Research on Remote Sensing User Demand Mining Analysis Method

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

A method of mining and analyzing remote sensing user needs is presented in this paper. With the increasing number and types of satellites in China, satellites are more and more widely used in national economic construction, which can be used in land and resources census, emergency natural disaster relief, urban construction management, earth environment monitoring and so on. Therefore, the number of remote sensing users expand quickly and the specific content of remote sensing observation needs and remote sensing image data needs is diverse. How to find out the user's remote sensing resources and data use rules from the large number of remote sensing needs is a key problem. The difficulty lies in that the rules representing the relationship between demand parameters are recessive rather than explicit, which can only be found by complex data analysis, especially under the condition of large amount of data. Data mining technology can find potential as well as valuable rules and knowledge from massive data. It is an effective tool to analyze massive data and an effective technical means to solve the above problems. In this paper, the association rules mining algorithm is adopted to mine and analyze mass of remote sensing user demand data, and find the potential association rules between the demand source, regional scope, time, demand type and other parameters.

Keywords: Remote sensing users, remote sensing demand, data mining, association analysis

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. Remote sensing satellite, as an extremely important strategic resource, has achieved a dominant role in many fields such as the land and resources survey, natural disaster emergency rescue, the urban construction and management, the earth's environment monitoring and so on. By mining and analyzing the accumulated users' remote sensing needs and users' remote sensing image data needs, the association rules among the demand source, the range of earth observation area, the time of earth observation, the type of sensor and other parameters can be found. According to these knowledge rules, users' needs can be automatically generated in advance, and the automatic intelligent submission of users' needs can be realized, so that the regular needs of users can be submitted to the satellite collaborative application management and control department as soon as possible. For example, users of our aid units in Sudan will submit remote sensing mission requirements in August every year in view of road damage in Sudan. According to this rule, observation mission requirements can be generated in advance and submitted to remote sensing satellite data exchange center as soon as

possible, so as to obtain greater opportunities for observation resource arrangement; In addition, users' interests and preferences in applying for remote sensing data can be found. For example, users applying for remote sensing images of Subi Reef in the South China Sea usually apply for remote sensing images of Huangyan Island and Mischief Reef there. According to this rule, users applying for remote sensing images of Subi Reef can actively recommend remote sensing images of Huangyan Island and Mischief Reef.

Due to the large number of users' demands, the specific contents of users' remote sensing task demands and remote sensing image data demands are diverse. It is a difficult problem to find out the rules of users' remote sensing resource use and data use from a large number of users' demands. The difficulty lies in that the rules representing the relationship between the demand parameters are implicit rather than explicit, which can only be found through complex data analysis, especially under the condition of large data scale. Data mining technology can find potential and valuable rules and knowledge from massive data. It is an effective tool to analyze massive data and an effective technical means to solve the above problems.

Next, the association rules mining algorithm of data mining technology will be used to mine and analyze large bodies of user demand to find the potential association rules between the demand source, regional scope, time, demand type and other parameters.

II. Classical Association Rules Mining Algorithm

2.1 Association analysis in remote sensing demand mining

Data mining is mainly through the purposeful extraction, sorting and classification of massive data information, mining the hidden useful information, which is used to provide decision support information for the business development of all walks of life. Association rule mining is a kind of problem of find the Internal connection between itemsets on large transaction data sets, which is a significant method and research field in data mining [1-4].

The association rules analysis method does not require any data analysis, and the results obtained are completely based on the data without any subjective assumption, which can truly and objectively reflect the nature of the data and have a strong conviction. The results obtained from the association rule analysis of the data can be regarded as the summary of the potential laws among variables in the data.

Association analysis is a popular method in data mining. Now many work involves the exploration of association rules theory, the perfection of previous algorithm and the design of new algorithm based on the latest application. This paper makes unremitting efforts in improving the efficiency and application of mining rule algorithm.

Let $I = \{I_1, I_2, \text{L } I_m\}$ be the set of all items. D is the data collection of all transactions, T is the collection of some items, T is included in I, and each transaction can be identified by a unique identifier TID. In this paper, D is the requirement of observation task set or remote sensing image data. Each record in the observation task set is an observation task requirement, including the requirement source, the range of the earth observation area, the earth observation time, the sensor type, the resolution and other parameter attributes; Each record is a remote sensing user demand, including the requirement source, image data number, data time, data type, data coverage area and other parameter attributes. $i_k = (k = 1, 2, \text{L } m)$ represents a specific value of an attribute. For example, for the observation task requirement, the requirement source can be "Peking University"; for the requirement of remote sensing image data, the number can be "123456".

The rule $X => Y$ in D is determined by support s and confidence c. 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.

