Adult Daily Life Analysis by Detecting Outliers and Top-K Queries

K. Ashesh
Research Scholar
Department of CSE
Gitam University
Visakhapatnam

G. Appa Rao, PhD
Professor
Department of CSE
Gitam University
Visakhapatnam

ABSTRACT
Detecting outliers has attracted the attention of researchers ever since it was found useful to uncover latent, unexpected, interested and previously unknown knowledge. There are many real world utilities of outlier detection such as change monitoring, rare event discovery, fraud detection, and event change detection, alarm systems, revealing trend occurrences and finding strange patterns. Time series analysis can help detect outliers and discover knowhow for efficient decision making. Outlier detection mechanisms that came into existence dealt with plethora of problems. In this paper, our focus is on analyzing strange behaviours in activities of daily living. We proposed two approaches towards adult daily life analysis. Both are data mining approaches that extract knowledge from underlying dataset. Top-K probabilistic analysis and outlier detection are the two mechanisms used to analyze adult life. The top-k analysis brings about the prevalent behavioural dynamics in adult life while the outlier detection throws light into observations that are peculiar and do not match with other observations. We used a time series dataset named Activities of Daily Living (ADLs) collected from UCI machine learning repository. We built a prototype to demonstrate the proof of concept. The empirical results are encouraging.

Keywords
Data mining, outlier detection, top-k queries, adult life analysis

1. INTRODUCTION
Detecting outliers is an important activity in data mining. Outlier is some item in a dataset which is very distant from other items [1]. In other words it is an observation which is far distant and peculiar or strange or abnormal when compared with other observations. Outlier detection has plethora of real world applications. They include change detection, studying abnormal behaviours, rare event discovery, fraud detection, event change detection, alarm systems, revealing trend occurrences and so on [5], [6]. The list is not exhaustive as outlier detection can be used in every domain to discover knowledge that can help in making well informed decisions.

Many research papers came into existence in exploring outlier detection mechanisms. Rasheed and Alhaji [27] proposed a framework that can detect periodic outlier patterns in time series data. Nouira and Trabelsi [23] proposed a mechanism that can detect outliers in intensive care data. Principal Component Analysis is used for outlier detection in [19]. Different approaches were explored in [17], [18], [20], [21], and [22] towards analyzing the benefits of outlier detection in the real world. DBSCAN and many other data mining approaches were explored in the literature. More details are provided in related works section.

Our contributions in this paper are as follows. We proposed an algorithm for top-k queries that will bring about hidden knowhow on underlying datasets. The algorithm provides top-k patterns. We also proposed outlier detection algorithm that throws light into the observations that are abnormally distant from other similar observations. Activities of daily living dataset from UCI [34] are considered for experiments. We built a prototype that demonstrates the proof concept. The empirical results are encouraging. The remainder of the paper is structured as follows. Section 2 reviews relevant literature on related works. Section 3 presents the proposed approaches for adult life analysis. Section 4 presents prototype implementation. Section 5 throws light into experiential results while section 6 concludes the paper besides giving directions for future work.

2. RELATED WORKS
This section provides review of literature that throws light into previous works in the similar work. Takeuchi et al. [25] proposed a unified framework that has provision for outlier detection and also change point detection from time series dataset. The two stage process proposed by them combines both outlier detection and change point detection. The change point detection outcome is given as input to the outlier detection. Different mechanisms of outlier detection are explored in [30] and [32]. Nouira and Trabelsi [23] proposed a mechanism that can detect outliers in intensive care data. Gibbs sampling approach is used in order to detect outliers. These outliers can be used in intelligent alarm systems. Zhi-zhong [24] explored machine learning techniques for outlier detection. Their solution was able to detect outliers efficiently. Similar to the approach explored in [25], Li et al., [25] also used a unified approach that contains two aspects such as change point detection and outlier detection. However, they used fuzzy approaches for performance improvement. Tang et al. [26] proposed a novel mechanism for detecting outliers in load curve data. This mechanism analyzed periodic patterns and used for cleaning load curve data besides detecting outliers.

