ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

A Survey on Activity Detection using Data Mining

Santosh S.Gurav, Prof. S. R. Todmal

Student, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India.

Professor, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India.

ABSTRACT: Today, various applications require the ability to monitor a continuous stream of fine-grained data for the occurrence of certain high-level activities. A number of computerized systems—including ATM networks, web servers, and intrusion detection systems—systematically track every atomic action we perform, thus generating massive streams of times tamped observation data, possibly from multiple concurrent activities. In this paper, we address the problem of efficiently detecting occurrences of high-level activities from such interleaved data streams. A solution to this important problem would greatly benefit a broad range of applications, including fraud detection, video surveillance, and cyber security. We define algorithms for insertion and bulk insertion into the tMAGIC index and show that this can be efficiently accomplished. We also define algorithms to solve two problems: the "evidence" problem that tries to find all occurrences of an activity (with probability over a threshold) within a given sequence of observations, and the "identification" problem that tries to find the activity that best matches a sequence of observations. We introduce complexity reducing restrictions and pruning strategies to make the problem—which is intrinsically exponential—linear to the number of observations. Our experiments confirm that tMAGIC has time and space complexity linear to the size of the input, and can efficiently retrieve instances of the monitored activities.

KEYWORDS: Activity detection, indexing, stochastic automata, times stamped data, Data mining applications.

I. INTRODUCTION

There are numerous applications where we need to monitor whether certain (normal or abnormal) activities are occurring within a stream of transaction data. For example, an online store might want to monitor the activities occurring during a remote login session on its Web site in order to either better help the user or to identify users engaged in suspicious activities. A company providing security in an airport might want to monitor activities in a baggage claim area or in a secure part of the tarmac in order to identify suspicious activities. A bank might want to monitor activities at its automatic teller machines for similar reasons [2].

It is well recognized that models of activities are likely to be uncertain. We can rarely predict exactly how a particular activity may be executed, especially as a large number of irrelevant activities might be intermixed together. As a consequence, though early models of activities were "certain" about what constituted an activity and used logical methods or context-free grammars, more recent activity detection is based on either graphical models, or stochastic automata in which vertices correspond to observable atomic events [2].

However, most existing work on stochastic activity recognition has two main limitations. First, they often do not account for the time between observations associated with an activity. For instance, Fig. 1 shows an example of an online bill payment activity [1].

II. RELATED WORK

Limitations of traditional database management systems in supporting streaming applications and event processing have prompted extensive research in Data Stream Management Systems (DSMSs). An early yet

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

comprehensive survey of relevant issues in data stream management was presented in [2]. Amongst the several systems resulting from research efforts in this direction, of particular relevance is Telegraph CQ [1], a streaming query processor that filters, categorizes, and aggregates flow records according to one or more CQL [2] continuous queries, generating periodic reports. Differently from traditional queries on static data collections, results of continuous queries on streaming data need to be periodically and incrementally updated as new data is received. A significant portion of research in this area has been devoted to optimization of continuous queries [2]. Other works target the recognition of events based on streams of possibly uncertain data [1].

Although the system we propose in this paper operates on streams of observation data, the scope of our work is drastically different from the scope of DSMSs. In fact, we are not interested in retrieving a set of data items satisfying (exactly) certain conditions and keeping this set up to date as new data items are received. Instead, we are interested in finding sets of records such that, with a probability above a given threshold, the records in each set together constitute the "evidence" that a given activity occurred in a specific time interval. Additionally, we want to track partially completed activity occurrences. To the best of our knowledge, DSMSs do not provide support for this type of probabilistic inference. Moreover, there has been limited work on efficient indexing to support probabilistic activity recognition. The aim of past work on indexing of activities was merely to retrieve previously recognized activities, not to recognize new ones. Such work includes that of Ben-Ariet al. [2] who use multidimensional index structures to store.

