Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 5, pp. 17218-17222, May, 2017

ISSN: 0976-3031

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

CRITICAL STUDY AND ANALYSIS FOR DECIDING SENSITIVE AND NON-SENSITIVE ATTRIBUTES OF MEDICAL HEALTHCARE DATASET THROUGH SURVEY AND USING ASSOCIATION RULE MINING

Devendra I Vashi[1]., H B Bhadka[2] and Kuntal P Patel[3]

[1]CE Department Nirma University Ahmedabad-382481 Gujarat, India
[2]Computer Science Department C U Shah University Wadhwan-363030 Gujarat, India
[3]School of Computer Studies Ahmedabad University Ahmedabad-380009 Gujarat, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0805.0307

ARTICLE INFO

Article History:

Received 18th February, 2017
Received in revised form 10th March, 2017
Accepted 06th April, 2017
Published online 28th May, 2017

Key Words:

Data mining, association rule, Apriory algorithm, Healthcare, support, confidence, antecedent, consequent, item set, PPDM, weka.

ABSTRACT

Association rule mining technique is useful for extracting useful relation from the given dataset. In today's scenario data mining technique is performed on medical dataset to find out a pattern on relative attributes. In this this paper Apriory algorithm is used of association rule mining technique to find out patterns. List of probable medical attributes has been finalized through survey for creating dataset. Dataset created mainly for the age group of 20 to 45 years. Total 26 attributes finalized and 131 input was taken through google form for crating database to decide highly sensitive, average sensitive, low sensitive and not sensitive attribute. This approach was studied using weka [8] tool.

INTRODUCTION

In privacy preserving data mining before providing database for data mining specific attributes should be secure enough during PPDM. Means there are some medical data or personal data which should not bed is closed to third party during data mining. For the same apriori algorithm of association rule can be useful to decide sensitive data by generating frequent item set rule. Means which attribute may be sensitive with occurrence of the other attribute.

Proposed Approach to Decide Sensitive Attribute

As per literature review and guidance taken from the researchers who are working in the field of privacy preserving data mining, only the individual who can decide the sensitive attribute of his/her medical or personal data. That is why for categorization of different kinds of sensitive attribute, online survey has been taken through google form. For the same different kinds of healthcare data has been studied and all possible common health care attributes and personal attribute which may be associated with it have been finalized for deciding the various sensitive attributes.

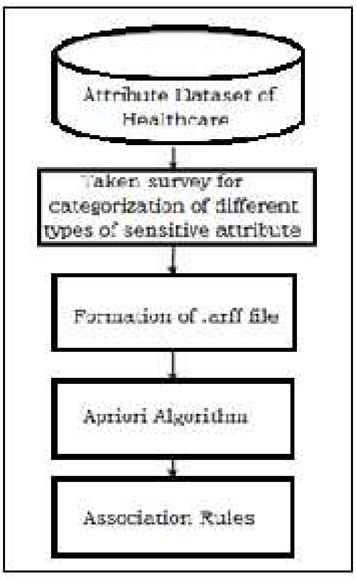

Attribute Dataset of Healthcare
Taken survey for categorization of different types of sensitive attribute
Formation of .arff file
Apriori Algorithm
Association Rules

Figure 1 Flowchart of the Proposed Approach

**Corresponding author: Devendra I Vashi*
CE Department Nirma University Ahmedabad-382481 Gujarat, India

These sensitive attribute will be providing for application of privacy preserving data mining through cryptography technique. In the below proposed approach [9] the analysis of created dataset has been done using apriorI algorithm of association rule mining using weka tool.

Types of Sensitive Data

1. ***Highly sensitive data***: In medical health care database, there are some attributes which are straightaway sensitive for patient if it reveals to society. Like if someone will come to know Email-id, mobile no, salary and PAN card of an individual then it may happen that their bank or social media account may hacked by unauthorized person.
2. ***Average sensitive data***: Based on survey taken, there are some attribute like age, Date of birth, blood group, marital status and dieses name are may not be associated directly to reveal personal information of an individual but it by combining these attributes patient may be identifiable. Such kinds of attributes are to be considered as average sensitive.
3. ***Low sensitive data***: Even by combining some quasi attributes these attributes may be located but still it may not me the problem for individual.
4. ***Not sensitive data***: Even these attributes will reveal to third party it will not create any problem for individual like name, gender, education detail and nationality.

