Enhancing Privacy and Security in Personalized Web Search

Priyanka A. Sonawane
PG Student
Department of Computer Engineering
SSBT's College of Engineering & Technology,
Bambhori, Jalgaon [M.S], INDIA

Satpalsing D. Rajput
Assistant Professor
Department of Computer Engineering
SSBT's College of Engineering & Technology,
Bambhori, Jalgaon [M.S], INDIA

ABSTRACT
Personalized Web Search is a promising way to improve the accuracy of web search and has been attracting much attention recently. Data classification and prediction using searching are used for many purposes. Privacy and security of Personal information has the more challenging task in web mining. In existing system greedy algorithm is used and it generates decision tree which stores split pattern. If split pattern is disclose then complete tree data can be retrieved. So this can compromise privacy due to which it is unsecured. In proposed system train the system with dataset and calculate the probability of output classes. The probability calculation is personalized to the training dataset and output is secured by providing enhanced privacy. In Proposed Approach, System improve the relevancy and prediction of the information in order to get more accurate result for effective personalized web search. Experimental evaluation shows that, Results obtained by using proposed approach are more precise and relevant than existing approach.

Keywords
Personalized Web Search, Privacy, Security, AES Encryption and Decryption.

1. INTRODUCTION
In Personalized Web Search is generates decision tree which stores split pattern. If split pattern is disclose then complete tree data can be retrieved.

1.1 Background
Personalized Web Search has demonstrated its effectiveness in improving the quality of various search services on the Internet. The Personalized Web Search framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. The unitize generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. In existing system used Two simple but effective generalization algorithms, GreedyDP and GreedyIL, to support runtime profiling. While the former tries to maximize the discriminating power (DP), the latter attempts to minimize the Information loss (IL). By exploiting a number of heuristics, GreedyIL outperforms GreedyDP significantly.

A health care organization may implement Knowledge Discovery in databases (KDD) with the help of a skilled employee who has good understanding of health care industry. Knowledge Discovery in databases can be effective at working with large volume of data to determine meaningful pattern and to develop strategic solutions. Health care analyst and policy makers can learn lessons from the use of Knowledge Discovery in databases in other industries and apply Knowledge Discovery in databases to problems of health care industry Health care data is massive. It includes patient centric data, resource management data and transformed data. Health care organizations must have ability to analyze data. However there can be a concern of patient privacy. It is more than clear that the role of data mining is not to practice medicine but to improve useful information and knowledge so that better treatment and health care be provided.

Diabetes is one of the common and rapidly increasing diseases in the world. It is a major health problem in most of the countries. Diabetes is a condition in which your body is unable to produce the required amount of insulin needed to regulate the amount of sugar in the body. This leads to various diseases including heart disease, kidney disease, blindness, nerve damage and blood vessels damage. Heart is the important part of our body. Life is itself dependent on efficient working of heart. If operation of heart is not proper, it will affect the other body parts of human such as brain, kidney etc. Health care industry contains huge amount of heath care data, which contains hidden information. This hidden information is useful for making effective decisions. Computer based information along with advanced Data mining techniques are used for appropriate results.

1.2 Motivation
Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. Data classification is most commonly used among all domains. They are robust and easily applicable and can be placed easily and require less memory and provides better results. Nowadays, data classification and prediction using searching are used for many purposes like,

get weather forecast, to predict future disease to person etc. Recently classification is also used at various places for predictions. The Prediction is done by learning past data and generating future prediction using the same. While classification there is certain problem due to which the system fails. One of the major problems in recognition of parameters and classification classes, It fails at the time when parameter changes. This problem mainly arises when In proposed system are working for more than one prediction in a single system.

2. RELATED WORK
Zhu et al. in [1], presented a bayes-optimal privacy notion to bound the prior and posterior probability of associating a user with an individual term in the anonym zed user profile set also propose a novel bundling technique that clusters user profiles into groups by taking into account the semantic relationships between the terms while satisfying the privacy constraint.

Roca et al. in [2], presented a novel protocol specially designed to protect the users privacy in front of web search profiling. This system provides a distorted user profile to the web search engine. Proposed protocol improves the existing solutions in terms of query delay. Statistical results of the protocols performance show that the presented scheme improves previous proposals.

Xu et al. in [3], has given to receive personalized web services, the user has to provide personal information and preferences, in addition to the query itself, to the web service.. online anonymity is dealing with unknown and dynamic web users who can get online and offline at any time..

