ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20210903062509
Original scientific paper

Internet Medical Privacy Disclosure Mining and Prediction Model Construction Based on Association Rules

Yong WANG

Abstract: In recent years, China's Internet medical industry has developed rapidly and the market scale has been expanding. Medical privacy is an important research point in the Internet medical field. If the patient cannot fully communicate with the doctor on the other end of the Internet, then it is obvious that the patient will not be well treated. Then it becomes very worthwhile to mine the factors affecting patients' privacy disclosure and predict patients' disclosure behavior. This paper uses the classical and improved multidimensional Apriori (MD-Apriori) to mine patient privacy disclosure factors, which proves that the improved MD-Apriori algorithm is more applicable in this study. In order to prove the validity and authority of the mining results, this paper uses SPSS to analyze 331 valid questionnaires. The results show that the privacy disclosure factors obtained by the two methods are almost the same. Finally, based on the above factors, we establish the Internet medical privacy disclosure intention prediction model, in order to guide the construction and improvement of internet medical.

Keywords: internet medical; internet medical privacy disclosure intention prediction model; MD-Apriori (multi-dimensional Apriori); questionnaire; SPSS

1 INTRODUCTION

The mismatch of demand and resource in China's medical and health industry is prominent, the aging of the population, the smaller family size, the normalization of sub-health and other phenomena have led to a surge in demand for medical help and health. In China, there is a shortage of practicing doctors per 10 000 people, a lack of general practitioners, a shortage of doctor resources per capita, and uneven distribution of medical resources. These factors lead to great problems in the allocation of medical resources. In order to solve the above problems, Internet Medical came into being. With its advantages in hierarchical diagnosis and treatment, medical informatization, health management, medical separation, and commercial insurance, it has become the main way to solve this problem [1].

Formed by the deep integration of the Internet as a carrier and information technology as a means with traditional medical and health services, Internet Medical is a new type of medical and health service model. It represents the new development direction of the medical industry and is conducive to solving the contradiction between the imbalance of medical resources and people's increasing demand for health care [2]. In recent years, China's Internet medical industry has developed rapidly and the market scale has continued to expand. The promulgation of relevant policies has further regulated and clarified the development direction of Internet Medical. During the COVID-19 epidemic, the online diagnosis and treatment, health consultation and other epidemic-related services provided by Internet medical services have effectively helped fight the epidemic. At the same time, it has also cultivated the habit of online consultation for the broad number of patients.

Internet medical websites represented by Haodf (https://www.haodf.com/) and Chunyu Doctor (https://www.chunyuyisheng.com/) rely on users' comprehensive, accurate, and standardized personal health data to provide services to users and ensure smooth communication with doctors. However, the development time of Internet Medicine is short, and there are still many problems; among them, personal health information is very important to patients' lives. Doctors can make correct suggestions based on relevant information to improve users' health status, but wrong diagnosis will put patients at risk [3]. In addition, patients' awareness of the protection of their own health information is constantly strengthened, and medical data is very important. Once tampered, damaged and leaked, it will pose a serious threat to the privacy and health safety of both doctors and patients. During the COVID-19 epidemic, China's health and medical networks, vaccine research institutions, etc. have frequently encountered cyber attack, some health sensitive data are illegally leaving the country. For example, a famous hospital in China has reached a cooperation agreement with a foreign company to illegally initiate scientific research on some sensitive data. Worldwide, the number of attacks on medical and health networks has increased by 5 times year-on-year. This shows that the improvement and development of Internet Medical are still facing huge challenges [4].

Factors affecting users' intention to disclose personal health information include but are not limited to trust in medical websites, privacy and security, information quality, material rewards, etc. For example, when patients are interrogated on an internet medical platform, when referring to age, weight and height, the general patients will disclose them truthfully, and often avoid similar problems like "AIDS history" and "drug abuse", then if the medical website has a good reputation and strong privacy and confidentiality, the possibility of disclosing privacy will increase. Moreover, some platforms have set up material rewards on some privacy issues, which will also encourage patients to disclose privacy. The starting point of this paper is to explore factors affecting patients' intention to disclose medical privacy, and establish a prediction model of intention to disclose medical privacy on this basis. This research is of great significance.

Due to the lack of basic medical data, all hospital databases have not been fully opened. Only 3% of hospitals have realized database interoperability. Using the Internet to help hospitals realize data interoperability can effectively solve the bottleneck faced by online medical websites running on user health database in health data collection and integration. Mining the influencing factors

of personal medical privacy disclosure intention, exploring the relevance of these factors, and making suggestions for the optimization of online medical website, is significant to the development of Internet Medical.

