A proposed strategy to secure web usage

Prof. Dr. Hillal Hadi Saleh

Dr. Soukaena Hassan Hasheem

University of Technology, computer sciences department.

Abstract:

With much data on the web, it can be difficult, frustrating, and seemingly impossible to find the exact information you need. There are many powerful search utilities on the web are called search engines, in addition to the ***visitor tracking*** in a web to study exactly the behavior of the web visitors, to improve the efficiency of that web. This research concentrates on a particular aspect, which is applying Data mining technique especially by association analysis algorithm on the encrypted web log files, that to ensure the privacy of the original data for these files. So since the input data introduced to mining algorithm is encrypted then the resulted association rules are encrypted that to ensure the privacy of the extracted knowledge. Then analyze the decrypted web log data file for the web usage, to study the visitor tracking. According to this study the server configurations and all the services will be improved.

Keyword:

Data Privacy, knowledge privacy, data mining, association analysis, web usage.

ستراتيجية مقترحة لحماية تعدين الويب

د.هلال هادي صالح د.سكينة حسن هاشم

الجامعة التكنولوجية/قسم علوم الحاسبات

الخلاصة:

مع الكميات الهائلة من المعلومات عبر الانترنيت من الصعوبة ايجاد مانريده بالضبط لذلك كان الدافع لبناء محركات البحث وتقنيات متابعة زائري المواقع على الويب لتحسين كفاءة المواقع لتقديم مايريده المستخدم تماما. هنا نقترح تطبيق تقنيات تعدين البيانات وخاصة خوارزمية تحليل القواعد المترابطة على قواعد البيانات الخاصة بزائري الويب والتي يجب ان تكون مشفرة اولا لحماية خصوصية زائري الويب بما ان البيانت المقدمة لخوارزمية التعدين مشفرة اذا القواعد المترابطة الناتجة ستكون مشفرة ايضا. بالتالي سيتم اعتماد هذه القواعد المشغرة ثم فك شفرتها وتحليلها ودراستها لتحسين مستوى الويب من ناحية خدمة الزائرين.

1- Introduction [1, 2, 3]:

With data mining the techniques of the search engines and visitor tracking called ***web mining***. The important task for web mining is web usage mining, which mines Web log records to discover user access patterns of Web pages. Analyzing and exploring regularities in Web log records identify potential customers for electronic commerce, enhance the quality and delivery of Internet information services to the end user, and improve Web server system performance. A web server usually registers a (Web) log entry, or Web log entry, for every access of a Web page. It includes the requested URL, the IP address from which the request is originated, and a timestamp. For Web-based e-commerce servers, a huge number of Web access log records are being collected. Popular Web sites may register the Web log records in order of hundreds of megabytes every day.

Since Web log data provide information about what kind of users will access what kind of Web pages, Web log information can be integrated with Web content and Web linkage structure mining. This information will help Web page ranking, Web document classification, and the construction of a multilayered Web information base as well. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or

KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications. Conceptually, the World Wide Web is comprised of three major components: the content of the Web, which encompasses documents available; the structure of the Web, which covers the hyperlinks and the relationships between documents; and the usage of the web, describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining.

2- The Proposed System:

In this research we describe a data mining framework for constructing web usage model. The key ideas are to mine web log record data for consistent and useful patterns of program and user behavior. So we could apply the web usage with high precision. We propose to use the association rules and frequent itemsets computed from logging data as the basis for guiding web usage. The general algorithm and all the details would be explained in the following sections briefly:

2.1- The Basic Algorithm:

To display the idea of this research in more clearly phase we would present the real implementation of that system as follow:

First, we must explain that this system would work as an offline system, because it in period collect the web log information and submit these information to this system to the study web tracking to improve the configuration of that web server according what it preferred by the visitor. The application program will collect the most important information about the visitor by the packet capture software see figure (1).

