3, {User B, Target 2, Jul.}:2, {User B, Target 2, 2020, Jul.}:2
Aug.	{Target 3, Aug.}:2, {Target 3, 2019, Aug.}:2
Sept.	{Target 1, Sept.}:2
2019	{User C, 2019}:2, {Target 3, 2019}:2
2020	{User B, 2020}:3, {User B, Target 2, 2020}:2, {Target 1, 2020}:2

Based on the discovered frequent itemsets, association rules can be obtained by calculate their support, confidence and lift. Some representative association rule analysis results are presented in Table 8.

Table 8 The analysis results of the generated association rules

No.	Association rule	Support	Confidence	Lift	Analysis result
1	{User A} → {SAR}	0.4	1	2	All demands of User A contain SAR, and in all SAR requirements, User A has a high proportion, which indicates that User A is interested in SAR images.
2	{Target 2} → {Jul.}	0.4	1	2	All observations of Target 2 are in July.
3	{User A} → {Target 2, Jul.}	0.2	0.5	1.25	User A usually requires Target 2 observation in July.
4	{User B} → {T2, 2020, Jul.}	0.2	0.67	2.23	The demands of User B are mainly Target 2 in Jul. 2020.
5	{Target 3} → {Elec.}	0.2	1	5	Demands of Target 3 are all required Elec. payload.

In these association rules, the support and the confidence all satisfy the thresholds sup_{min} = 0.2 and $conf_{min}$ = 0.5, in addition, the lift measure is larger than 1, both of which illustrate the given association rules are meaningful.

6. Conclusions

In this paper, a novel association rule analysis method based on directed item graph-based double-layer FP-tree is proposed to address the data mining issue of the remote sensing user demand. This method features the discovery of the intrinsic relationship between demand items from a huge number of historical remote sensing demand data. Firstly, by analyzing the logical relationship between demand items, the directed term graph of user, target, payload and observation date is designed. Secondly, a double-layer FP-tree based on the directed term graph is presented, and an association rule analysis method based on FP-growth algorithm is proposed. Finally, a group of historical remote sensing user demand transactions are calculated and analyzed with the proposed method, and its effectiveness is verified. The advantages of the proposed method are mainly reflected in the compression of remote sensing demand data by FP-tree, scanning the data set only twice by FP-growth algorithm, no need to generate candidate itemsets and no repeated frequent itemset generation. In addition, the directed term graph has the function of pre-pruning for the construction of double-layer FP-tree, and such a tree features short search path and small search range. For a large number of remote sensing demand in the future, the proposed demand mining method may be an effective one to analyze remote sensing user demands and observation target characteristics, which can be used in the recommendation and management of remote sensing data services.

Acknowledgements

Author Contributions: Conceptualization, Z.Z. and J.Z.; methodology, Z.Z., J.Z and H.Y.; validation, Z.Z., J.Z. and H.Y.; investigation, Z.Z. and J.Z.; writing—original draft preparation, Z.Z. and J.Z.; writing—review and editing, J.Z. and H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Y. Hajjaji, W. Boulila, I.R. Farah, I. Romdhani, A. Hussain, “Big data and IoT-based applications in smart environments: A systematic review,” Comput. Sci. Rev., vol. 39, pp. 100318, 2021.
[2] X. Pan, F. Yang, L. Gao, Z. Chen, B. Zhang, H. Fan, J. Ren, “Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms,” Remote Sens., vol. 11, pp. 917, 2019.
[3] M. Alkhelaiwi, W. Boulila, J. Ahmad, A. Koubaa, M. Driss, “An Efficient Approach Based on Privacy-Preserving Deep Learning for Satellite Image Classification,” Remote Sens., vol. 13, pp. 2221, 2021.
[4] B.U. Choudhury, A. Sood, S.S. Ray, P.K. Sharma, Panigrahy, “Agricultural Area Diversification and Crop Water Demand Analysis: A Remote Sensing and GIS Approach,” Journal of the Indian Society of Remote Sensing, vol. 41, pp. 71-82, 2013.
[5] S. Xu, J. Chen, G. Gao, “Remote sensing ocean data analyses using fuzzy C-Means clustering,” International Society for Optics and Photonics, 2009.
[6] R.P. Sishodia, R.L. Ray, S.K. “Singh, Applications of Remote Sensing in Precision Agriculture: A Review,” Remote Sens., vol. 12, pp. 3136, 2020.
[7] C. Quintano, A. Fernández-Manso, A. Stein, W. Bijker, “Estimation of area burned by forest fires in Mediterranean countries: A remote sensing data mining perspective,” Forest Ecology and Management, vol. 262, pp. 1597-1607, 2011.

[8] I. Dochev, P. Gorzalka, V. Weiler, J.E. Schmiedt, “Calculating Urban Heat Demands: An analysis of two modelling approaches and remote sensing for input data and validation,” Energy and Buildings, vol. 226(2020), pp. 110378, 2020.

[9] Q. Tong, Y. Xue, L. Zhang, “Progress in Hyperspectral Remote Sensing Science and Technology in China Over the Past Three Decades,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, pp. 70-91, 2014.

[10] I. Chebbi, N. Mellouli, I.R. Farah, M. Lamolle, “Big Remote Sensing Image Classification Based on Deep Learning Extraction Features and Distributed Spark Frameworks,” Big Data Cogn. Comput., vol. 5, pp. 21, 2021.