Before this paper, it has many contributions in mining the influencing factors of information and privacy disclosure in many fields。Zang et al. [5] introduced the persuasive knowledge theory on the basis of privacy computing theory to mine the impact of their own knowledge on personal information disclosure intention. Hua et al. [6] used the S-O-R model and the Perceived Risk-benefit theory to explore the process of information disclosure behavior. Li et al. [7] explored the comprehensive influence of gender-based positioning, individualization, and social benefits on LBS information disclosure intention. Fan et al. [8] constructed the influencing factor model of personal information disclosure intention of social media based on privacy calculus and planned behavior theory. Cheng et al. [9] constructed a social media user privacy disclosure intention model from the dual perspectives of individual rational and perceptual factors based on privacy computing theory, communication privacy management theory, etc.

Compared with the existing work, the contributions of this paper can be summarized as follows:

(1) Using the classical Apriori and the improved MD-Apriori algorithm, we have mined the relationship between many factors that affect the user's intention to disclose privacy.

(2) Using questionnaire and SPSS to analyze the factors affecting user privacy disclosure intention, and compare them with the data mining results.

(3) Mining factors that affect users' privacy disclosure intention, and constructing an Internet medical privacy disclosure intention prediction model based on Bayesian network

The following parts of this paper are organized as follows: Chapter 2 lists the results of previous studies on the influencing factors of privacy disclosure intention. Chapter 3 shows the processes of data acquisition, data processing, and selection of risk factors. Chapter 4 introduces the methods used in this paper, and Chapter 5 introduces the experimental process and conclusions, Chapter 6 summarizes the work of this paper.

2 RELATED WORK

Before this paper, the research on the influence of users' personal information disclosure intention has some achievements, which can be roughly divided into several categories: (1) Works on the analysis of influencing factors of privacy disclosure (2) Works on the use of statistical methods for the study of information disclosure, (3) Works based on data mining techniques to mine association rules (4) Works to build predictive models based on influencing factors.

Some works analyse the influencing factors of privacy disclosure. For example, Li et al. [19] apply the general linear modelling approach to blogging data converted with a coding scheme, and intended to investigate the effect of users' demographics, social network site experience, personal social network size, and blogging productivity on privacy disclosure behaviors by analyzing the data collected from social network sites. Yu et al. [20] analyse how perceived privacy risks and privacy concerns affect the disclosure intention and the actual information disclosure behavior of Internet users. By applying meta-analyses and SEM on 104 independent studies with 42 256 samples from existing empirical studies, this paper attempts to systematically reveal the relationship between privacy cognition and information disclosure.

Some works are based on privacy disclosure using statistical methods, for example, Xia [10] carried out structural equation model test facing the mobile commerce environment on the collected sample data using questionnaire and SPSS; the research explored the key technical characteristics and potential interaction of privacy feedback from two perspectives of innovation diffusion and signal transmission theory; it also revealed the impact mechanism of these technical characteristics on users' personal information disclosure intention, and constructed the impact mechanism model of privacy feedback technical characteristics on Mobile Commerce Users' personal information disclosure intention from the perspective of innovation diffusion theory and taking users' deep psychological state and psychological comfort as the intermediary. Xie et al. [11] constructed the influencing factor model of users' intention to disclose privacy under the condition of command library from the perspective of users, using privacy computing theory, privacy concern theory, communication privacy theory and behavior planning theory, and verified the model hypothesis through questionnaire. Wang [12] constructed the influencing factor model of users' intention to disclose health information in virtual health community based on privacy computing theory, social exchange theory and trust theory. Taking users who have used medical and health websites as the research object, he collected 264 effective samples through questionnaire, and then conducted an empirical study on the influencing factor model by using SPSS 20.0 and AMOS 21.0.

Some works use data mining technology to mine association rules, for example, Chen et al. [14] applied association rule in tourism study to mine association rules of different tourists focusing on different variables in scenic spots to provide suggestions for the improvement of scenic spots. We observed that association rule was more applied to the medical field. For example, Fang [15] and Qi [16] used association rule to mine the medication laws for some diseases to promote the development of modern medicine.

Some works build prediction model based on influencing factors, for example, Wang et al. [17] proposed a short-term wind speed prediction method of particle swarm optimization limit learning machine based on principal component attribute reduction clustering. The PCA is used to calculate the eigenvalues of each component, the K-means is used to cluster the wind speed samples, and then the particle swarm optimization algorithm is used to optimize the limit learning machine to construct the wind speed combination prediction model. Ye [18] builds the traffic accident risk prediction model to predict the occurrence of traffic accidents based on the obtained accident risk factors and Bayesian network.

It can be observed that the above research methods mainly have one or more of the following problems:

(1) Ignoring the relationship between many factors that affect users' intention to disclose information, for example, factors A and B have a positive impact on the intention to disclose, and factor C has a negative impact on the intention to disclose, but the comprehensive impact of A, B and C has not been excavated.

(2) Using a single questionnaire method, although the questionnaire has a strong randomness, it is restricted by objective conditions such as the lack of valid questionnaires and the difficulty of evaluating the quality of questionnaires, which will lead to insufficient accuracy of the results.

(3) Mining the factors that affect users' intention to disclose privacy, but not applying them further.

This paper comprehensively uses social science statistical analysis, data mining and deep learning techniques and is based on the data of the current popular internet medical platform, uses the classical Apriori and the improved MD-Apriori algorithm to mine the association rules of these factors. and uses questionnaire and SPSS to verify its reliability; finally, the Internet medical privacy disclosure intention prediction model is constructed in order to be beneficial to the optimization of Internet Medical.

3 DATA PROCESSING

3.1 Influencing Factors

Referring to previous studies, users often pay more attention to the strength and authority of medical websites, as well as the effectiveness of doctors' suggestions, and users will also have the impulse to disclose when facing reward stimuli. In addition, we also add psychological factors such as Perceived Risk, demographic characteristics such as age and gender. Therefore, the factors affecting privacy disclosure in this paper include information quality, personalized service, trust in medical website, trust in doctors, Perceived Risk, material reward and other variables. Tab. 1 introduces the specific meaning of each factor.

Table 1 Influencing factors and specific meaning

Risk factors	Specific meaning
information quality	Users' positive expectations for the high quality of doctors' suggestions
personalized service	If the medical website provides different services for different individual situations
trust in medical websites	Users' positive expectations for the authority and reliability of medical websites
trust in doctors	Positive expectations for doctors' ability, moral quality and behavior
Perceived Risk	Users cannot predict whether their decision is correct or worry about serious consequences
material rewards	Does the medical website give material rewards
other variables	Age and gender

3.2 Data Acquisition

The experimental data of this paper comes from the doctor-patient communication content shown by Chunyu Doctor (https://www.chunyuyisheng.com/), Haodf (https://www.haodf.com/), uses web crawler to obtain, and combines the policies of the platform, as well as the user's evaluation of platform services, information quality, drug purchase, etc. As shown in Tab. 2, in this data, the patient is a male, 20 years old, who does not trust the website and doctor, non-members have no personalized service and have a strong Perceived Risk, the platform does not set material rewards, and the final result is no privacy information disclosure.

Table 2 One of the acquired data

Patient: Hello, doctor. How long can paint dermatitis get better (male, 20 years old)
Dr. Diao XX: Hello, friend. How long have you been? Is there anything uncomfortable
Patient: on the 16th
Dr. Diao XX: did you have an operation on the 16th?
Patient: I haven't done it. I just touch lacquer tree when cutting firewood, and I have acne all over my body
Dr. Diao XX: consider it a skin disease
Patient: (the patient's speech is fierce and should not be displayed)
The service was poorly evaluated and there was no drug purchase
Privacy disclosure: no photos, no detailed condition information

When patients use internet medical services, they will not directly give some clear information, such as "I trust this doctor" or "I think the suggestion is very useful", etc. Therefore, we select the following characteristics as the measurement of the above indicators, as shown in Tab. 3.

Table 3 Risk factors and their measurement indicators

Risk factors	measurement indicators
information quality	Whether to adopt doctor's suggestions and user feedback
personalized service	It depends on the specific policy of the medical website
trust in medical websites	It depends on the user evaluation
trust in doctors	Whether there is drug purchase behavior, user evaluation and consultation time
Perceived Risk	Whether there is drug purchase behavior, and keywords such as "effective"
material rewards	It depends on the specific policy of the medical website
other variables	Whether the user is older than 46, male or female

If the patient purchases drugs according to the doctor's suggestions, it indicates that the patient trusts the doctor. Among them, according to the previous studies [13], people over the age of 46 are more reluctant to disclose personal information, and men are more likely to allow third-party institutions to use desensitized health information for research purposes. Therefore, these factors are also taken into account in this paper, some websites such as Chunyu Doctor can enjoy personalized services by opening a membership, and other measurement indicators are listed in detail.

3.3 Final Data

In this paper, we use keyword extraction technology, combined with the above risk factor measurement standards, and filter out the data with missing values. We get 1256 data. See Tab. 4 for some of them, where "yes" represents the presence of the factor and "no" represents the absence of the factor.

4 METHOLOGY

4.1 Data Mining

Data mining refers to the process of searching the information hidden in the data through algorithms.

Through data mining technology, data value-added can be realized and the hidden knowledge of data can be mined. Data mining technology has many important applications. This paper uses data mining to describe the relationship between variables and the eigenvalues of data sets or some data sets, that is, association rules. The following will introduce the basic concepts and applications of association rules.