Association Rule Mining

Definition 1: Let item set I = {i1, i2, i3 …in} be a set of items. D is a database of transactions. Each transaction T is usually given a transaction id called TID. Each transaction T is a subset of I. [1][3] [22]

Definition 2: Association rules are kind of if then else statement which helps to find out relationship between unrelated data [17]. The major objective of association rule mining is to generate rule with minimum support [7][26] and confidence. The only problem of Association rule mining is it works nicely with categorized data not with numerical data.

Association rule has two parts: antecedent and consequent.

Example: prescriptions=Highly Sensitive and medical or blood Reports=Highly Sensitive 41 ==> Laboratory results=Highly Sensitive 41, meaning of this rule is if the person believes that prescriptions and medical or blood Reports are highly sensitive than any Laboratory results is also highly sensitive form him and 41 instances of LHA and 41 instances of RHS has been found in the dataset.

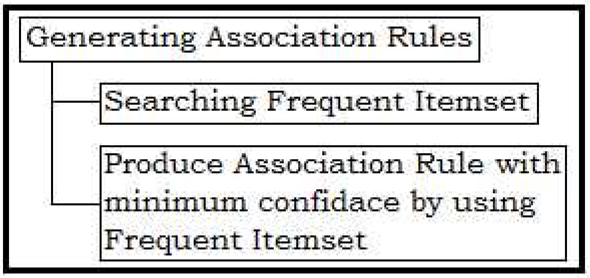

Generating Association Rules
Searching Frequent Itemset
Produce Association Rule with minimum confidace by using Frequent Itemset

Figure 2 Association Rule Generation[15]

Association rules are basically to fulfill the minimum support and minimum confidence which defines by the user. Association rule is used to explore relations between various fields or variables in data mining. To compute correlation between them there are two key terms [13][19][20].

1. Minimum Support
2. Minimum Confidence

Minimum Support

Support means how frequently the combinations of item (item set) noticeable in Transaction or Database.

Minimum Confidence

Confidence is a relative or dependent support of the item.

Example

Association Rule: X ==>Y means if customer buy then he will also buy item y. and we can compute the Support = (Frequency of transaction of X & Y together) / N
Where N is the total no of transactions.
We can also compute the
Minimum Confidence = (Frequency of transaction of X & Y together) / Frequency of transaction of X)

Classification of Association Rule [25]

1. Single Dimensional association rule: In single dimension association rule have one dimension or predicator that is items in the rule refers to only one dimension or predicate.
   Example: Pencil ==> Eraser
   Dimension is *shopping*
2. Multidimensional association rule: as the name indicates it has two or more dimension. In multidimensional association rule the dimension and predicates should not be the repeated.
   Example: prescriptions=Highly Sensitive diagnosis codes=Highly Sensitive 38 ==> Laboratory results=Highly Sensitive 38
3. Hybrid Dimensional association rule: In dimensional association rule the dimension and predicates can be the repeated.
   Example: time (4 o'clock), buy (pencil) ==>buy (eraser)

Why association rule mining technique is used for medical health care dataset?

Association Rule Mining Algorithm	Advantages	Disadvantages
Apriori	1. Requires less memory. 2. Easy to implement. 3. Because of pruning concept very less item sets left to check for further support checking. 4. Work with distributed database	1. Scanning of database is more. 2. Uses only one minimum support threshold. 3. This algorithm is good for small database.

Apriori Algorithm: This algorithm works only with categorical dataso if any dataset contains numerical values then we must convert it in to nominal values. Apriori algorithm calculate all possible rule which satisfy with minimum support and be better than confidence.