Shen et al. in [4], has given systematically examine the problems of privacy preservation in personalized search. In future there will be different levels of privacy protection provided by search engines depending on a user's preference for the tradeoff between the privacy concern and the improved search service quality.

Dou et al. in [5], given that straightforward click-based personalization strategies perform consistently and considerably well. It improves the search accuracy on some queries, but they also harm many queries. Since these strategies are far from optimal.

Kumari et al. in [6], discussed Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes. It used datasets for diabetes disease from the machine learning laboratory at University of California, Irvine. All the patients. data are trained by using SVM.

Ghumbre et al. in [7], given an intelligent system based support vector machine along with a radial basis function network is presented for the diagnosis Results obtained show that support vector machine can be successfully used for diagnosing heart disease. In future to apply on different techniques to check the result.

Rani et al. in [8], presented Neural Networks are much efficient in classification of the data. Accuracy is much important in the field of medical science. Using the neural networks multi layer perceptron classification for choosing the best and accurate classification model researchers get much accuracy.

Shouman et al. in [9], discussed Decision Tree is one of the successful data mining techniques used.. Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease. Yet its accuracy is not perfect.

Jesmin et al.in [10], has given is very easy to protest brain cancer before affected and reduce treatment cost. Implement software through online so that any person can easily check their brain cancer risk level.

3. PROPOSED SYSTEM

The proposed solutions introduced a new approach train the system with dataset and calculate the probability of output classes. The probability calculation is personalized to the training dataset and output is secured by providing enhanced privacy.

3.1 Proposed Naive Bayes Architecture

In Figure 1 shows architecture of the proposed system. Train the system with dataset and calculate the probability of output classes. The probability calculation is personalized to the training dataset. Proposed system mainly consists of two units - Training phase and Testing Phase.

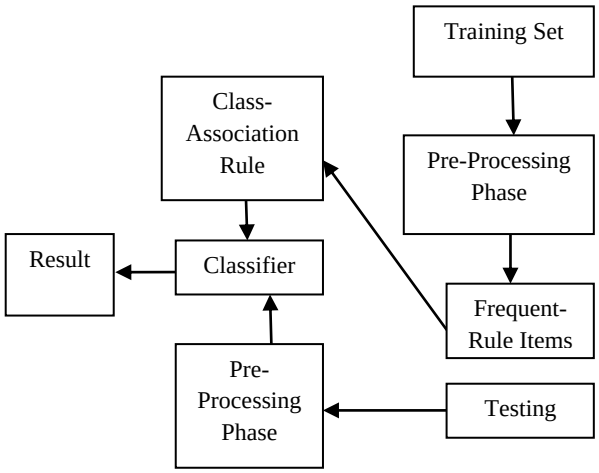

Figure 1. Proposed Naive Bayes Architecture

3.1.1 Training Set:-

Training set to trained the dataset for different combination. There are 3 dataset contains Diabetes, Brain and Heart. also 20 parameters contains Age, Gender, HBAC, BP, Plasma Glucose, Cholesterol, pulse rate, Hypertension, Hereditary, Foot ulcer, jaw pain, toothache, Headache, Nausea, Chest Pain, Radiation, Omitting, Vision, Swelling and Severity and 14 classes present in training phase.

3.1.2 Pre-Processing Phase:-

In pre-processing phase take the input parameters with possible values Yes, No, Low, medium, High, moderate etc. It is pre- process the data.

3.1.3 Frequent Rule Items:

In frequent rule items calculate individual probability for each parameter and also calculate the group probability.

3.1.4 Class Association Rule:-

All rules in the database is find out by Association rule mining that satisfy some minimum support and minimum confidence constraints. Class Association Rule is important part in Proposed Naive Bayes. It is easily solve the problems.

3.1.5 Testing Set

In Pre-processing phase calculate individual and group probability for all classes.

3.1.6 Classifier:

Naive Bayes is used as the classifier because of its simplicity and good performance in document and text classification. The classifier works by taking a document to be classified and calculating the probability that the documents belongs to each of the categorized with which the system has trained. The training category that is found to have calculate the posterior probabilities for each class. The class with the highest posterior probability is the outcome of the prediction

3.1.7. Results

Finally, Training and testing predictable result is generated.