Rasheed and Alhaji [27] proposed a framework that can detect periodic outlier patterns in time series data. Suffix tree based algorithm is used in order to achieve this. Projective clustering approach for outlier detection [1], Modified Stahel Dohono [2] for multivariate outlier detection, outlier detection based on distribution of data using univariate dataset [3], detection of multivariate outliers on survey data [4], heart rate variability analysis using outlier detection [7], DBSCAN for exploring outlier detection [8], outlier detection on temporal data [9], survey of outlier detection methodologies [10], [11], fast outlier detection [15], structural outlier detection [14], applications of outlier detection and techniques [13], [16] are some of the researches that focused on outlier detection.

Geo and Tan [12] explored the conversion of outputs of outlier detection algorithms into probability estimates. Principal Component Analysis is used for outlier detection in [19]. Different approaches were explored in [17], [18], [20], [21], and [22] towards analyzing the benefits of outlier detection in the real world. Latent outlier detection problem is explored in [33] besides low precision problem. Jury based grading systems are explored in [31] for outlier detection. Snake validation and principal component analysis were explored in [29]. This paper explores the combination of both outlier detection and top-k queries for analyzing activities of daily living.

3. PROPOED APPROACH

In this paper we proposed and implemented a mechanism that combines outlier detection and also top-k queries to analyze activities of daily living. The activities of daily living dataset is taken from UCI machine learning repository [34]. It is a time series data which has various house hold events or the actions of adults in daily living captured by sensors. The dataset is essentially a time-series dataset which is explored in this paper to find outliers besides making use of top-k analysis. The aim of the proposed system is adult daily life analysis using outlier detection and top-k queries on time serried data.

Algorithm: Adaptive Top-K Algorithm

Purpose: To process top – k queries on dataset generated by sensors

Inputs : dataset, top-k query

Outputs: Top – K results

STEP 1: PRE-PROCESSING

Load given dataset into a reader object

Write the dataset content to a relational table

STEP 2: IDENTIFY PROBABLE TOP K COLUMN

For Each Column in Dataset

IF column is top k probable

Choose column as Candidate for Top K Processing

END IF

END

Initialize F to hold unique field values

For all values in chosen Column

Add unique value to ***F***

END

STEP 3: Processing Top K Queries

Initialize ***R*** for holding top K Results

Populate F on UI

Take Top K Input from End User

Extract all rows from dataset that satisfy user selection

Compute ranking for rows

Sort rows in ascending order by rank

Populate top K rows into ***R***

Return ***R***

Figure 1 – Proposed algorithm

As can be seen in Figure 1, it is evident that the algorithm operates on ADL dataset which is a time series dataset that captures adult life with respect to various household events and the behaviour of humans. The top-k analysis provides the important tuples that reflect the adult life dynamics. From the results, the outlier detection becomes easier as the outliers are nothing but abnormal observations that are very distant from other observations.

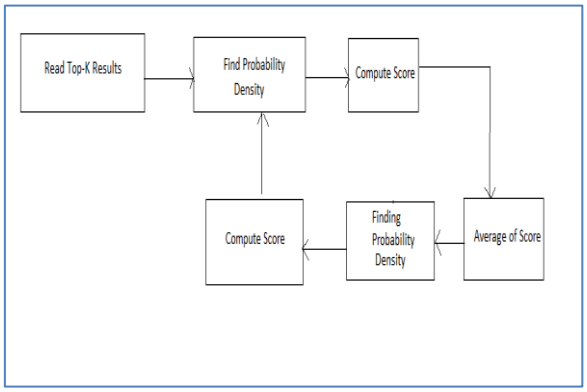


Figure 2 – Mechanism for outlier detection

As shown in Figure 2, it is evident that the top-k results are being used in the process of outlier detection. The top-k results are useful in order to reduce the search space and also speed up the outlier detection process. The mechanism used here finds probability and then computes score. This process is made iteratively every time finding average score. Finally the outliers are detected and the results are presented in the following section.

4. EXPERIMENTS AND RESULTS

4.1. Data Set Used for Experiments

The dataset used for experiments in this paper is known as Activities of Daily Living (ADL). The dataset is time series dataset which is obtained from UCI machine learning repository [34]. The dataset contains activities of daily living and it reflects adult life dynamics in households. It has 2747 instances. The data reflects adult life in two households for 35 days. In this paper, this time-series dataset is used in order to detect outliers. It is also used to analyze adult life using top-k queries.