Steps to implement apriori algorithm

- This algorithm is used to mine frequent [23] item set [6] from transactional database for a Boolean association rules.
- This algorithm is the based on the idea that "a subset of a frequent item set must also be a frequent item set[24]. Example: if {I1, I2} is a frequent item set, then {I1} and {I2} also should be a frequent item set". It is an iterative approach to find a frequent item sets. And it uses the frequent item set to generate the association rules.
- The algorithm uses the level wise search, where *k-item sets* are used to explore *(k+1)-item sets*.
- Frequent subsets are extended one item at a time. This step is known as *candidate generation* step. Then groups of candidates are tested against the data. It identifies the frequent individual items in the database and extend them to a larger item sets as long as those item sets appear sufficiently often in the database. [27] [28]
- This algorithm determines frequent item sets that can be used to determine association rules which highlight general trends in the database.
- This algorithm takes an advantage of the fact that any subset of frequent item set is also a frequent item set. The algorithm can therefore, reduce the number of candidates being considered by only exploring the item sets whose support count is greater than the minimum support count. All infrequent item sets can be *pruned* if it has an infrequent subset

How to do prune?

Prune [21] those item sets candidates that have no hope to be large, consequently the unnecessary effort of counting those item sets can be avoided.

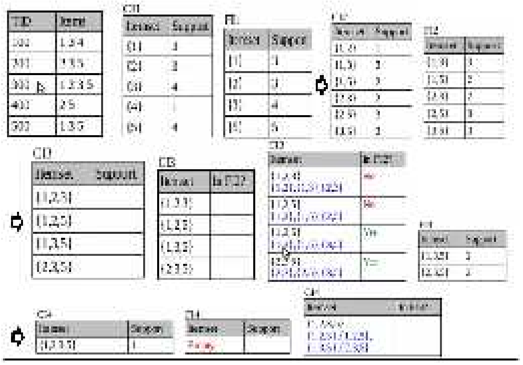


Figure 3 Example of Apriori algorithm [2] [4] [12]

Dataset Used In Weka

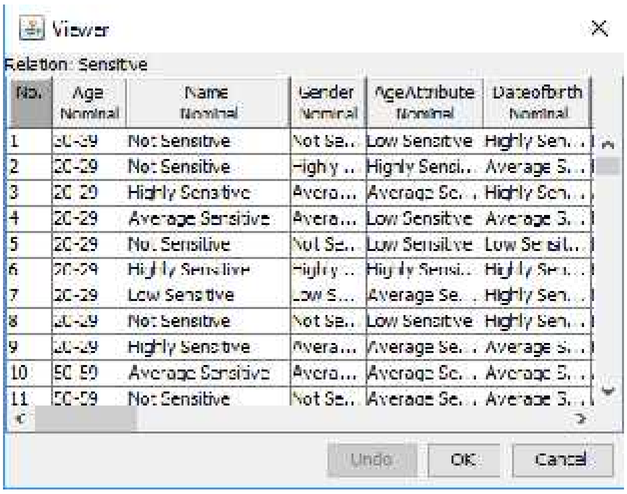
Table 1 Dataset Description

Dataset	No of Attributes	No of Instances
Common List of Attributes of Healthcare	26	131

Table 2 List of Attributes in Dataset][14]

Age	Nationality	Salary	Laboratoryresults
Name	EducationDetails	Photos	Reports
Gender	MaritalStatus	PancardNumber	Diagnosiscodes
Age	AddressHome	Disease	DoctorName
Dateofbirth	AddressWork place	Medicationallergies	Date&Timeofvisit
BloodGroup	MobileNo	Treatments	
DisabilityDetails	Emailid	prescriptions	

Sample Dataset



Relation: Sensitve

No.	Age Nominal	Name Nominal	Gender Nominal	AgeAttribute Nominal	Dateofbirth Nominal
1	30-39	Not Sensitive	Not Se...	Low Sensitve	Highly Sen...
2	20-29	Not Sensitive	Highly ...	Highly Sensi...	Average S...
3	20-29	Highly Sensitive	Avera...	Average Se...	Highly Sen...
4	20-29	Average Sensitive	Avera...	Low Sensitve	Average S...
5	20-29	Not Sensitive	Not Se...	Low Sensitve	Low Sensit...
6	20-29	Highly Sensitive	Highly ...	Highly Sensi...	Highly Sen...
7	20-29	Low Sensitive	Low S...	Average Se...	Highly Sen...
8	30-39	Not Sensitive	Not Se...	Low Sensitve	Highly Sen...
9	30-39	Highly Sensitive	Avera...	Average Se...	Average S...
10	50-59	Average Sensitive	Avera...	Average Se...	Average S...
11	50-59	Not Sensitive	Not Se...	Average Se...	Average S...