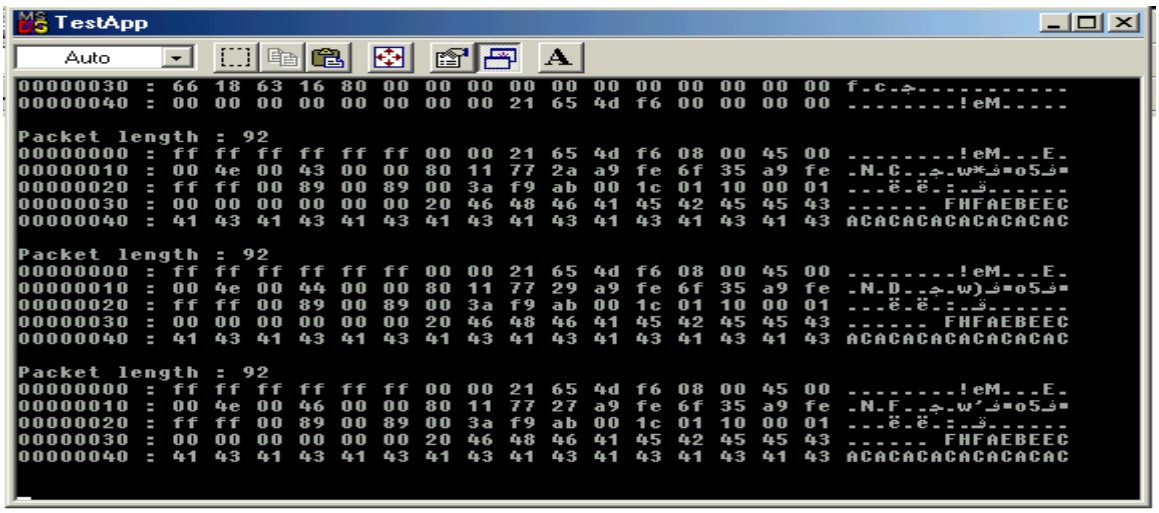

Figure (1) packet capture software

The software responsible is to take the hexadecimal values of local address, local port, remote address, remote port, state, time-stamp, and protocol type from the packet. And then this information would show in understandable information for the administration.

See figure (2), which represent the database of web log files, this database which will be the mined database. Let *A* be a set of attributes (local IP (A), local port (B), remote IP (C), remote port (D), state (E), type (F) and time stamp (G)), and *I* be a set of values on *A*, called items. Any

subset of *I* is called an itemset. The number of items in an itemset is called its length. Let *D* be a database with *n* attributes (columns). Define *support*(*X*) as the percentage of transactions (records) in *D* that contain itemset *X*. An association rule is the expression $X \rightarrow Y$, *c, and s*. Here *X* and *Y* are itemsets, and $X \cap Y = \Phi$. *s* = *support*(*X U Y)* is the support of the rule, and *c* = *support*(*X U Y)* / *support*(*X*) is the confidence.

TID	Local IP (A)	Local port (B)	Remote IP (C)	Remote port (D)	State (E)	Type (F)	Time stamp (G)
1	122.22.3.18	139	33.56.233.77	80	Listen	Tcp	2:30
2	122.22.3.18	139	44.56.78.22	50	Listen	Udp	5:50
.							
..							
.							
.							
.							

Figure (2) the database (D) for the web login/out file.

2.2- The Encryption Algorithm:

Apply encryption algorithm for all values of selected attribute we suggest to encrypt the values by asymmetric encryption algorithm since we deal with dynamic environment, internet, (here we will select RSA encryption algorithm), see figure (3).

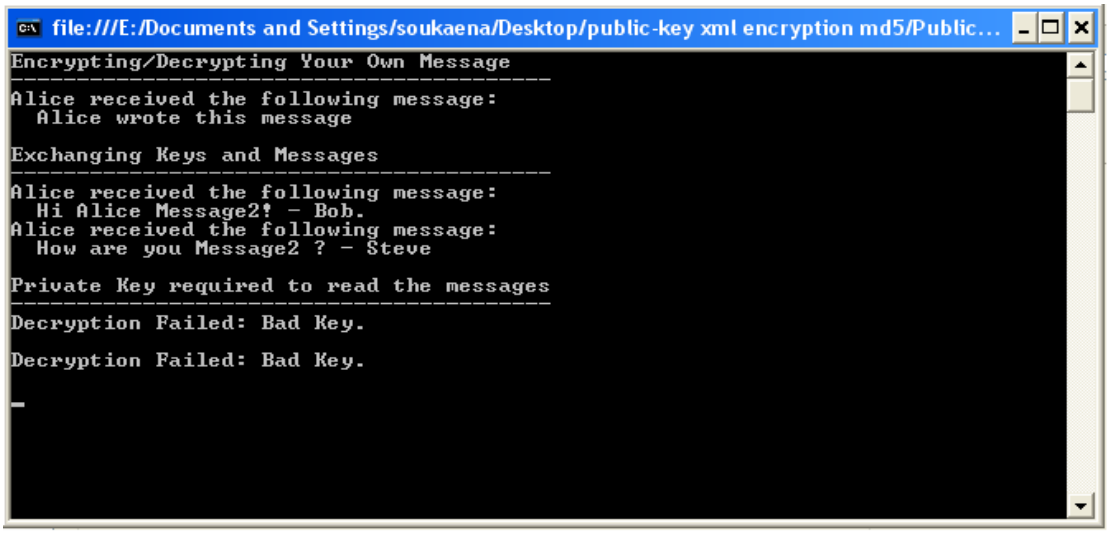

Figure (3): RSA algorithm.

After applying the RSA encryption algorithm on all the values of all the attributes of the original database, the last one will be as in the following figure (4).