Start time	End time	Activity
2011-11-28 02:27:59	2011-11-28 10:18:11	Sleeping
2011-11-28 10:21:24	2011-11-28 10:23:36	Toileting
2011-11-28 10:25:44	2011-11-28 10:33:00	Showering
2011-11-28 10:34:23	2011-11-28 10:43:00	Breakfast
2011-11-28 10:49:48	2011-11-28 10:51:13	Grooming
2011-11-28 10:51:41	2011-11-28 13:05:07	Spare_Time/TV
2011-11-28 13:06:04	2011-11-28 13:06:31	Toileting
2011-11-28 13:09:31	2011-11-28 13:29:09	Leaving
2011-11-28 13:38:40	2011-11-28 14:21:40	Spare_Time/TV
2011-11-28 14:22:38	2011-11-28 14:27:07	Toileting
2011-11-28 14:27:11	2011-11-28 15:04:00	Lunch
2011-11-28 15:04:59	2011-11-28 15:06:29	Grooming
2011-11-28 15:07:01	2011-11-28 20:20:00	Spare_Time/TV
2011-11-28 20:20:55	2011-11-28 20:20:59	Snack
2011-11-28 20:21:15	2011-11-29 02:06:00	Spare_Time/TV
2011-11-29 02:16:00	2011-11-29 11:31:00	Sleeping
2011-11-29 11:31:55	2011-11-29 11:36:55	Toileting
2011-11-29 11:37:38	2011-11-29 11:48:54	Grooming
2011-11-29 11:49:57	2011-11-29 11:51:13	Showering
2011-11-29 12:08:28	2011-11-29 12:18:00	Breakfast
2011-11-29 16:34:17	2011-11-29 17:08:07	Spare_Time/TV

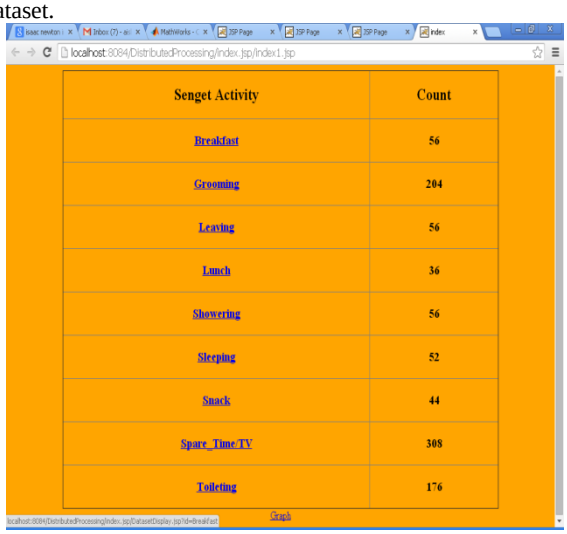
Figure 3 – An excerpt from ADL dataset

As shown in Figure 3, it is evident that the dataset has time series data set collected from UCI machine learning repository

[34]. This dataset is used for analyzing adult life using outlier detection and top-k queries.

4.2. Prototype Implementation

In order to explore outlier detection mechanism proposed by us and analyze the adult life using top-k queries, we built a prototype using Java platform. The proposed mechanisms are applied to the time-series dataset which captures adult daily life. The dataset contains details pertaining to various adult activities and the time dynamics of the same. The aim of this paper is to detect outliers and also analyze adult life using top-k queries. The methodology used is described here. First of all the dataset is loaded into the application and then the data is subjected to top-k queries. The top-k queries can provide useful adult life dynamics. Then the results of top-k are given as input for outlier detection. Thus the quality of outlier detection gets improved. The algorithm for top-k analysis and the mechanism for outlier detection are applied on the chosen dataset.

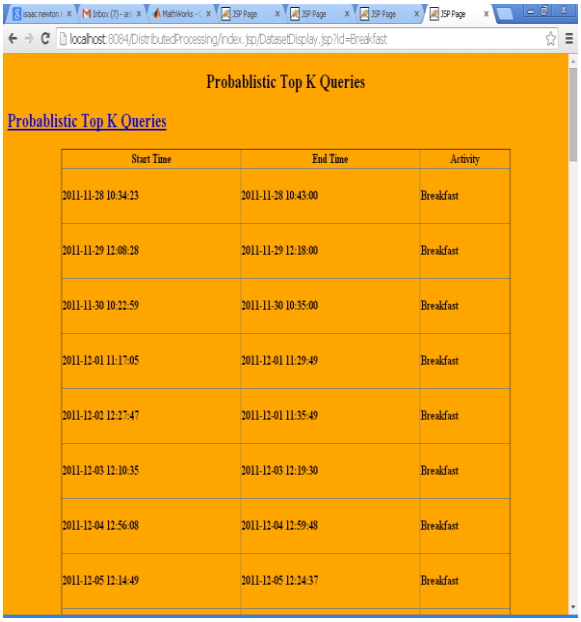


Senget Activity	Count
Breakfast	56
Grooming	204
Leaving	56
Lunch	36
Showering	56
Sleeping	52
Snack	44
Spare_Time/TV	308
Toileting	176

Graph

Figure 4 – Adult activities and the frequency

As shown in Figure 4, it is evident that the algorithm produces top-k events and the corresponding frequencies. The results are intermediary results and finally the algorithm produces top-k results based on the event that has been selected.



Probablistic Top K Queries

Probablistic Top K Queries

Start Time	End Time	Activity
2011-11-28 10:34:23	2011-11-28 10:43:00	Breakfast
2011-11-29 12:08:28	2011-11-29 12:18:00	Breakfast
2011-11-30 10:22:59	2011-11-30 10:35:00	Breakfast
2011-12-01 11:17:05	2011-12-01 11:29:49	Breakfast
2011-12-02 12:27:47	2011-12-01 11:35:49	Breakfast
2011-12-03 12:10:35	2011-12-03 12:19:30	Breakfast
2011-12-04 12:56:08	2011-12-04 12:59:48	Breakfast
2011-12-05 12:14:49	2011-12-05 12:24:37	Breakfast

Figure 5 – Shows top-k results for the "Breakfast" event

As show in Figure 5, the selected event is taken and the final results are produced by the algorithm. This will help analyze different top-k events reflected in the activities of adult life dataset.

4.3. Results

The prototype we built is used to make experiments. The top-k analysis and outlier detection are the combined activities as presented in the methodology of the proposed work. Three different experiments are made and the results are presented here.

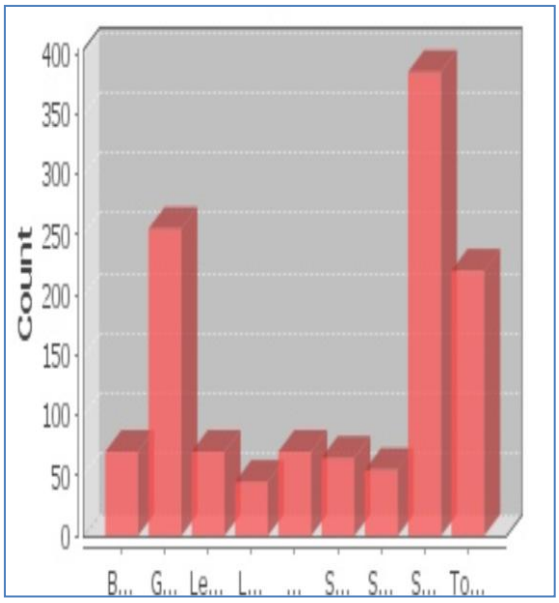


Figure 6 – Results of Experiment 1

As seen in Figure 6, the results provide the results of combined method that is the combination of top-k analysis and also outlier detection. The results reveal the dynamics of adult life and the events that are far from other events and that are treated as outliers.

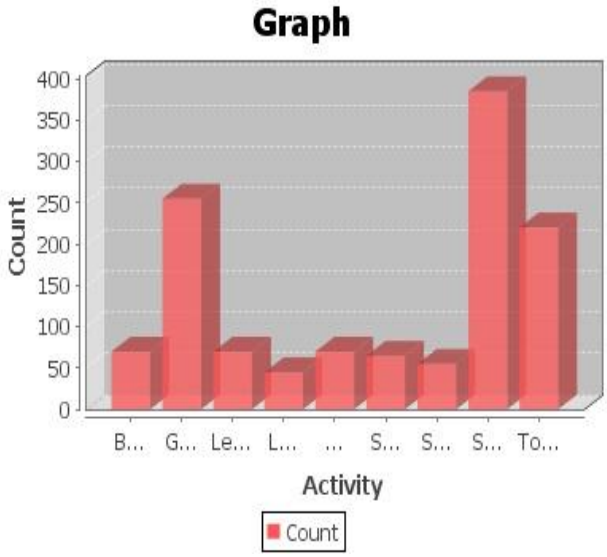


Figure 7 – Results of Experiment 2

As seen in Figure 6, the results provide the results of combined method that is the combination of top-k analysis and also outlier detection. The results reveal the dynamics of adult life and the events that are far from other events and that are treated as outliers.

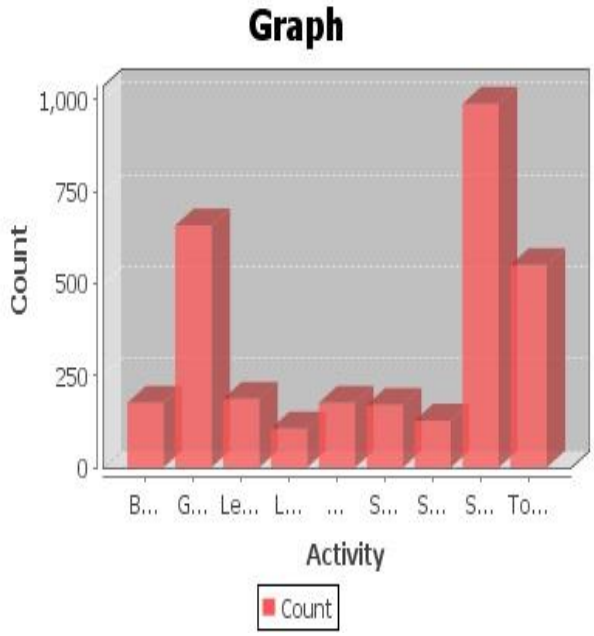


Figure 8 – Results of Experiment 1

As seen in Figure 8, the results provide the results of combined method that is the combination of top-k analysis and also outlier detection. The results reveal the dynamics of adult life and the events that are far from other events and that are treated as outliers.

5. CONCLUSIONS AND FUTURE WORK

In this paper we studied the outlier detection mechanisms that are applied in different domains. Outlier detection has many advantages in real world applications including anomaly detection, rare event discovery, fraud detection, change detection and so on. Many approaches came into existence for detecting outliers in the data mining domain. They are used to solve different problems in real world applications. In this paper we proposed a combined approach that includes outlier detection and top-k analysis in order to analyze adult daily life. Adult daily life dataset is considered for experiments. The dataset is obtained from UCI machine learning repository. We built a prototype to demonstrate the proof of concept. The knowhow obtained through outlier can provide business intelligence that can be used to make well informed decisions. Experiments are made with the prototype application using the ADL dataset. The empirical results are encouraging.

6. REFERENCES

[1] R.Parimaladev. (2012). PROJECTIVE CLUSTERING FOR OUTLIER DETECTION IN HIGH DIMENSIONAL DATASET. *ISSN*. 1 (9), p.23-30.

[2] WADA, Kazumi. (2012). Parallel computation of modified Stahel–Donoho estimators for multivariate outlier detection. *detection*.p.44-55.

[3] Mark j.p. (2010). Distribution based outlier detection in univariate.*IJSIM*.p.56-63.

[4] Valentin Todorov. (2004). Detection of Multivariate Outliers in Business Survey Data with Incomplete Information. *Nations Industrial Development Organization*.p.66-77.

[5] Tan, Steinbach, Kumar. (2008). Data Mining Anomaly Detection. *Tan, Steinbach, Kumar*.p.88-99.

[6] Sanjay Chawla and Pei Sun. (2006). Outlier Detection: Principles, Techniques and Applications. *pakkd*.p.56-70.

[7] L Y Ji, Y J Yang. (1996). Robust Time Series Processing for Heart Rate Variability Analysis in Daily Life. *ISSN*. p.77-88.

[8] Samson Sifael Kiware. (2009). Detection of Outliers in TimeSeriesData. *detection*.p.66-78.

[9] Manish Gupta. (2014). Outlier Detection for Temporal Data: A Survey.*IEEE*. 25 (4), p.77-88.

[10] Victoria J. Hodge. (2004). A Survey of Outlier Detection Methodologies.*(vicky@cs.york.ac.uk)*. p.88-98.

[11] Victoria J. Hodge. (2004). Survey of Outlier Detection Methodologies.*victort*. p.23-30.

[12] Jing Gao. (2005). Converting Output Scores from Outlier Detection Algorithms into Probability Estimates. *gaojing2@msu.edu*. p.56-63.

[13] Hans-Peter Krieg. (1949). Outlier Detection Techniques. *IJSIM*. p.56-70.

[14] Meike Klettke. (2013). Schema Extraction and Structural Outlier Detection for JSON-based NoSQL Data Stores. *detection*.p.77-88.

[15] Spiros Papadimitriou. (2003). LOCI: Fast Outlier Detection Using the Local Correlation Integral. *IEEE*. 5 (8), p.23-30.

[16] Karanjit Singh. (2012). Outlier Detection: Applications And Techniques.*IJSIM*. 9 (3), p.44-55.

[17] Irad Ben-Gal. (2004). OUTLIER DETECTION. *detection*. p.77-88.

[18] Varun Chandola. (2006). Outlier Detection : A Survey. *detection*. p.66-78.

[19] Xin He,. (2007). An Outlier Detection Based Approach for PCB Testing.*colostae*. p.88-99.

[20] CHARU C. AGGARWAL. (2006). OUTLIER ANALYSIS. *IBMT*. p.54-66.

[21] Thomas Seidl. (2011). Statistical selection of relevant subspace projections for outlier ranking. *IJSIM*. p.78-99.

[22] Jun-ichi Takeuchi. (2006). A Unifying Framework for Detecting Outliers and Change Points from Time Series. *IEEE*. 18 (4), p.44-55.

[23] Kaouther Nouira. (2006). Time Series Analysis and Outlier Detection in Intensive Care Data. *IEEE*.p.34-45.

[24] Liu Fang. (2011). An online outlier detection method for process control time series. *IEEE*. p.45-55.

[25] Zhi Li. (2006). A Unifying Method for Outlier and Change Detection from Data Streams. *IEEE*.p.66-75.

[26] Guoming Tang. (2014). From Landscape to Portrait: A New Approach for Outlier Detection in Load Curve Data. *IEEE*. 5 (4), p.45-60.

[27] Faraz Rasheed and Reda Alhajj. (2014). A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences. *IEEE*. 44 (5), p.44-55.

[28] Hancong Liu. (2004). On-line outlier detection and data cleaning.*elsver*. 25 (1), p.1635–1647.

[29] Baidya Nath Saha,. (2009). Snake Validation: A PCA-Based Outlier Detection Method. *IEEE*. 16 (6), p.77-88.

[30] Joanne B.; Ciulla. (2013). Introduction to Outlier Detection. *(www.biomedware.com)*.p.44-55.

[31] Mary Kathryn Thompson. (2013). Statistical Outlier Detection for Jury Based Grading Systems. *Atalnta*. n.d (n.d), p.55-66.

[32] Hans-Peter Kriegel. (2010). Outlier Detection Techniques.*http://www.dbs.ifi.lmu.de*. p.77-88.

[33] Fei Wang. (2013). Latent Outlier Detection and the Low Precision Problem. *IJSIM*. p.34-45.

[34] Kevin Bache. (2007). Machine Learning Repository. UCI. p.12-19. Retrieved from https://archive.ics.uci.edu/ml/datasets.html.