Table 4 Processed data display

id	information quality	personalized service	trust in medical websites	trust in doctors	Perceived Risk	material rewards	Age > 46	gender	disclosure
1	good	no	yes	yes	no	yes	no	male	yes
7	bad	no	yes	no	yes	yes	yes	female	no
56	good	yes	yes	yes	no	yes	no	female	yes
455	good	no	yes	yes	no	yes	no	male	yes
1026	good	no	yes	yes	no	yes	no	female	no

4.2 Association Rules

Association rule extraction is an important application in data mining. At the beginning of the big data era, China has been in the dilemma of "massive data and lack of information". Even if it has strong data, it cannot find useful information in all kinds of data. In the e-commerce industry, they bundle the goods according to the personalized characteristics of users by analyzing the purchase habits and needs of users, such as the famous cases of beer and diapers. In the transportation field, the causes of traffic accidents are diverse, such as drunk driving, overspeed, etc. Through mining association rules, we can deeply explore the combined influencing factors leading to traffic accidents. This paper uses association rule to mine the relationship of factors affecting users' intention to disclose privacy.

Algorithm 1 MD-Apriori

Input：data D，minsupport;
Output：Frequent itemsets *L*;
1. C_1**= {candidate 1 -itemsets};** // Find all candidate itemsets of 1-term frequent itemsets, here, $K = 1$, $L_K = \emptyset$
2. $L_1 = \{ c \in C_1 \mid$ **c.support ≥ minsupport};**// Find all 1-item frequent itemsets
3. for(*K* **= 2,** L_{K-1} **!= Null,** *K*++**){** // Connect steps, until the maximum frequent itemset cannot be generated
4. C_K **= sc_candidate (** L_{K-1} **);** // Generate frequent itemsets of *K* elements
5. if the frequent items in C_K **are not in [attribute set];**
// Pruning frequent itemsets, attribute set refers to the influencing factors in this paper, //such as "information quality"
6. delete this frequent item;
7. for each *t* **in** *D* **{**
8. C_t **=subset(** C_K **,** *t***);** // Get a subset of *t*
9. for each candidate $c \in C_t$ **:** // Each frequent itemset in *t*
10. c.count = c.count + 1;} // Count candidate itemsets
11. c.support = c.count/len(*D***);** // Calculate support of *c*
12. $L_K = \{ c \in C_k \mid$ **c.support ≥ minsupport};**// Obtain *K*-term frequent itemsets
13. for each item in L_K **{**
14. if the attribute of item is not in [result set]: // Result set contains the intention to disclose privacy
15. delete this frequent item;}}
16. Return *L*;// Returns all frequent itemsets

4.3 Apriori

Apriori is the most basic and widely used algorithm in association rule mining. The core idea of Apriori is to mine the frequent itemset of transactions, use multiple iterations to calculate the frequent itemset in the data and find out the strong association relationship. This paper uses Apriori to mine the influencing factors of patients' intention to disclose private information. The related concepts of association rule algorithm will be introduced in detail below. The two most important concepts in association rules are support and confidence.

(1) Support

Support indicates the probability that the next attribute value X of a dimension appears in all records D, as shown in the following formula:

$$\text{Support}(X) = \text{Count}(X) / \text{Count}(D) = P(X) \tag{1}$$

(2) Confidence

Confidence represents the probability that attribute X and Y appear simultaneously in all records D, as shown in the following formula:

$$\text{Confidence}(X \to Y) = \text{Support}(X \cup Y) / \text{Support}(Y) = P(Y \mid X) \tag{2}$$

(3) Candidate itemset

Candidate itemset refers to the itemset obtained by downward merging and is represented as C_K. K refers to the number of items.

(4) Frequent itemset

Frequent itemsets represent itemsets whose support is greater than or equal to the given minimum support, expressed as L_K. The algorithm has the property of downward closure, that is, any nonempty subset of frequent itemset is also frequent itemset.

(5) Pruning step

The pruning step means that according to the property of Apriori, all nonempty subsets of frequent itemsets must also be frequent itemsets, so the itemsets that do not meet this property will not exist in. This process is pruning.

The steps of mining the influencing factors of patients' intention to disclose private information using Apriori are briefly summarized as the following two steps:

Step 1：Find all frequent sets. The process includes: A. Scanning all data. B. Count the value of each factor. C. Compare the statistical support with the minimum support threshold. D. Generate frequent itemsets. E. Connect, prune and generate candidate itemsets for each frequent itemset. F. Repeat a-e until a larger frequent itemset cannot be found.

Step 2：Generate association rules. For each frequent itemset, the confidence of all nonempty subsets is calculated. If it is greater than the minimum threshold, the rule is output.