Figure 3 Classification of Dataset [18]

Summary and Analysis of Result Obtain

In this approach following default settings for Apriori algorithm was set:

With a minimum support 0.1% to generate 10 rules or the support falls below 10%. And minimum confidence is 90%.

Minimum support: 0.1 (13 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 2

Generated sets of large item sets:

Size of set of large item setsL (1): 87
Size of set of large item setsL (2): 1135
Size of set of large item setsL (3): 3769
Size of set of large item setsL (4): 6502
Size of set of large item setsL (5): 8928
Size of set of large item setsL (6): 9980
Size of set of large item setsL (7): 8334
Size of set of large item setsL (8): 4835
Size of set of large item setsL (9): 1804
Size of set of large item setsL (10): 397
Size of set of large item setsL (11): 50
Size of set of large item setsL (12): 4

RESULT ANALYSIS

In this study 10 best association rules were generated using apriori algorithm for the 100% confidence.

Best rules found [11]:

Some Exception Found During Experiments

While implementing of apriori algorithm on the attribute dataset following exceptions were found [10]:

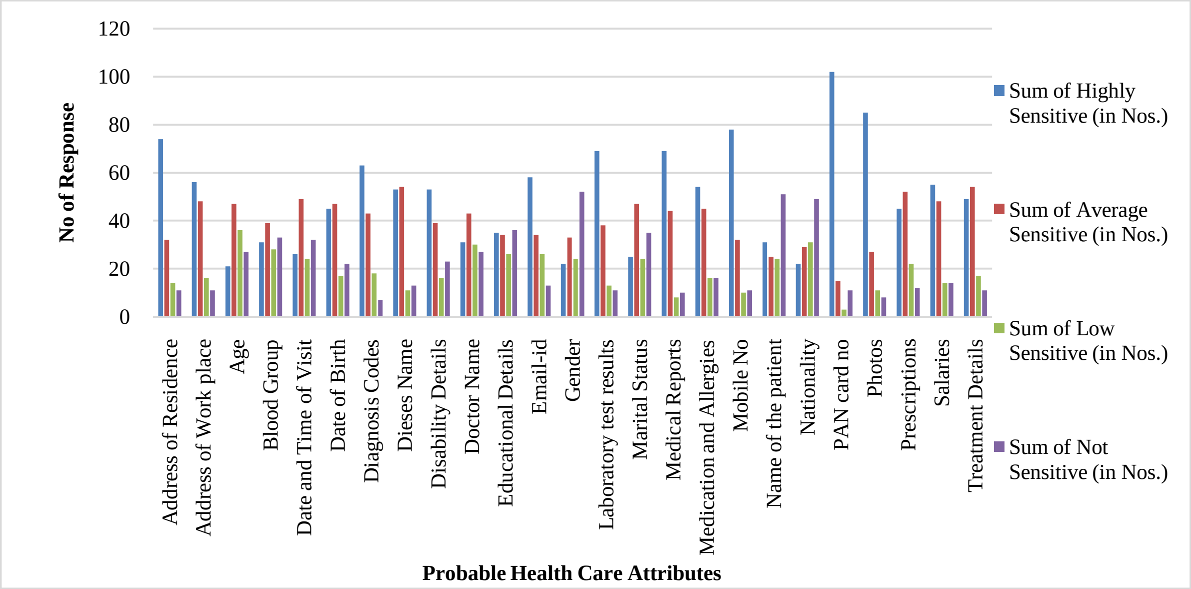
1. Sometimes not generating ***useful*** or interesting rules
2. Sometimes number of rules generated ***more***