3.2 Algorithm of Proposed Naive Bayes: Training Phase of Proposed Naive Bayes Algorithm

Require: N: No of parameters , M[N]: Matrix of N, P[N]: Probability of N, c: Classes, Pi: Individual Probability,

P_{max}: Maximum Probability, C_n: Number of classes, P_n: Number of probability.

Step 1: Initialize an array M[N] for N number of parameters.

N is between 1 to 20 i.e. 1 <= N <= 20

N is real number.

Step 2: Assume P[N] be array of possible values in M[N].

P[N] = {1,2,3,..}

Step 3: Learn values for each parameter N for all values of P[N]

$P_i = P(N)$

Where 1 <= N < 20

i <= N <= 14

Step 4: Learn Group values for all combinations

$P_n = P(n \mid n+c)$

Where n and c is no. of classes.

Testing Phase of Proposed Naive Bayes Algorithm:-

Require: N : Number of parameters, M[N] : Matrix of N, P[N]: Probability of N, c : classes, P_i : Individual Probability, C_n : Number of classes, P_n : Number of probability.

Step 1: Initialize an array M[N] for N no. of parameters

N is between 1 to 20 i.e. 1 <=N <=20

N is real number

Step 2: Assume P[N] be array of possible values in M[N]

P[N]= {1,2,3,…}

Step 3: for (i = 1; i <= C_n ; i++)

For (j = 1; j <= C_n ; j++)

Step 4: A] Calculate individual probability Pi for all classe

$P_i = P(C_n)$

Step 5: B] Calculate group probability for all combinations

$P_n = P (n \mid n+c)$

Where n and c is no. of classes

Step 6: C] Calculate prediction from individual and group

$P(C_i \mid P_n) > P(C_j \mid P_n)$

Step 7: D] Calculate maximum probability from the prediction

$P(C_i > P_n) = P(N \mid C_i) P(C_j) \mid P(N)$

$P_{max} < P (C_i \mid P_n)$

$P_{max} = P (C_i \mid P_n)$

End for

End for.

3.3 Advanced Encryption Standard:-

Advanced Encryption Standard is a slightly updated version of Rijindael algorithm. In this algorithm AES uses 256 bits. The key length consist variable 128, 192 and 256 bits.

There are several steps for Encryption.

3.1.1 Sub Byte

Sub byte 1st takes the byte and converts into its Hex value. After converting the Hex value the first bit of Hex value compare with the Look Up table in which 1st Hex bit is column and second Hex bit is Row. E.g. ASCII of A-hex 0 X 42 means here 4th column and 2nd row value in the lookup table used for substitution. This process completely used for confusion purpose.

3.1.2. Shift Rows

Shift row method is used for shifting the columns in the sub byte. In which first row is placed as it is 2nd column shifted to the 2nd block and 3rd column shifted to the 3rd block.

3.1.3. Mix Column

In Mix Column different values come after the shift rows these four byte each column in the state are mixed with as a 4-byte number and transformed to another 4 byte number.

3.1.4. AddRoundKey

In AddRoundKey the each byte on a state is XOR with Sub Key.

3.1.5. AES Decryption:

The process of decryption of an AES cipher text is similar to the encryption process in the reverse order. Each round consists of the four processes conducted in the reverse order –

- Add round key
- Mix columns
- Shift rows
- Byte substitution

Since sub-processes in each round are in reverse manner, the encryption and decryption algorithms need to be separately implemented, although they are very closely related.

4. RESULTS AND DISCUSSION

4.1 Performance Metrics:-

Table 1. shows that Time complexity of Existing GreedyIL algorithm.

Table 1. Time Complexity Table

Sr. No	Algorithmic Steps	Execution Times
1	if DP (q, R) < μ	logn
2	Insert <t, IL(t)> into Q for all t $T_H(q)$	n
3	Set s <- par(t, G_i)	1
4	Process prune-leaf G_i -> G_{i+1}	n
5	if t has no siblings, Insert <s, IL(S) >to Q	n
6	if t has siblings, Merge t into	n

	shadow-siblings	
7	if No operations on t's siblings in Q, Insert <s, IL(s)> to Q	1
8	Update the IL-values for all operations on t's siblings in Q	n

Time complexity for Existing GreedyIL algorithm:- nlogn.

In Table 2 shows that Time complexity for Training phase of Proposed Naive Bayes Algorithm.