4.4 MD-Apriori

When the classical Apriori algorithm processes multiple dimension data, it is easy to connect different attribute values of the same dimension, resulting in useless rules. The classical Apriori algorithm will produce many useless rules. In this study, we explore the impact of some factors on patients' willingness to disclose privacy. Therefore, the conclusion of the rule is uniquely specified. Although the original algorithm can ensure that the conclusion is unique, it cannot specify the conclusion. Therefore, we propose MD Apriori algorithm. For example, we hope to obtain such useful rules, such as material reward ("yes")^information quality ("good")^personalized service ("yes") −−→ whether to disclose privacy ("yes"), rather than such useless rules, Material reward ("yes")^material reward ("no")^information quality ("good")^personalized service ("yes") −−→ whether to disclose privacy ("yes"), In this case, different attribute values ("yes", "no") are connected to each other in the data of the same dimension (material reward dimension), or such useless rules as material reward ("yes")^information quality ("good") −−→ personalized service ("yes"), that is, the final rule is not what we want. Moreover, the calculation process of classical Apriori needs to scan all data frequently. When the data file is large, the calculation efficiency is low.

Considering the above problems, we propose a MD-Apriori algorithm. On the basis of the original algorithm, MD-Apriori adds constraints, adds the judgment of whether the frequent itemset contains multiple dimensions in the connection step, and only considers the generation of frequent items of "whether to disclose privacy" in the rule tree pruning stage. For example, the occurrence of D event may be affected by factors A, B and C. in MD-Apriori algorithm, two or more of the three dimensions A, B and C will appear in the association rules mined, it makes the expression of rules richer, reduces the number of useless rules in one dimension, and does not appear in association rules like A −−→ B. Since the frequent items are fixed, the rules needed in this experiment can be found faster and more succinctly. The improved pseudo code is shown in Algorithm 1, codes are shown in bold, annotations are added where necessary.

MD-Apriori adds the generation limit of frequent itemsets based on the original algorithm. See line 5 of the pseudo code. Here, by judging whether the items in the current frequent itemset are independent variables, the useless rule that the dependent variables appear in the independent variable set is avoided. Line 14 of the pseudo code determines whether the items of the current frequent itemset appear in the result set, so as to avoid the useless rule that the independent variable appears in the dependent variable set.

4.5 SPSS

At present, scholars usually use three methods to study the relationship between variables. One is the traditional statistical analysis represented by SPSS, the other is the structural equation model based on covariance represented by Amos, and the other is the partial least squares path model represented by PLS-Graph. Ordinary regression analysis is suitable for studying a single potential variable, and cannot deal with multiple simultaneous equations at the same time to solve the problems of intermediary and regulatory effects at one time. Structural equation model cannot only realize traditional analysis such as factor analysis and regression analysis, but also deal with multiple potential variables at the same time to provide model fitting evaluation. The path model of partial least squares method is robust to collinearity problems and abnormal distribution data, and is suitable for small samples. Therefore, when the sample size is greater than 300 and the latent variables do not show too much collinearity, we use SPSS to analyze the sample data.

4.6 Bayesian Network

Bayesian network, based on Bayesian statistics, is a directed acyclic graph, including nodes and directed edges. Nodes can represent any random variable. The directed edges between nodes represent the relationship between nodes, and the strength of the relationship is expressed by conditional probability. If there is no parent node, it is expressed by a priori probability. Bayesian network is one of the most effective models in the field of uncertain knowledge and reasoning. It can also reason for incomplete or uncertain information. Some concepts are introduced below.

(1) Conditional probability

Conditional probability refers to the probability that event A occurs under the condition that another event B has occurred. It is defined as: if A and B are two events and $P(B) > 0$, it is called

$$P(A|B) = P(AB) / P(B) \tag{3}$$

where, it is the probability of event A under the condition that event B occurs.

(2) Prior probability

Prior probability is the probability of various events determined according to historical data or subjective judgment. It is defined as: $A_1, A_2, A_3 \ldots, A_n$ is n events in the sample space, A_i can be obtained by previous data

analysis or estimation based on prior knowledge, it is called $P(A_i)$, a priori probability whose value has been determined before the experiment.

(3) Bayes Rule

A and B are two events, $P(A)>0$, $P(B)>0$, according to the multiplication theorem, $P(AB)=P(A|B)P(B)=P(B\mid A)P(A)$, we can get:

$$P(A|B)=P(B\mid A)P(A)/P(B) \tag{4}$$

This formula is called Bayesian formula.

The Bayesian rule of event form is defined as: $A_1, A_2, A_3 \ldots, A_n$ is a collectively exhaustive events of E, and $P(A_i)>0$, B is any event of E, then

$$P(A_i|B)=P(A_i)P(B\mid A_i)/\sum_{j=1}^{n}P(A_i)P(B\mid A_i) \tag{5}$$

where, it is Bayes Rule, where $P(A_i)$ is a priori probability, $P(A_i\mid B)$ is a posteriori probability.