Table 4 Best Rules found for sensitive attribute dataset

Rule No.	Defined rule as LHS ==> RHS [16]	Confidence
1	prescriptions=Highly Sensitive Reports=Highly Sensitive 41 ==>Laboratory results=Highly Sensitive 41	conf:(1)
2	prescriptions=Highly Sensitive diagnosis codes=Highly Sensitive 38 ==>Laboratory results=Highly Sensitive 38	conf:(1)
3	Age=20-29 prescriptions=Highly Sensitive Reports=Highly Sensitive 36 ==>Laboratory results=Highly Sensitive 36	conf:(1)
4	prescriptions=Highly Sensitive Reports=Highly Sensitive diagnosis codes=Highly Sensitive 36 ==>Laboratory results=Highly Sensitive 36	conf:(1)
5	Treatments=Highly Sensitive prescriptions=Highly Sensitive 35 ==>Laboratory results=Highly Sensitive 35	conf:(1)
6	PancardNumber=Highly Sensitive prescriptions=Highly Sensitive Reports=Highly Sensitive 35 ==>Laboratory results=Highly Sensitive 35	conf:(1)
7	PancardNumber=Highly Sensitive Treatments=Highly Sensitive Reports=Highly Sensitive 34 ==>Laboratory results=Highly Sensitive 34	conf:(1)
8	Photos=Highly Sensitive prescriptions=Highly Sensitive 33 ==> PancardNumber=Highly Sensitive 33	conf:(1)
9	Photos=Highly Sensitive prescriptions=Highly Sensitive 33 ==>Laboratory results=Highly Sensitive 33	conf:(1)
10	Age=20-29 prescriptions=Highly Sensitive diagnosis codes=Highly Sensitive 33 ==>Laboratory results=Highly Sensitive 33	conf:(1)

Table 5 Summary of no of response from the age group: 20 years to 60 years

Name of Attribute	Highly Sensitive (in Nos.)	Average Sensitive (in Nos.)	Low Sensitive (in Nos.)	Not Sensitive (in Nos.)	Conclusion
Name of the patient	31	25	24	51	Not Sensitive
Gender	22	33	24	52	Not Sensitive
Age	21	47	36	27	Average Sensitive
Date of Birth	45	47	17	22	Average Sensitive
Blood Group	31	39	28	33	Average Sensitive
Disability Details	53	39	16	23	Highly Sensitive
Nationality	22	29	31	49	Not Sensitive
Educational Details	35	34	26	36	Not Sensitive
Marital Status	25	47	24	35	Average Sensitive
Address of Residence	74	32	14	11	Highly Sensitive
Address of Work place	56	48	16	11	Highly Sensitive
Mobile No	78	32	10	11	Highly Sensitive
Email-id	58	34	26	13	Highly Sensitive
Salaries	55	48	14	14	Highly Sensitive
Photos	85	27	11	8	Highly Sensitive
PAN card no	102	15	3	11	Highly Sensitive
Dieses Name	53	54	11	13	Average Sensitive
Medication and Allergies	54	45	16	16	Highly Sensitive
Treatment Details	49	54	17	11	Average Sensitive
Prescriptions	45	52	22	12	Average Sensitive
Laboratory test results	69	38	13	11	Highly Sensitive
Medical Reports	69	44	8	10	Highly Sensitive
Diagnosis Codes	63	43	18	7	Highly Sensitive
Doctor Name	31	43	30	27	Average Sensitive
Date and Time of Visit	26	49	24	32	Average Sensitive



CONCLUSINS AND FUTURE WORK

Association rule is widely used in finding the frequent item set in the medical healthcare data. In this research article, by taking the online survey all four kinds of attributes of medical heath care were categorized. For the same to support the analysis of categorical attribute, apriori algorithm of association rule mining is used to generate the rule to identify the sensitive attribute. This result may vary depends on the age group. Means sensitivity [5] can be decide only based on age group only as age increase definition of sensitive attribute may vary for different age group people.

This result analysis will be used further in privacy preserving data mining in healthcare data using cryptographic approach.