Assume that the minimum support threshold and confidence threshold are respresented by $support_{min}$ and $confidence_{min}$, respectively. The user generally decides these thresholds. The rules which satisfies these corresponding thresholds are called strong association rules, conversely called weak association rules. The task of discovering association rules is to discover the effective rules with confidence and support greater than or equal to a given threshold from the database.

Based on the above concepts, the following conclusions can be drawn:

(a) The necessary condition for a k-itemset X_k to be a frequent itemset is that all (k-1)-itemsets it contains are frequent itemsets too, denoted as X_{k-1};

(b) If any (k-1)-itemset X_{k-1} of a k-itemset X_k is not a frequent itemset, then the k-itemset X_k itself is not the largest data itemset.

The mining association rules problem can be divided into two as follows:

(a) Find all itemsets in transaction database D that are greater than or equal to the minimum support, and the itemset with the minimum support is known as the maximum itemset. The support of a itemset refers to the number containing the itemset.

(b) Once the frequent itemsets are found from the transactions in database D, the following formula can be adopted to generated association rules:

$$confidence(A => B) = \frac{count(A \cup B)}{count(A)} \tag{1}$$

Where, $count(A \cup B)$ and $count(A)$ are the number of transactions containing $A \cup B$ and the number of transactions containing A, respectively.

2.2 Apriori association rules mining algorithm

This section will analyze Apriori association rules mining algorithm which is often cited as a classical one in order to improve it and introduce our application research [5-8].

Apriori association rules mining algorithm serves as a classical algorithm of discovering association rules. Its basic idea is as follows: to begin, scan the data set and then a large candidate data itemset is generated and the occurrence times of each candidate data item are calculated. Next, frequent 1-itemsets are generated based on the given minimum support, which is recorded as L_1; Then, the frequent 2-itemsets are generated by L_1; Using the similar method, until generating frequent n-itemsets, it is no longer possible to generate (n-1)-itemsets satisfying the minimum support. Finally, the rules are exported from the big data itemset.

The following is what the pseudo-code is described:

Algorithm: the frequent itemsets are found by layer by layer iteration based on candidate sets, and among them, L_k denotes frequent k-itemsets, and C_k denotes a candidate.

Input: transaction set D and support threshold.

Output: frequent itemset L.

Method:

$(k-1)$-itemsets

$\qquad C_k$ = apriori_gen(L_{k-1});   //Call function apriori_gen, and C_k is generated by L_{k-1}

$\qquad$ for $t \in D$ do

$\qquad\qquad C_k$ = subset(C_k, t);   //The candidates contained in t

$\qquad\qquad$ for $c \in C_k$ do c_count++;

$\qquad$ end;

$\qquad L_k$ = { $c \in C_k$ | c_count $\geq$ support threshold};

$\qquad$ end;

$\quad L = \text{U}_k L_k$;   // L is the output result

Fig 1: Apriori association rule mining algorithm

The apriori_gen algorithm realizes two actions of connection and pruning, and C_k is obtained from L_{k-1}. Generate possible candidates in the connection section. The Apriori property is used in the pruning section to remove candidates with infrequent subsets. Testing of infrequent subsets can be implemented by writing a procedure. Subset returns all k-dimensional data items that are included in t in the candidate set C_k.

III. Remote Sensing Task Demand Mining and Analysis Algorithm

According to the data characteristics of remote sensing task requirements, a remote sensing demand mining and analysis algorithm is designed. The algorithm is divided into two steps. Firstly, a FP-Tree is generated according to the database of remote sensing task requirements. The second step uses the remote sensing task demand mining and analysis algorithm to produce all frequent itemsets according to the FP-Tree generated in the first step. The following introduces how to establish FP -Tree tree and how to use remote sensing task demand mining analysis algorithm [9-10].

The first step is to generate FP_Tree. A formal description of the algorithm for creating a FP-Tree from a database is as follows.

Input: a transaction database D and a minimum support threshold.
Output: FP_Tree

Step 1: Scan the database D again. The set F of frequent items and the support degree of each frequent item are obtained. Put F in descending order of support, and the result is denoted as L.

Step 2: Create the root node of the FP_Tree, mark it as 'null', and do the following operation for each Transaction in D.

The frequent items in Transaction are selected and sorted according to the order in L. The Transaction out of order sign frequent items listed submission for [p|P], where p is the first element and P is the rest. Then, call the function insert_tree ([p|P], T).

Function insert_tree ([p|P], T) function as follows:

If *T* has a child of *N*, increase the count value of *N* by 1; otherwise, create a new node *N* with count 1, its parent *T*, and string its node_link with those that have the same item_name field.

FP_Tree is a compressed data structure that stores all the information needed for subsequent frequent set mining with less space.

Step 2 builds on the FP_Tree generated in the Step 1. It calls itself recursively and calls the newly generated FP_Tree repeatedly. Then, the FP-growth algorithm can be adopted to generate all frequent itemsets.

The followings are specific examples to explain how to use the mining algorithm to discover the correlation law among the parameters such as demand source, earth observation area scope, earth observation time, sensor type and so on. Assume that there are four years of cumulative user observation task requirements as shown in Table 1.

Table 1 Example of user observation task requirements

TID	User	Application Time	Observation Time	Observation Target
1	DHJD	2010-3-2	2010-3-20	DYD
2	GFDX	2010-4-2	2010-4-12	XWY
3	HJJ	2010-5-2	2010-5-17	GD
4	HJZBYJY	2010-6-2	2010-6-21	ZSB
5	NBZQ	2011-1-2	2011-1-2	HXH
6	DHJD	2011-3-3	2011-3-22	DYD
7	GFKD	2011-3-20	2011-3-31	FLB
8	PYS	2011-8-10	2011-8-21	HSW
9	DHJD	2012-3-4	2012-3-21	DYD
10	NHJD	2012-7-2	2014-7-28	HYD
11	D1JTJ	2012-9-2	2012-9-21	PTJ
12	GFDX	2012-12-2	2014-12-20	GD
13	DHJD	2013-3-2	2013-3-24	DYD
14	D27JTJ	2013-7-5	2014-7-12	WH
15	D38JTJ	2013-9-15	2013-9-20	HT
16	D39JTJ	2013-11-16	2014-11-20	XTY

That is, the item or itemset whose occurrence frequency is greater than or equal to 2 is considered to be frequent item or frequent itemset.

Step 1: Scan the above user remote sensing task demand sample table, and calculate the occurrence frequency of various user observation demand parameter values, as shown in Table 2.

Table 2 Candidate collection C_1

No.	Itemsets	Frequency	No.	Itemsets	Frequency
1	User: DHJD	4	26	Observation Time: Early January	1
2	User: GFDX	2	27	Observation Time: Late March	5
3	User: HJJ	1	28	Observation Time: Early April	1
4	User: HJZBYJY	1	29	Observation Time: Mid May	1
5	User: NBZQ	1	30	Observation Time: Late June	1

6	User: GFKJDX	1	31	Observation Time: Mid July	1
7	User: PYS	1	32	Observation Time: Late July	1
8	User: DHJD	1	33	Observation Time: Late August	1
9	User: NHJD	1	34	Observation Time: Late September	1
10	User: D1JTJ	1	35	Observation Time: Late November	1
11	User: D27JTJ	1	36	Observation Time: Late December	1
12	User: D38JTJ	1	37	Observation Target: DYD	4
13	User: D39JTJ	1	38	Observation Target: XWY	1
14	Application Time: Early January	1	39	Observation Target: GD	2
15	Application Time: Early March	4	40	Observation Target: ZSB	1
16	Application Time: Late March	1	41	Observation Target: HXH	1
17	Application Time: Early April	1	42	Observation Target: KLK	1
18	Application Time: Early May	1	43	Observation Target: HSW	1
19	Application Time: Early June	1	44	Observation Target: HYD	1
20	Application Time: Early July	1	45	Observation Target: PTJ	1
21	Application Time: Early August	1	46	Observation Target: HSW	1
22	Application Time: Early September	1	47	Observation Target: WH	1
23	Application Time: Mid September	1	48	Observation Target: HT	1
24	Application Time: Mid November	1	49	Observation Target: XTY	1
25	Application Time: Early December	1			

The frequent itemset is shown in Table 3.

Table 3 Frequent items collection L_1

No.	Itemset	Frequency
1	User: DHJD	4
2	User: GFDX	2
3	Application time: Early March	4
4	Observation time: Late March	5
5	Observation target: DYD	4
6	Observation target: GD	2

Candidate itemset C_2 is shown in Table 4.

Table 4 Candidate collection C_2

No.	Itemset	Frequency
1	User: DHJD; Application time: Early March	4
2	User: DHJD; Observation time: Late March	4
3	User: DHJD; Observation target: DYD	4
4	User: DHJD; Observation target: GD	0
5	User: GFDX; Application time: Early March	0
6	User: GFDX; Observation time: Late March	0
7	User: GFDX; Observation target: DYD	0
8	User: GFDX; Observation target: GD	1

9	Application time: Early March; Observation target: DYD	4
10	Application time: Early March; Observation target: GD	0
11	Observation time: Late March; Observation target: DYD	4
12	Observation time: Late March; Observation target: GD	0

The frequent itemset L_2 is shown in Table 5.

Table 5 Frequent items collection L_2

No.	Itemset	Frequency
1	User: DHJD; Application time: Early March	4
2	User: DHJD; Observation time: Late March	4
3	User: DHJD; Observation target: DYD	4
4	Application time: Early March; Observation target: DYD	4
5	Observation time: Late March; Observation target: DYD	4

Candidate itemset C_3 is shown in Table 6.

Table 6 Candidate collection C_3

No.	Itemset	Frequency
1	User: DHJD; Application time: Early March; Observation time: Late March	4
2	User: DHJD; Application time: Early March; Observation target: DYD	4
3	User: DHJD; Observation time: Late March; Observation target: DYD	4
4	Application time: Early March; Observation time: Late March; Observation target: DYD	4

Frequent itemset L_3 is shown in Table 7.

Table 7 Frequent items collection L_3

No.	Itemset	Frequency
1	User: DHJD; Application time: Early March; Observation time: Late March	4
2	User: DHJD; Application time: Early March; Observation target: DYD	4
3	User: DHJD; Observation time: Late March; Observation target: DYD	4
4	Application time: Early March; Observation time: Late March; Observation target: DYD	4

Candidate itemset C_4 is shown in Table 8.

Table 8 Candidate collection C_4

No.	Itemset	Frequency
1	User: DHJD; Application time: Early March; Observation time: Late March; Observation target: DYD	4

Frequent itemset L_4 is shown in Table 9.

Table 9 Frequent items collection L_4

No.	Itemset	Frequency
1	User: DHJD; Application time: Early March; Observation time: Late March; Observation target: DYD	4

Rules can be generated from frequent itemset L_4 :

"User: DHJD; Observation time: late March; Observation target: DYD "=>" Application time: early March ".

According to the above statistics and the total number of user demands in C_1, the confidence and support of this rule are as follows:

$$support = \frac{count(\text{DHJD; Early March; Late March; DYD})}{count(\text{User demand})} = \frac{4}{16} = 0.25 \qquad (2)$$

$$support = \frac{count(\text{DHJD; Early March; Late March; DYD})}{count(\text{DHJD; Late March; DYD})} = \frac{4}{4} = 1 \qquad (3)$$

Here, set the threshold of support as 0.2 and the threshold of confidence as 0.2. Therefore, this rule is a strong association rule.

Based on this rule, an observation application for DYD observation target can be automatically generated for a remote sensing user before the arrival of March every year, and submitted to the remote sensing satellite data exchange center, in order to obtain a greater opportunity to arrange observation resources.

IV. Conclusion

This paper mainly studies and designs a remote sensing user demand mining analysis method. Users on the accumulation of remote sensing data mining on the remote sensing image data requirements and user needs analysis, found that source of demand, the demand of earth observation area, earth observation time, the link between the sensor type parameters, such as law, on the basis of these knowledge rule in advance can be automatically generated user requirements, implement automatic intelligent user needs to submit, Ensure that the regular requirements of users can be submitted to the satellite cooperative application control department as soon as possible. The research and application of remote sensing demand mining analysis method is a relatively new research topic. This paper only preliminarily discusses the technical process of remote sensing data mining and knowledge discovery, but the realization of system prototype, there are some key technologies need to be further discussed.

Reference

[1] M.Y. He, et al., "Research progress in feature mining of hyperspectral remote sensing data," Space Return and Remote Sensing, vol. 02, pp. 1-10, 2013.

[2] G.J. He, "Research on intelligent processing method of satellite remote sensing information based on data mining mechanism," Science Technology and Engineering 12, 2005.

[3] Z.G. Zheng et al., "Remote sensing data users need fusion processing technology," Journal of National University of Defense Technology, vol. 25, pp. 9, 1-3, 2018.

[4] M. Leng, "Analysis method and application of satellite earth observation requirements," National University of Defense Technology, vol. 03, 2015.

[5] Q. Li, "Research on association analysis algorithm in data mining," Harbin Institute of Technology, vol. 01, 2010.

[6]   Y.B. Li, et al., "Frequent pattern mining algorithm based on improved FP Tree," Computer Application, vol. 01, 2011.

[7]   Y. Cui, et al., An overview of association rule mining computer application research, vol. 32, no. 02, pp. 2, 2016.

[8]   K. Laib, A. Korniienko, M. Dinh, G. Scorletti, F. Morel, "Hierarchical robust performance analysis of uncertain large scale systems," IEEE Trans. Autom Control, vol. 63, pp. 2075–2090, doi: 10.1109/TAC. 2017.2762468, 2018.

[9]   Y. Zhang, Y. Yuan, Y. Feng, X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," IEEE Trans. Geosci. Remote Sens, pp. 1–14, doi:10.1109/TGRS.2019. 2900302, 2019.

[10]   R.J. L, J. Z, P.J. S, F.J. S, G.F. L, "An agglomerative hierarchical clustering based high-resolution remote sensing image segmentation algorithm," In proceedings of 2008 International Conference on Computer Science and Software Engineering, Wuhan, China, 12–14 December 2008, pp. 403–406, 2008.