Table 2 Time Complexity Table for Training Phase

Sr. No	Algorithmic Steps	Execution Times
1	Initialize an array M[N] for N no.of parameters	1
2	Assume P[N] be array of possible values.	1
3	Learn values for each Parameter.	logn
4	Learn group values for all combination	logn

Training phase time complexity is = logn

In Table 3. shows that Time complexity for Testing phase of Proposed Naive Bayes Algorithm.

Table 3 Time Complexity Table for Testing Phase

Sr. No	Algorithmic Steps	Execution Times
1	Initialize an array M[N] for N number .of parameters	1
2	Assume P[N] be array of possible values	1
3	Calculate individual probability	n
4	Calculate group probability	n
5	Calculate prediction from individual and group	1
6	Calculate maximum probability prediction	1

Testing phase Time complexity is :- O(n)

Overall time complexity of Proposed Naive Bayes: - logn + O(n)

4.2 Experimental Evaluation:-

The Table 4. Consists for simulation results sample values are taken into consideration according to values provided the results are generated.

Table 4. Example of Existing GreedyIL System and Proposed Naive Bayes System

	GreedyIL(Severity)	Proposed Naive Bayes(Severity)
For Heart	4	1
For Brain	3	1
For Diabetes	3	3

In Table 5 indicates that if the Time comparison of Existing GreedyIL and Proposed Naive Bayes algorithm.

Table 5. Result of GreedyIL and Proposed Naive Bayes for Time

Dataset Size	GreedyIL(Time)	Proposed Naive Bayes(Time)
50	0.04	0.01
100	0.063	0.029
200	0.079	0.042
500	0.092	0.057
1000	0.12	0.072

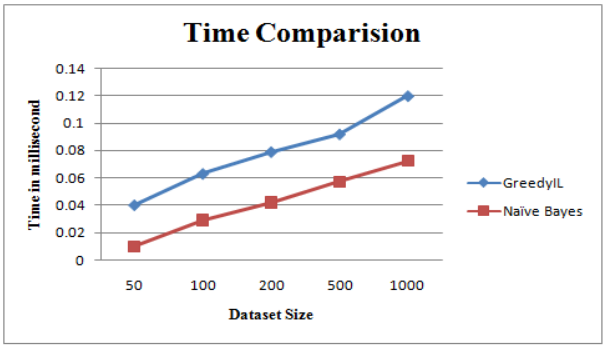

Figure 2. Dataset Size Vs Time

In Figure 2. shows the comparison of time for Existing GreedyIL and Proposed Naïve Bayes Algorithm and it is observed that, For every dataset size, 10 iterations are considered and average of 10 iterations is mentioned in above given table. Time is calculated by difference between the submission process is done to the results generated. Proposed Naive Bayes algorithm consumes less time as per existing GreedyIL algorithm. So Proposed Naive Bayes algorithm is better than existing GreedyIL algorithm.

Time = $\sum Time_{Result} - \sum Time_{Process}$

= 794ms - 507ms

=0.287 millisecond

Where , T_{Result} = Time in milliseconds from 1 Jan 1970 to time when results got generated.

$T_{Process}$= Time in milliseconds when process has stored.

In Table 6 shows Average Iteration result for GreedyIL algorithm and Proposed Naive Bayes algorithm.

Table 6. Result for Average Iteration Vs Yahoo

Iteration	GreedyIL	Proposed Naive Bayes
Distinct	85	8
Medium	75	48

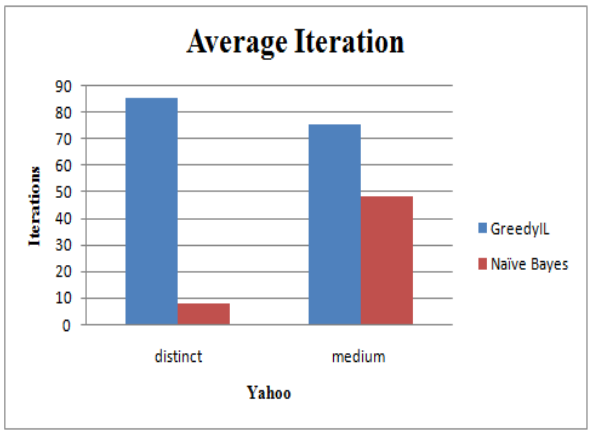

Figure 3. Graph for Average Iterations Vs Yahoo

In Figure 3 shows the comparison of average iteration for Existing GreedyIL and Proposed Naive Bayes Algorithm. In this figure proposed naive bayes distinct iteration is less than existing GreedyIL. And also medium iteration is also less than existing GreedyIL Algorithm. So Proposed Naive Bayes algorithm is better than Existing GreedyIL Algorithm.

In Table 7 shows Privacy Results of Existing GreedyIL algorithm and Proposed Naive Bayes algorithm. In this table proposed naïve bayes algorithm provide more privacy than existing GreedyIL algorithm. Proposed Naïve Bayes is more secure than Existing GreedyIL algorithm.

Table 7. Results for Privacy Threshold Vs. Average Precision

Privacy Threshold	GreedyIL	Proposed Naive Bayes
0.1	45	67
0.2	55	69
0.3	57	72
0.4	54	70
0.5	40	68
0.6	48	65

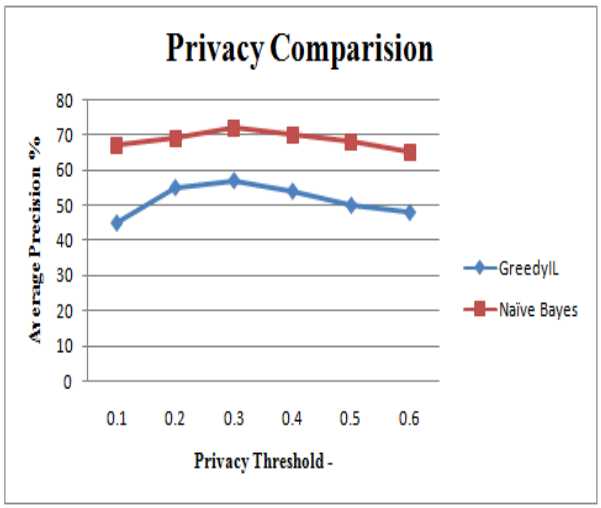

Figure 4. Graph for Privacy Threshold Vs Average Precision

The graph in Figure 4 represents Privacy Result of GreedyIL algorithm and proposed naive bayes algorithm. It is observed that Proposed Naive Bayes algorithm privacy is more than existing GreedyIL algorithm.

5. CONCLUSION

Personalized web search (PWS) is a general category of search techniques to providing better search results, which are tailored for individual user needs. Privacy and security of Personal information has the more challenging task in web mining. In proposed system used 3 Dataset, 20 parameters and 14 classes. As per experimental result shows that proposed naive bayes generates better results that GreedyIL algorithms on particular given scenario. Also Analysis of result proved that proposed naive bayes is better and efficient for time, iteration and privacy parameters used for comparison.

In future research will test proposed naive bayes system with different classification algorithm also can combine various algorithms to test the output results. Also it needs to be tested on more diseases with different datasets.

6. REFERENCES

[1] Y. Zhu, L. Xiong, and C. Verdery, “Anonymizing user profiles for personalized web search," ACM, April 2010.

[2] J. Castell-Roca, A. Viejo, and J. Herrera-Joancomart, “Preserving users privacy in web search engines," Elsevier, p. 1541- 1551, 2009.

[3] Y. Xu, K. Wang, G. Y. Ada, and W. C. Fu, “Online anonymity for personalized web services," ACM, November 2009.

[4] X. Shen, B. Tan, and C. Zhai, “Privacy protection in personalized search," ACM, pp. 4-17, June 2007.

[5] Z. Dou, R. Song, and J. Wen, “A large scale evaluation and analysis of personalized search strategies," ACM, pp. 581-590, May 2007.

[6] V. A. Kumari and R.Chitra, “Classification of diabetes disease using support vector machine," Journal of Computer Science, vol. 3, pp. 1797-1801, April 2013.

[7] S. Ghumbre, C. Patil, and A. Ghatol, “Heart disease diagnosis using support vector machine," ICCSIT, pp. 84-88, December 2011.

[8] S. Rani and M. S. Girdhar, "Analyse the heart disease and diabetes using artificial neural networks," IJARCSSE, vol. 4, pp. 1155-1157, August 2014.

[9] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," Australasian Data Mining Conference, vol. 121, pp. 23-29, 2011.

[10] T. Jesmin and K. Ahmed, "Brain cancer risk prediction tool using data mining," ACM, 2013.