References

1. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. "Data mining: an overview from a database perspective." *IEEE Transactions on Knowledge and data Engineering* 8.6 (1996): 866-883.
2. Park, Jong Soo, Ming-Syan Chen, and Philip S. Yu. *An effective hash-based algorithm for mining association rules*. Vol. 24. No. 2. ACM, 1995.
3. Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." *Acmsigmod record*. Vol. 22. No. 2. ACM, 1993.
4. Agrawal, Rakesh, *et al*. "Fast Discovery of Association Rules." *Advances in knowledge discovery and data mining* 12.1 (1996): 307-328.
5. Agrawal, Rakesh, and John C. Shafer. "Parallel Mining of Association Rules: Design." *Implementation, and Experience* (1996).
6. Agarwal, Ramesh C., Charu C. Aggarwal, and V. V. V. Prasad. "A tree projection algorithm for generation of frequent item sets." *Journal of parallel and Distributed Computing* 61.3 (2001): 350-371.
7. Ma, Bing Liu Wynne Hsu Yiming, and Bing Liu. "Integrating classification and association rule mining." *Proceedings of the 4th*. 1998.
8. Parvathi, I., and Siddharth Rautaray. "Survey on data mining techniques for the diagnosis of diseases in medical domain." *International Journal of Computer Science and Information Technologies* 5.1 (2014): 838-846.
9. Tuba, P. A. L. A., İbrahim YÜCEDAĞ, and Hasan Biberoğlu. "Association rule for Classification of Breast Cancer Patients." *Sigma* 8.2 (2017): 155-160.
10. Ordonez, Carlos, Cesar A. Santana, and Levien De Braal. "Discovering Interesting Association Rules in Medical Data." *ACM SIGMOD workshop on research issues in data mining and knowledge discovery*. 2000.
11. Doddi, AchlaMarathe, SS Ravi, David C. Torney, Srinivas. "Discovery of association rules in medical data." *Medical informatics and the Internet in medicine* 26.1 (2001): 25-33.
12. Ullah, Irshad. "Data mining algorithms and medical sciences." *International Journal of Computer Science & Information Technology (IJCSIT)* 2.6 (2010).
13. Rashid, Mahmood A., MdTamjidulHoque, and Abdul Sattar. "Association rules mining based clinical observations." *arXiv preprint arXiv: 1401.2571* (2014).
14. Hu, Ruijuan. "Medical data mining based on association rules." *Computer and Information Science* 3.4 (2010): 104.
15. Jain, Divya, and SumanlataGautam. "Implementation of Apriori Algorithm in Health Care Sector: A Survey." *International Journal of Computer Science and Communication Engineering* 2.4 (2013).
16. Rane, Nandita, and Madhuri Rao. "Association Rule Mining on Type 2 Diabetes using FP-growth association rule." *International Journal of Engineering and Computer Science* 2.8 (2013): 2481-85.
17. Devi, M. Renuka. "Applications of Association Rule Mining in Different Databases." *Journal of Global Research in Computer Science* 3.8 (2012): 30-34.
18. Nofal, A. A. D. M., and SuliemanBani-Ahmad. "Classification based on association-rule mining techniques: a general survey and empirical comparative evaluation." *Ubiquitous Computing and Communication (UBICC) Journal* 5.3 (2010).
19. Kaur, Jagmeet, and NeenaMadan. "Association Rule Mining: A Survey." *International Journal of Hybrid Information Technology* 8.7 (2015): 239-242.
20. Pazhanikumar, K., and S. Arumugaperumal. "Association rule mining and medical application: a detailed survey." *International Journal of Computer Applications* 80.17 (2013).
21. Slimani, Thabet. "New Approach to Optimize the Time of Association Rules Extraction." *arXiv preprint arXiv: 1312.4800* (2013).
22. Pawar, Priyanka, Sachin Deshpande, and VipulDalal. "Association Rule Mining with Hybrid-Dimension Datasets." *International Journal of Computer Science and Information Security* 10.2 (2012): 81.
23. Rob Sullivan. "Representing Data Mining Results", Introduction to Data Mining for the Life Sciences, 2012
24. Rout, Trilochan, MamataGaranayak, Manas Ranjan Senapati, and Sushanta Kumar Kamilla. "Big data and its applications: A review", EESCO, 2015.
25. Fang, Luo, and QiuQizhi. "The Study on the Application of Data Mining Based on Association Rules", 2012 International Conference on Communication Systems and Network Technologies, 2012.
26. Jingkai Zhou. "The application of association rule mining to remotely sensed data", Proceedings of the 2000 ACM symposium on Applied computing - SAC 00 SAC 00, 2000
27. Trupti A. Kumbhare, Prof. Santosh V. Chobe, "An Overview of Association Rule Mining Algorithms", *International Journal of Computer Science and Information Technologies*, Vol. 5 (1) , 2014
28. Zhang, Yiyun, and Guolong Chen. "A forensics method of web browsing behavior based on association rule mining." *Systems and Informatics (ICSAI), 2014 2nd International Conference on*. IEEE, 2014.

****