5 EXPERIMENT

5.1 Study Process

Fig. 1 shows the research process of this paper. Firstly, the paper uses classical Apriori and improved MD Apriori to obtain the influencing factors and association rules affecting patient privacy disclosure. Then, the paper analyzes 331 valid questionnaires through SPSS, and obtains the results as the verification of data mining results. Finally, this result is used to establish the prediction model. Next, we will introduce the process in detail.

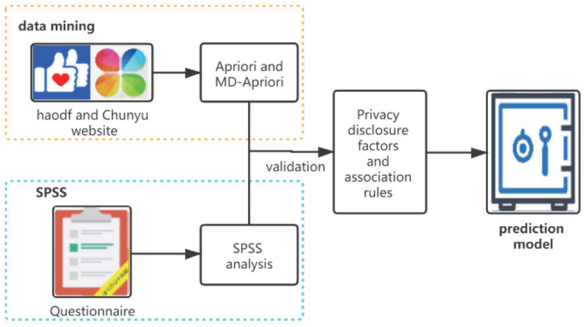

Figure 1 The research process of this paper

5.2 Data Mining

The comparative experiment was carried out on windows 10. The experimental data used the data shown in Chapter 3 of this paper, a total of 1256 data. The original Apriori and the improved MD-Apriori algorithm were used to mine association rules for the factors affecting the intention of online medical privacy disclosure. In the experiment, the support was set to 0.5% and the confidence was set to 70%. Some results are shown in Tabs. 5 and 6.

Because there are many independent variables and rules in this study, only some of them are selected for display. It is obvious that the improved MD-Apriori algorithm is better than the original Apriori algorithm in rule acquisition.

Table 5 Some results obtained by Apriori algorithm

Some results obtained by Apriori algorithm
1. Information quality (good), trust in doctors (yes), material reward (yes)−−→disclosure (yes), 0.8456
2. Information quality (good)−−→trust in doctors (yes), 0.7002
3. Trust in medical websites (yes), Perceived Risk (no)−−→disclosure(yes), 0.7563
4. Trust in medical websites (yes), trust in doctors (yes)−−→material reward (yes), 0.8795

Table 6 Results obtained by MD-Apriori algorithm

Results obtained by MD-Apriori algorithm
1. Information quality (good), trust in doctors (yes), trust in medical websites (yes), material rewards (yes)−−→disclosure (yes) 0.8952
2. Information quality (good), trust in doctors (yes), material reward (yes) −−→disclosure(yes), 0.8861
3. Trust in doctors (yes), trust in medical websites (yes), material rewards (yes), gender (male) −−→disclosure(yes), 0.7231
4. Trust in doctors (yes), material rewards (yes) −−→disclosure(yes), 0.7005
5. Information quality (good), trust in doctors (yes), personalized service (yes) −−→disclosure(yes), 0.8214
6. Perceived Risk (yes), trust in doctors (yes), material reward (yes) −−→disclosure(yes), 0.7452
7. Perceived Risk (no), trust in doctors (yes), material reward (yes) −−→disclosure(no), 0.7289
8. Trust in doctors (yes), material rewards (yes), gender (female) −−→disclosure (yes), 0.7305

From Tab. 6, we can observe that factors such as information quality, trust in doctors and material rewards have a positive correlation with disclosure intention, while trust in medical websites, age and gender have little impact on the results, and the impact of perceived risk on the results is uncertain, which needs to be further implemented in the questionnaire.

5.3 Questionnaire

This paper adopts the questionnaire method to collect data, and uses the existing questionnaire on the intention of online medical website users to disclose personal health information [13]. The reliability and validity test of the questionnaire shows that the questionnaire fully meets the psychometric requirements. The questionnaire is designed by three methods: (1) all measurement items are derived from the existing literature, then modified and supplied in combination with the situation of China's Internet medical website, (2) at the beginning of the formation of the questionnaire, invite experts in structural equation model to modify the questionnaire, (3) on the basis of expert modification, randomly distribute 50 questionnaires to online medical website users for small sample test to further supply the questionnaire.

Since the impact direction of perceived risk on information disclosure is not clear, based on the conclusion of data mining, this section divides perceived risk into five dimensions, namely physical risk, privacy risk, functional risk, psychological risk and social risk. The questionnaire includes 12 potential variables and 38 measurement items: information quality, personalized service, trust in medical websites, trust in doctors, physical risk, privacy risk,

functional risk, psychological risk, social risk, material reward, other variables, intention to disclose personal health privacy, the measurement items are in the form of Likert 7-point scale, and the measurement range changes from "very disagree" to "very agree".