III. PROCESS OF DATA MINING

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years. Wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

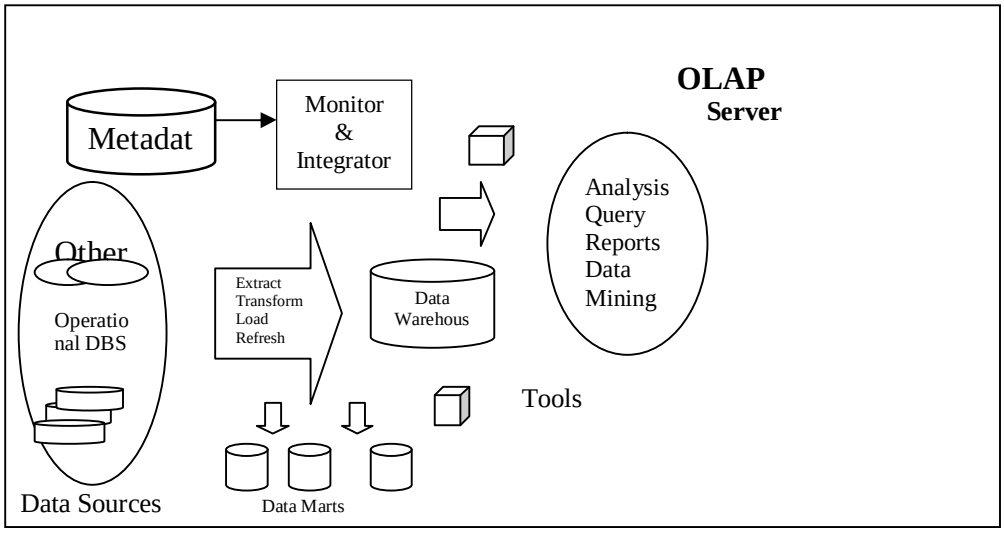

Figure 1: Data mining process

Data mining is about finding insights which are statistically reliable, unknown previously, and actionable from data (Elkan, 2001). This data must be available, relevant, adequate, and clean. Also, the data mining problem must be well-defined, cannot be solved by query and reporting tools, and guided by a data mining process model (Lavrac et al, 2004). The term fraud here refers to the abuse of a profit organization's system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the legitimacy of and

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms.

Evolved from numerous research communities, especially those from developed countries, the analytical engine within these solutions and software are driven by artificial immune systems, artificial intelligence, auditing, database, distributed and parallel computing, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, neural networks, pattern recognition, statistics, visualization and others. There are plenty of specialized fraud detection solutions and software1 which protect businesses such as credit card, e-commerce, insurance, retail, telecommunications industries.[4]

IV. BENEFITS OF DATA MINING TECHNIQUES [3]

- Problems with large databases may contain valuable implicit regularities that can be discovered automatically.
- Difficult-to-program applications, which are too difficult for traditional manual programming.
- Software applications that customize to the individual user's preferences, such as personalized advertising

.

There are several reasons why data mining approaches plays a role in these three domains. First of all, for the classification of security incidents, a vast amount of data has to be analyzed containing historical data. It is difficult for human beings to find a pattern in such an enormous amount of data. Data mining, however, seems well-suited to overcome this problem and can therefore be used to discover those patterns.[3]

V. TEMPORAL STOCHASTIC ACTIVITY MODEL

The difference between text mining and data mining is based on source of data. In text mining, basically input is the unstructured file while for data mining input is of structured data. That means patterns are extracted from unstructured text in text mining while in data mining, structured data is used.

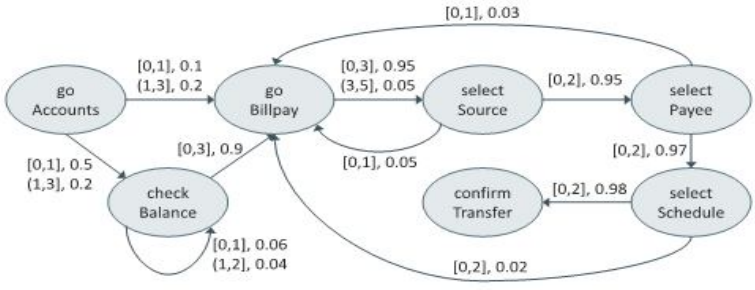

Figure 1: Example of temporal stochastic activity

Fig. 1 shows an example temporal stochastic activity modeling a bill payment process in an online banking system. A user will first access her accounts page (goAccounts) and either check her balance (checkBalance) or continue directly to the bill payment page (goBillpay). Assuming a time granularity of minutes, the edges between goAccounts and its successors are interpreted as in Example 2.1, e.g., there is a 0.5 probability that the checkBalance observation will occur in less than1 minute, and a 0.2 probability that it will occur in 1-3 minutes. The rest of the activity requires users to select an account (selectSource), choose a payee (selectPayee), schedule the amount and date of payment (selectSchedule), and finally confirm the transfer (confirmTransfer).At each stage of the process, a user can cancel the sequence and return to the bill payment page.

Also, the data mining problem must be well-defined, cannot be solved query and reporting tools, and guided by a data mining process model (Lavrac et al, 2004).