TID	(A)	(B)	(C)	(D)	(E)	(F)	(G)
1	2FD39499A 0FF66BEBD BAA22CF6 674C03	BB1 B6D 70D6 5A74 D0	DF1AAF1E8 AA6D7F706 6D5CBB1B A6B887	33FDC0 B7B440 A666	60A7F DF750 521B4 3	5E574 0A615 452FA 4	8C69 D136 0BFB A08 D
2 . . .	2FD39499A 0FF66BEBD BAA22CF6 674C03	BB1 B6D 70D6 5A74 D0	90A85BE972 72C01689CE 23BF833E79 1B	EF15B4 76C4EA 8E5E	60A7F DF750 521B4 3	09228 D7F10 285679	D3B4 8996 BB85 87B7

Figure (4) the encrypted database (D) for the web login/out file.

2.3- The Web Usage Algorithm:

Now after completing the suiting of the database for the mining and encrypting this database, the next step will deal with that new D, which is represent un clear web log data (encrypted database). The dealing is to analyze the data of web log included in the new D for study and trace the visitors tracking, and according to these observations the web administrators would improve their servers and all their services. The analysis is done by the apriori DM algorithm, which is explained in the previous section.

We look for correlation among values of different features (attributes), and the (preprocessed) log data usually has multiple columns of features, each with a large number of possible values.

For example, visitor tracking can give the time periods encrypted patterns for which visitors access your site:

The encrypted patterns:

F96F0C5D8D201F6E00860522110F4924388C383198D63BDE8EB43E63D2659282A4E2D13EC30E3EAD

- 719BDBB713330230F3A46D578709F0B78DC9510922EEADF45FC445FEFDCA0D28416F3D44ACE72AD0B9308659D3F94399C657F2A57B84225FF63A99030E8D030748122759EEBF7F8FEF6EE45BCD9EB456E1115827CF0A08A18A9C7465FD4F48FE723D20E34305BE21
- 64F66B87BF6C6BA9BE14F79BF5D6B8DC1CC004F86B8AE52D5FC445FEFDCA0D28CCD8EC2CB723298446B2B6440ADC133239A63CF014C117C232C80FA1EA8DF95366DEF763ADBAC1C6

- AF7EF0E076CB761C0D2BC188A7615BB69999F1C5685872C85FC445FEFDCA0D289565C474F3E89526EAF5AB8BF6773CD30382E4107D9719C823599045A320FEFB
- 057F8D5B53D5388A32551782B53D6CEAF4AA66227AC61EE55FC445FEFDCA0D2870E5E684439FD2031A6B205B48FBCEBB4E518E85B8E12BE9F4B90750272B0565106DF26C476DDAB861C5222CCE662F9B1B2C9A4089B7A06861C6C3B9F07F332F
- 238B1E94479F595E91CB1EAE6CAADFBA05B7ED677A94ABC31182EF2F4F985A6892EA7DD3F17449D1B1194F1C422DAEB6445A9870AF7C8BBEEA077A7042C6EF16B14934DB97A73DF1F4DE277BC9B9042849C2B093EAC738CD2442049327481AF3C90A4BCCF2AAE70F9CF777F3A906345D
- 486EE164219C7CF3F81FBB902709A8E29DBC9081487D5A557159C0E0FB13C94F29D5343B1992F5F4C6B209CF524708CA

The decrypted and analyzed patterns:

For the 24 hours covered on 13-apr-99:

- 5 a.m. – 6 a.m. = 23 visitors enter by UDP protocol on port 80 from EUR countries only for games.
- 6 a.m. – 7 a.m. = 35 visitors from different countries for email.
- 7 a.m. – 8 a.m. = 42 visitors from Arabian for newsgroup.
- 8 a.m. – 9 a.m. = 69 visitors almost those are merchants dealing only with e-commerce.
- 9 a.m. – 10 a.m. = 105 visitors those using searching techniques for some of their interested subjects.
- 10 a.m. – 11 a.m. = 323 visitors shopping.

From these discovered patterns by DM algorithm the web site administration would be able to depend on these patterns for reconfigure their site servers according to the way preferred by visitors, and almost make their predictions about the track of visitors in the coming times.

3- Conclusion:

In context with the results of the present study it can be concluded that:

- Implementation of the web usage is a great help for support the reputation of the site, this by tracking the visitors and reconfigure the server organization according to it.
- Building web usage on the web log data after apply the encryption on the original data of these web log files, then the extracted knowledge pattern also will be encrypted, that will present high level of privacy for web usage.
- We suggest in the future work on this field making the keys of the RSA algorithm updated periodically, such as each week.

References:

1- M. S. Chen, J. Han, and P. S. Yu. **"Data mining: An overview from a database perspective"**. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

2– U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. **"Advances in Knowledge Discovery and Data Mining"**. AAAI/MIT Press, 1996.

3- J. Han and M. Kamber. **"Data Mining: Concepts and Techniques"**. Morgan Kaufmann, 2000.