5.4 SPSS Analysis

A total of 349 questionnaires were distributed in this paper. Excluding 18 questionnaires with too short filling time and the same 10 consecutive options, a total of 331 valid questionnaires reached the number of samples required by the theory. We use SPSS to analyze and obtain the correlation coefficient between each risk factor and the intention to disclose private information. The results are shown in the Tab. 7.

Table 7 Correlation coefficients between factors and disclosure intention

Independent variable	dependent variable	Pearson Correlation Coefficient
information quality	disclosure intention	0.896
personalized service	disclosure intention	0.464
trust in medical websites	disclosure intention	0.127
trust in doctors	disclosure intention	0.362
physical risk	disclosure intention	−0.684
privacy risk	disclosure intention	−0.725
functional risk	disclosure intention	−0.486
psychological risk	disclosure intention	−0.128
social risk	disclosure intention	−0.089
Material rewards	disclosure intention	0.563
other variables	disclosure intention	0.179

Where, when the absolute value of Pearson Correlation Coefficient is 0.8-1.0, there is a strong correlation between the two; when it is 0.6-0.8, there is a strong correlation between the two; when it is 0.4-0.6, there is a medium correlation between the two; when it is 0.2-0.4, there is a weak correlation between the two; when it is 0.0-0.2, there is a very weak correlation or no correlation between the two. The results of questionnaire and SPSS show that information quality, personalized service, trust in doctors and material rewards have a positive impact on disclosure intention, among which information quality has a greater impact on doctors' trust. Physical risk, privacy risk and functional risk negatively affect disclosure intention, among which privacy risk has a greater impact. Trust in medical website, psychological risk, social risk, gender and age have little effect on disclosure intention.

The results show that different types of Perceived Risk have different effects on disclosure intention, which explains the problems left in the data mining experiment. The conclusion of the data mining method used in this paper is basically consistent with the conclusion of questionnaire and SPSS analysis, and the result is reliable.

5.5 Model Prediction

In this paper, Bayesian network is used to establish the prediction model of medical privacy information disclosure intention. The set of prediction variables is the factors with strong correlation with disclosure intention in the above results, such as information quality, Privacy Risk, etc. the prediction value range is $A=\{0,1\}$, "1" indicates that the intention to disclose behavior occurs, "0" indicates that it does not occur. Due to the occurrence of disclosure behavior is affected by many factors and is random and uncertain, its a priori probability is difficult to determine. Therefore, this paper defaults that whether the disclosure behavior occurs as an equal probability event, the probability is 0.5. Given the values of a group of influencing factors, the posterior probabilities of occurrence and non-occurrence of disclosure behavior are calculated respectively. By comparing the two posterior probabilities, we can judge whether disclosure behavior will occur.

In order to illustrate the effectiveness of Bayesian network, SVM support vector machine binary classification model is selected for comparative experiment. The experimental results are shown in Tab. 8.

Table 8 Model prediction results

prediction model	train set	test set
SVM	85%	79%
Bayesian network	93%	83%

where Bayesian network obtains 93% accuracy in the training set and 83% accuracy in the test set. The accuracy of Bayesian method in predicting disclosure intention is higher than that of SVM method in both data sets, and the accuracy in the test set reaches 83%, which is a relatively high value, which proves that the prediction model in this paper is effective.

6 CONCLUSION

Based on the relevant data of internet medical websites, this paper uses the classical Apriori and improved MD Apriori data mining technology to mine the association rules of the factors affecting users' intention to disclose privacy. It is proved that the improved MD Apriori algorithm is better than the classical Apriori algorithm in this experiment. Using questionnaire and SPSS for statistical analysis, the results are consistent with the conclusions of using data mining, which proves that the risk factors mined in this paper are reliable. At the same time, it also provides suggestions for the construction of internet medical websites, that is, we should pay attention to improving the privacy protection and information quality of the website, and can increase appropriate personalized services and material rewards. Then, based on Bayesian network, this paper uses the mined association rules to establish a prediction model of online medical privacy disclosure intention. Through the comparative test with SVM model, it can be concluded that this model has good usability in predicting medical website users' privacy disclosure intention.

7 REFERENCES

[1] iResearch Interprets the Six Trends of Internet in China's Medical and Health Industry 2015.Shanghai iResearch Market Consulting Co.(eds.). *IResearch consulting series research reports*, *8*, 173-233.

[2] Li, Y.Q. & Pan, X.H. (2020). Analysis of the development of Internet plus medical health in China. *China Statistics*, *11*, 68-70. https://doi.org/CNKI:SUN:ZGTJ.0.2020-11-028

[3] Liang, X., Barua, M., Lu, R., Lin, X., & Shen, X. S. (2012). Health Share: Achieving secure and privacy-preserving

health information sharing through health social networks. *Computer Communications*, *35*(15), 1910-1920. https://doi.org/10.1016/j.comcom.2012.01.009

[4] Zeng, L. L. (2021). National network information security management measures for medical institutions will be issued. *Economic Information Daily*, *5*.

[5] Zang, G. Q., Han, M. X., & Zhang, K. L. (2021). Research on the Influencing Factors of Personal Information Disclosure Willingness from the Perspective of Persuasion Knowledge Management. *Information studies: Theory & Application*, *6*.

[6] Hua, H. C., Zhang, T. W. Y., Chen, Y. J., & Peng, J. Y. (2021). Research on the influencing factors of users' information disclosure behavior in the information age. *China New Telecommunications*, 7, 78-82. https://doi.org/CNKI:SUN:TXWL.0.2021-07-037

[7] Li, Y., Mou, J., Ye, L., Long, J., & Huang, W.W. (2021). An empirical investigation of the utilitarian, social benefits in LBS information disclosure - The moderating effect of the gender based social role theory. *International Journal of Information Management*, *56*, 102243. https://doi.org/10.1016/j.ijinfomgt.2020.102243

[8] Fan, A., Wu, Q., Yan, X., Lu, X., Ma, Y., & Xiao, X. (2021). Research on Influencing Factors of Personal Information Disclosure Intention of Social Media in China. *Data and Information Management*, *5*(1), 195-207. https://doi.org/10.2478/dim-2020-0038

[9] Cheng, H. P., Wen, X. Y., & Su, C. (2020). A Model of Factors Influencing Privacy Disclosure Intention of Social Media Users: An Empirical Study. *Library and Information Service*, *64*(16), 92. https://doi.org/10.13266/j.issn.0252-3116.2020.16.010

[10] Xia, H. M. (2018). *An empirical study on the influence of technical features of privacy feedback on mobile commerce user's personal information disclosure intention*. Master's thesis, Central China Normal University.

[11] Xie, Z. & Yang, J. L. (2020). Empirical Research on Users' Willingness to Disclose Privacy Information from Perspective of Smart Libraries. *Library Tribune*, *9*, 69-78.

[12] Wang, Y. C. (2018). Research on the influencing Factors of Users' Health Information Disclosure Intention in Online Medical Community. *Journal of Information Resources Management*, *1*, 93-103+113. https://doi.org/10.13365/j.jirm.2018.01.093

[13] Jiang, Y. Q. (2017). *Research on Influencing Factors of Personal Health Information Disclosure Intention of User in Online Medical Websites*. Master's thesis, Wuhan University.

[14] Chen, T. Y., Zhang, C. Y., Li, Y., & Xie, S. Y. (2017). Research on Tourist Satisfaction on 5A Scenic Spots in Hubei Province Based on the Association Rules. *Resource development and market*.

[15] Fang, H. Y., Zhu, Z. L., Feng, Z. L., Hou, W. X., & Zhang, H. C. (2021). Study on Prescription Medication Law of Lung Carbuncle (Lung Abscess) Based on Data Mining. *Journal of China-Japan Friendship Hospital*, *4*, 223-225. https://doi.org/CNKI:SUN:ZRYH.0.2021-04-007

[16] Qi, X. Z. & Ji, Z. Z. (2021). Analysis of Medication Law of Yu Zhiqiang in the Treatment of Hypertension Based on Data Mining. *Journal of Traditional Chinese Medicine*, *8*, 19-22. https://doi.org/0.16808/j.cnki.issn1003-7705.2021.08.006

[17] Wang, S. J., Fan, Y. X., Pan, C., Zhao, T. Y., Han, C. C., & Du, L. (2021). Short-Term Wind Speed Combined Forecasting Based on Optimized Elm of Principal Component Reduction Clustering. *Acta Energiae Solaris Sinica*, *8*, 368-373. https://doi.org/10.19912/j.0254-0096.tynxb.2019-0564

[18] Ye, Y. J. (2018). *Research on Mining Algorithm and Prediction Model of Traffic Accident Risk Factors Based on News Data.* Master's thesis, Beijing University of Technology.

[19] Li, K., Lin, Z., & Wang, X. (2015). An empirical analysis of users' privacy disclosure behaviors on social network sites. *Information & management, 52*(7), 882-891.

[20] Yu, L., Li, H., He, W., Wang, F. K., & Jiao, S. (2020). A meta-analysis to explore privacy cognition and information disclosure of internet users. *International Journal of Information Management, 51*, 102015.

Contact information:

Yong WANG
School of Economics and Management,
Beijing Jiaotong University,
Beijing 100044, China
E-mail: 17113137@bjtu.edu.cn