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

The term fraud here refers to the abuse of a profit organization's system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms.[2]

VI. IDENTIFICATION AND EVIDENCE PROBLEM

This section formalizes the Evidence and Identification problems. Without loss of generality, we assume that observations are stored in a single relational observation table, denoted D. Each tuple t 2 D corresponds to a single observation, denoted t:obs, which is observed at a given time, denoted t:ts. When our framework is used for real-time activity detection, our proposed insertion algorithm (which will be described in Section 4.1) processes each observation as it is received, updates the index, and stores a tuple in the observation table. Conversely, when the framework is used to detect activities in a previously acquired body of data, our bulk insertion algorithm can pull all the observation tuples from the table and build the whole index [2].

Additionally, in some applications, each observation may be associated with context information (e.g., IP address, full name, spatial location), which might help discriminate between observations belonging to different activity occurrences. However, we do not assume this information to be available in general. For instance, in an intrusion detection system, multiple attackers engaged in different activities, may need to perform some common steps, and they may appear to come from the same origin if they use proxies to conceal their real identities. We use t: context to denote context information for observation tuple t, and propose a restriction where two tuples are considered to be part of the same activity occurrence only if their context information is "equivalent." Note that t: context can generally be used to represent the result of the evaluation of a given predicate on t.

VII. TEMPORAL MULTIACTIVITY GRAPH INDEX

Temporal multiactivity graph index creation abbreviated as tMAGIC. In order to monitor an observation table for occurrences of multiple activities, we first merge all temporal activity definitions from into a single graph. We use to denote a unique identifier for activity A and IA to denote the set[2].

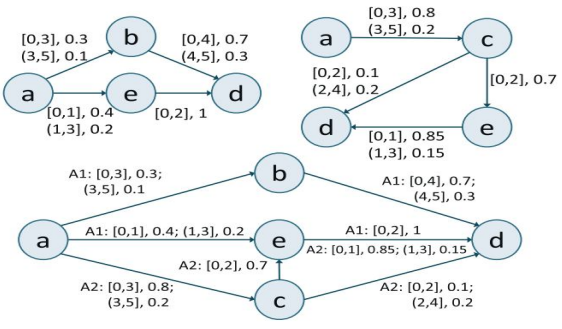

Figure 2: Temporal stochastic activities (top) and corresponding multiactivity graph (bottom)

A temporal multiactivity graph merges a number of stochastic activities. It can be graphically represented by labeling nodes with observations and edges with the ids of activities containing them, along with the corresponding timespan distributions. The temporal multiactivity graph can be computed in time polynomial in the size of A. Furthermore, the temporal multiactivity graph has to be computed only once before building the index. Fig. 5 shows two temporal stochastic activities and the corresponding multiactivity graph.

It is not feasible to find all activity occurrences each observation may be considered as being connected to many occurrences. However, in the real world, each tuple could have been generated by only one activity. Finding all the

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

identifiable activity occurrences is therefore both infeasible and undesirable, as it would lead to a number of identified occurrences much greater than the actual number of occurrences in the observation table. Hence, we define reasonable restrictions on what constitutes a valid occurrence in order to reduce the number of possible occurrences. We propose three restrictions minimal span, maximal probability, and earliest action applicable in most real-world scenarios. We do not claim these restrictions to be exhaustive: many others could be easily defined, depending on the application's needs, and added to our framework. Moreover, we will show that the most significant complexity reduction in our framework is achieved by introducing a pruning strategy that leverages.

Improving Time and Space Performance
We now propose two pruning strategies that improve the time and space performance of the tMAGIC index while guaranteeing the correctness of the results. The first strategy is called Time Frame pruning. It is based on the observation that the number of "recent" records in a tMAGIC index, i.e., those records whose corresponding observations still have a chance of being linked to a new one, is basically independent of the size of the observation table [2].

Time Frame Pruning:
Pruning strategy avoids scanning the entire predecessor table when most of the records in tables cannot be linked to the other table because too long has passed since their corresponding observations were made, causing the overall probability to be zero. The following propositions ensure that the strategy is correct and analyses the resulting time complexity [2].

VIII. MISUSE AND ANOMALY DETECTION USING DATA MINING TECHNIQUES[1]

A. Misuse Detection Using Supervised Learning**:**
Misuse detection methods, a model based supervised method make use of a classifier that has to be trained with labelled patterns [7]. The training patterns are labelled as normal or attacks. After the classifier is trained, it can classify or label new unlabeled patterns. These methods are also able to detect previously known attacks with good accuracy but also have some disadvantages. They are unable to detect new emerging threats and the labelling procedure of the training data is expensive and time consuming.

