DATABASE SEMINAR REPORT

"Detecting Anomalous Access Patterns in Relational Databases"
Ashish Kamra, Evimaria Terzi, Elisa Bertino

Sergey Chernokozinskiy
sc242@buffalo.edu

GENERAL OVERVIEW

Intrusion Detection (ID) techniques are essential component of any strong security system. However, the area of Intrusion Detection for DBMS has not been extensively researched. Furthermore, already proposed solutions of ID systems for operating systems or networks are not suitable for DBMS. Mainly, the reasons are DBMS specific architecture and different scope of tasks performed by DBMS.

A novel approach for ID systems tailored to DBMS is proposed in the paper. The approach is mainly aimed to detect insider threats, thus threats that originate from the legitimate users.

The main idea underlying the approach is to build profiles of normal user behavior and later use these profiles to detect anomalous queries issued by users. In order to build profiles the information from the database log files are extracted and processed. There is an assumption in the paper that each audit log record represents exactly one SQL command issued by the user. Every single SQL command from the audit log files is transformed to the new representation which is called quiplet.

Quiplets are basic units for forming profiles and may have three different levels of representation depending on the amount of information in every quiplet. Each quiplet has five fields: SQL COMMAND, PROJECTION RELATIOIN INFORMATION, PROJECTION ATTRIBUTE INFORMATION, SELECTION RELATION INFORMATION, SELECTION ATTRIBUTE INFORMATION).

The architecture of the proposed ID system has two phases: training phase and detection phase. In the training phase the information from the database log files are extracted and used to build profiles. In the detection phase every new single query issued by a user is compared against profiles to detect anomalous behavior.

However, the precise architecture of an ID system depends on two different scenarios: the first one when the database has a Role Based Access Control (RBAC) and thus each user has a role specifying his privileges. In the second scenario no RBAC is available.

When DBMS has a RBAC in place, in the training phase profiles are formed using user's roles and thus the classification is not required. For the detection phase the Naïve Bayes Classifier is used.

In the case when no RBAC is available for the training phase all users are clustered in groups either by k-centers or k-means algorithms. For the detection phase can be used either Naïve Bayes Classifier or detection outlier methodology.

The paper is concluded by the set of experimental results.

DISCUSSION

It should be mentioned that the paper is one of few papers published in the area of ID systems for DBMS. It uses the Machine Learning techniques for the both training and detection phases.

One of the main questions of the proposed approach is its practical application. It's not quite frequent that users interact with a database thru "pure" SQL queries, i.e. queries that are not really written but formed automatically with the help of the application interface (e.g. drop-down menus, visual selection and etc.). In such cases a user is bound to the interface. For example if the user has only "select" clause in a drop-down menu he obviously can't choose "insert" or "update" to formulate the new query.

Another possible issue is that the authors considered only simple SQL queries. They don't take into account any nested queries and queries with HAVING and GROUP BY clauses. It narrows down the applicability of the approach and makes impossible to query a database in more sophisticated ways. It should be mentioned that queries that give the same result (the same amount of information) may be formulated in many different ways. But having just simple queries with counting of the number of relations and attributes don't capture this aspect.

The proposed in the paper methodology requires adjustment and tuning for many parameters, e.g. m for the Naïve Bayes Classifier, D for the statistical test, K for the number of clusters. Some of these parameters play very important role in the overall performance of an ID system. Thus the adjustment of parameters might result in the maximum performance of an ID system or might reduce it to zero. Even more the wrong tuning might decrease the overall performance of DBMS when many normal queries are dropped and not executed. It's not quite clear how to select those parameters. There is no algorithm or any suggestions described in the paper. However it's a very crucial point. For example if it's fuzzy how to determine the optimal number of clusters (parameter K) for the case when no RBAC is in place. If a database has a large user population then the small number of clusters will result in high rate of False Positive types of error. On the other hand with a too large value of K the False Negative error rate will be high. There is no way to determine the value of K on a regular basis.

The paper doesn't touch any possible attacks to the ID system. Probably it's beyond the scope of the paper but every security solution should be evaluated for possible attacks, otherwise this opportunity gets the potential intruder. Theoretically at least one possible attack looks quite feasible. The main idea of the attack is to pose queries iteratively over the DBMS. These queries are not outliers but very close to them. For every such a query the profile (or cluster) might be updated, thus comprise the query and increase the size of profile. On a certain step when the profile is big enough the outlier query will not be recognized.

For the experimental evaluation have been used both synthetic and real data sets. However the experimental evaluation doesn't look to be very robust. In order to generate synthetic data set Zipf distribution has been chosen. There is no clear motivation in the paper why Zipf distribution is the one that the authors chose and if there are any other possible alternatives. It might be the reason that some strange results appeared during the experiments, e.g. when the False Negative error rate is high and lies in the range of 40-100 percent.

Nevertheless that thru the seminar's discussion many improvements and suggestions came up, the amount of work has been made by authors is considerable and fit in the scope of a conference paper. The reading of the paper requires fair knowledge of SQL and DBMS but good understanding of Machine Learning techniques. The overall organization is logically reasonable; the information is presented in the rational pace and does not have logical inconsistency.