[11] M. Wurm, A. Droin, T. Stark, C. Geiß, W. Sulzer, H. Taubenböck, “Deep Learning-Based Generation of Building Stock Data from Remote Sensing for Urban Heat Demand Modeling,” ISPRS Int. J. Geo-Inf., vol. 10, pp. 23, 2021.

[12] R. Anuradha, N. Rajkumar, “A Novel Approach in Mining Specialized Coherent Rules in a Level-Crossing Hierarchy,” International Journal of Fuzzy Systems, vol. 19, pp. 1782–1792, 2017.

[13] A.K. Chandanan, M.K. Shukla, “Removal of Duplicate Rules for Association Rule Mining from Multilevel Dataset,” Procedia Computer Science, vol. 45, pp. 143-149, 2015.

[14] M.A. Bhuiyan, M.A. Hasan, “An Iterative MapReduce Based Frequent Subgraph Mining Algorithm,” IEEE Transactions on Knowledge & Data Engineering, vol. 27, pp. 608-620, 2015.

[15] X. Zhao, Y. Chen, C. Xiao, Y. “Ishikawa, J. Tang, Frequent Subgraph Mining Based on Pregel,” The Computer Journal, vol. 59, pp. 1113-1128, 2016.

[16] C. Jiang, F. Coenen, M. Zito, “A Survey of Frequent Subgraph Mining Algorithms,” The Knowledge Engineering Review, vol. 28, pp. 75-105, 2013.

[17] H. Daschiel, M. Datcu, “Information Mining in Remote Sensing Image Archives: System Evaluation,” IEEE Transactions on Geoscience and Remote Sensing, vol. 43, pp. 188-199, 2005.

[18] J. Han, Y. Fu, “Mining Multiple-Level Association Rules in Large Databases,” IEEE Transactions on Knowledge and Data Engineering, vol. 11, pp. 798-805, 1999.

[19] T. McIntosh, S. Chawla, “High Confidence Rule Mining for Microarray Analysis,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, pp. 611-623, 2007.

[20] P.N. Tan, V. Kumar, “Mining Association Patterns in Web Usage Data,” In Proc. of the Intl. Conf. on Advances in Infrastructure or e-Business, e-Education, e-Science and e-Medicine on the Internet, L’ Aquila, Italy, January 2002.

[21] D. Barbará, J. Couto, S. Jajodia, N. Wu, “ADAM: A Testbed for Exploring the Use of Data Mining in Intrusion Detection,” SIGMOD Record, vol. 30, pp. 15-24, 2001.

[22] H. Xiong, X. He, C. Ding, Y. Zhang, V. Kumar, S.R. Holbrook, “Identification of Functional Modules in Protein Complexes via Hyperclique Pattern Discovery,” In Proc. of the Pacific Symposium on Biocomputing, Maui, January, 2005.

[23] J. Pei, J. Han, B. Mortazavi-Asl, H. Zhu, “Mining Access Patterns Efficiently from Web Logs,” In Proc. of the 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, Kyoto, Japan, April, pp. 396-407, 2000.

[24] R. Agrawal, T. Imielinski, A. Swami, “Database Mining: A performance perspective,” IEEE Trans. on Knowledge and Data Engineering, vol. 5, pp. 914-925, 1993.

[25] M.J. Zaki, “Generating Non-Redundant Association Rules,” In Proc. of the 6th Intl. Conf. on Knowledge Discovery and Data Mining: Boston, MA, August, pp. 34-43, 2000.

[26] R. Agrawal, T. Imielinski, A. Swami, “Mining association rules between sets of items in large databases,” In Proc. ACM SIGMOD Intl. Conf. Management of Data, Washington, DC, pp. 207-216, 1993.

[27] B. Dunkel, N. Soparkar, “Data Organization and Access for Efficient Data Mining,” In Proc. of the 15th Intl. Conf. on Data Engineering, Sydney, Australia, March, pp. 522-529, 1999.

[28] M.J. Zaki, “Efficiently Mining Frequent Trees in a Forest,” In Proc. of the 8th Intl. Conf. on Knowledge Discovery and Data Mining, Edmonton, Canada, July, pp. 71-80, 2002.

[29] J. Han, J. Pei, Y. Yin, “Mining Frequent Patterns without Candidate Generation,” In Proc. ACM-SIGMOD Int. Conf. on Management of Data, Dallas, TX, May, pp. 1-12, 2000.

[30] J. Han, J. Pei, Y. Yin, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach,” Data Mining and Knowledge Discovery, vol. 8, pp. 53-87, 2004.

[31] R.C. Agarwal, C.C. Aggarwal, V.V.V. Prasad, “A Tree Projection Algorithm for Generation of Frequent Itemsets,” Journal of Parallel and Distributed Computing (Special Issue on High Performance Data Mining), vol. 61, pp. 350-371, 2001.

[32] P.N. Tan, M. Steinbach, V. Kumar, A. “Karpatne, Introduction to Data Mining,” Posts & Telecom Press 2006.

[33] S. Brin, R. Motwani, C. “Silverstein, Beyond and Market Baskets: Generalizing association rules to correlations,” In Proc. ACM SIGMOD Intl. Conf. Management of Data, Tucson, AZ, pp. 265-276, 1997.