B. Anomaly Detection Using Supervised Learning**:**
The supervised anomaly detection approach train a classifier with pure "normal" labelled patterns. Anomalies (a subset of which is attacks) are detected as significant deviations from this model of normal behaviour. The arguments for this approach are that normal data is far easier to come by than are labelled attacks that a pure anomaly detector is unbiased towards any set of pre-trained attacks, and, therefore, it may be capable of detecting completely novel attacks. The counter arguments are that hostile activities which appear similar to normal behaviour are likely to go undetected, that it fails to exploit prior knowledge about a great many known attacks, and that, to date, false alarm rates for pure anomaly detection systems remain unusable high.

C. Misuse Detection Using Unsupervised Learning**:**
As is known unsupervised learning is based not on the predefined training data set misuse detection is done mostly by using supervised learning and the unsupervised learning is not been preferred for misuse detection .

IX. APPLICATIONS

Data mining is a most popular technology used for extracting data from huge amount of database. And by using data mining activity detection gets boots for its output following are the applications activity detection-

1) Security applications: Used for monitoring and analyzing activities transitions while buying something from ecommerce website.[5]
2) Bank application: every bank need to monitor every transaction done by users of bank it may be internet banking or ATM machines.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

3) Customer Relationship Management (CRM):Text mining is also useful in Customer Relationship Management (CRM) for supplying immediate answers to frequently asked questions
4) Activity detection is also used in following sectors [6]-
   a. Monitoring media tweets activities or news activities etc.
   b. Telecommunications, energy and other services industries.
   c. Information technology sector and Internet.
   d. Banks, insurance and financial markets.
   e. Pharmaceutical and research companies and healthcare.

X. MERITS

- Monitoring activities gets improvement in business of ecommerce and manage anomaly users to detect and protect system from such a unauthorized tasks.
- As data is growing rapidly in any organization, so it not possible to store that data in database due to its size limitation. So most of organization stores data in the form of text. Text mining is applied on that data for pattern extraction. [5]

XI. DEMERITS

- Data collection requires handling a lot of unstructured text in Data-mining.
- The use of natural language texts contains ambiguities and requires human intervention.
- To analyze unstructured text, there is no any program available that handle this text for text mining.

XII. CONCLUSION

This paper studies the problem of automatically and efficiently detecting activities in very large observation databases collected by systems such as web servers, banks and security installations.

We proposed temporal stochastic automata to model activities of interest and defined a data structure, called a temporal multi activity graph, to merge multiple activity graphs together and enable concurrent monitoring of multiple activities. We introduced the temporal multi activity graph index to index very large numbers of temporal observations from interleaved activities.

REFERENCES

[1] [1] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V.S. Subrahmanian, P. Turaga, And O. Udrea, A Constrained Probabilistic Petri Net Framework For Human Activity Detection In Video, Ieee Trans. Multimedia, Vol. 10, No. 8, Pp. 1429-1443,Dec. 2008.
[2] [2] Massimiliano Albanese, Andrea Pugliese, And V.S. Subrahmanian, Fast Activity Detection: Indexing For Temporal Stochastic Automaton-Based Activity Models , Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 2, February 2013
[3] [3] Clifton Phua,Vincent Lee, Kate Smith & Ross Gayler A Comprehensive Survey Of Data Mining-Based Fraud Detection Research .
[4] [4] R. Chellappa, N. P. Cuntoor, S. W. Joo, V. S. Subrahmanian, And P. Turaga, Understanding Events: How Humans See, Represent, And Act On Events. Oxford University Press, January 2008, Ch. Computational Vision Approaches To Event Modeling.
[5] [5] M. Albanese, V. Moscato, A. Picariello, V.S. Subrahmanian, And O.Udrea, Detecting Stochastically Scheduled Activities In Video, Proc. 20th Int L Joint Conf. Artifical Intelligence (Ijcai 07),Pp. 1802-1807, Jan. 2007.
[6] [6] Www.Ise.Bgu/~Hanj/Pdf
[7] En.Wikipedia.Org/Wiki/Data_Mining.

BIOGRAPHY

Santosh S. Gurav is a Student of computer engineering Department ,JSPM's Imperial college of engineering & research(ICOER),Wagholi, Pune.

Prof. S. R. Todmal is a Professor in Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune.