Data Slice Search for Local Outlier View Detection: A Case Study in Fashion EC

Takumi Matsumoto, Yuya Sasaki, Makoto Onizuka
{matsumoto.takumi,sasaki,onizuka}@ist.osaka-u.ac.jp
Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

ABSTRACT

The exploratory data analysis is one of the current research trends for business data analysis, by which we can identify interesting results by executing a large number of OLAP queries. The queries are generated by changing the analytical viewpoints (aggregation attribute or group-by attribute) and/or target data slices (data subset extracted by the select operation). Existing research effectively detects globally unexpected trends (global outliers), however, it cannot detect locally unexpected trends (local outliers), which are known useful in many applications. In this paper, we describe an analysis framework named D4C. D4C detects top-n data slices that generate local outlier results of automatically generated OLAP queries. We also introduce how to use the analysis results in a practical use case of fashion EC. Since there are various types of users and items at fashion EC site, it is useful to investigate the bias of the sales trend and identify meaningful results. We also show how the analysis results help making decisions on sales strategies.

KEYWORDS

Exploratory data analysis, OLAP, Outlier detection

1 INTRODUCTION

Many companies have collected or accumulated enormous and diversified data. Data analysts start their analysis work by extracting useful information (insight and exceptional data) from collected and accumulated data for decision making to make social or economic impact. Generally, in a business data analysis work, OLAP (online analytical processing) technology is frequently used with visualization tools, such as Tableau [1, 19]. The analysis workflow consists of two steps, (1) issuing a query that specifies an analysis pattern and a target data slice (data subset extracted by the select operation), and (2) investigating the query result (view). An analysis pattern is expressed with a combination of group-by attribute, measure attribute, and aggregate function. A target data slice is specified by WHERE clause. However, the analysis work is heavy burden for the analyst because the number of OLAP queries increases as the cardinality of data and the number of columns increase, and the number of times the analyst repeatedly performs (1) and (2) accordingly increases.

For solving the above problem, the exploratory data analysis is promising research area [7, 8, 13, 17, 18, 20, 22]. In these studies, analysis axes and data slices that generate exceptional query results are automatically identified so that analysts can easily find interesting views. They are categorized into two types of dual data analysis over OLAP queries, data slice search and query search. Let D be a database for analysis. In the query search case, given a data slice S of D, we identify query q among various OLAP queries that maximizes the deviation between $q(D)$ and $q(S)$. In contrast, in the data slice search case, given OLAP query q, we identify data slice S among various data slices that maximizes the deviation between $q(D)$ and $q(S)$. So, they are effective to detect globally unexpected trends (global outliers) by computing the distance between multiple query results, however, they cannot detect locally unexpected trends (local outliers). The local outlier factor [2] is a well-known concept in many applications areas, such as in fraud detection by detecting unusual usage of credit cards, in customized marketing for identifying the unexpected behavior of customers, or in medical analysis for finding unusual responses to various medical treatments [6].

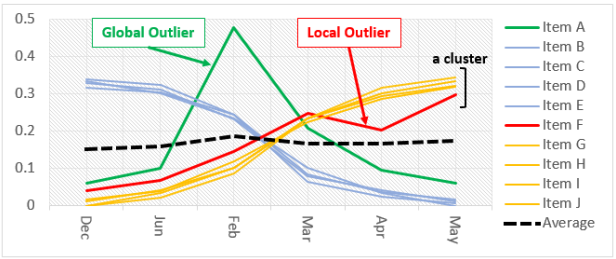

Figure 1: Monthly sales of items: X axis and Y axis indicate months and normalized sales sum, respectively. There are two clusters (yellow lines and blue lines). Item A (green line) is a global outlier and Item F (red line) is a local outlier.

In this paper, we describe D4C (Dimensionally Deviated Divisional Data Captor), a framework for automatically identifying top-n local outliers among OLAP query results. D4C automatically generates OLAP queries from a query template specified by users, executes those queries, and then identifies unexpected trends by computing LOF value for each query result.

As an application example of D4C, consider a case that a sales strategy is decided based on the sales trend over the last six months. Fig. 1 shows the sales trend of ten items (data slices), A to J. Each line represents the monthly sales of each item, and the black dash line represents the average of monthly sales of the whole items. This example has two clusters (1) items sold well during winter (A, B, C, D, and E) and (2) items sold well during spring (F, G, H, I and J). Items A is global outliers because they are deviated largely from the average. Notice that item F (red line) has the smallest global outlier factor among the items, however it has the largest local outlier factor among the items that are mainly sold in spring season. Therefore, it may be possible to increase the sales of item F by investigating reasons why item F is a local outlier and by changing the sales strategy according to the investigation (e.g., a discount sale should be made for item F in April, since its sales is deviated down from the yellow cluster in April).

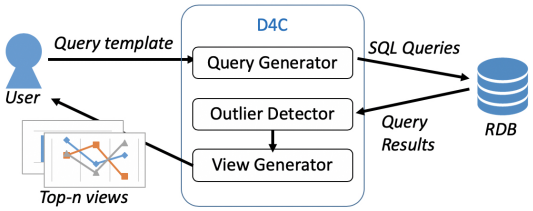

Figure 2: D4C Architecture

We apply the D4C to the sales data provided from one of the largest fashion EC site in Japan. The data contains transactions from April 2015 to March 2016 at the site. The analysis results give us interesting observations and helps making decisions on sales strategies.

Organization This paper is organized as follows. We explain local outlier factor as preliminaries in Section 2. Then, we describe outlier detection for data slices in Section 3 We give analysis results for a sales data of fashion EC site by applying D4C in Section 4. We describe related work in Section 5 and conclude this paper in Section 6.

2 LOCAL OUTLIER FACTOR

As preliminary, we introduce an outlier detection technique, local outlier factor (LOF) [2]. LOF is based on the idea of local density in multi-dimensional space. For each data point, we can compute LOF value that indicates the outlierness among its nearest neighbors. Intuitively, the LOF value of data point A is high if its local density is low and those of A's nearest neighbors are high. Consider data point A, which is represented by a pair of N positive real numbers a_i $(1 \leq i \leq N)$:

$$A := [a_1, a_2, \cdots, a_N] \tag{1}$$

We denote $N_k(A)$, the set of k or more nearest neighbors of A, which is defined as follows:

$$N_k(A) := \{B \in \mathbb{P} - \{A\} \mid d(A,B) \leq k\text{-}distance(A)\} \tag{2}$$

where $\mathbb{P}$ is a set of data points, $d(A,B)$ is the Euclidean distance between two data points A and B, $k\text{-}distance(A)$ is the Euclidean distance between A and the kth closest data point to A. The LOF value of A is defined as follows:

$$LOF(A) := \frac{\sum_{B \in N_k(A)} lrd_k(B) \ / \ |N_k(A)|}{lrd_k(A)} \tag{3}$$

That is, $LOF(A)$ is the ratio of the average density of A's nearest neighbors $(B \in N_k(A))$ to A's density. The density of A, $lrd_k(A)$, is the inverse of the average reachable distance from A's k nearest neighbors to A, which is defined as follows:

$$lrd_k(A) := \frac{|N_k(A)|}{\sum_{B \in N_k(A)} reach\text{-}dist_k(A,B)} \tag{4}$$

where $reach\text{-}dist_k(A,B)$ is a reachable distance from B to A defined next:

$$reach\text{-}dist_k(A,B) := \max\{d(A,B), \quad k\text{-}distance(B)\} \tag{5}$$

In Equation (5), $k\text{-}distance(B)$ is a term for reducing statistical fluctuation [2]. Additionally, the value of k should be 10 or more to remove unwanted statistical fluctuations.

3 OUTLIER DETECTION FOR DATA SLICES

We extend the technique of the exceptional view detection [13] with the notion of the local outlier factor. We describe a problem of identifying data slices that generate views with the largest LOF values for a given query template. We give our problem definition in Section 3.1. Then, in Section 3.2, we present our framework to solve the problem.

3.1 Problem definition

Let D be a set of records and C be a set of dimension attributes in a database. We define *data slice* S as a subset of D cuboid sliced by choosing a single value Y for dimension attribute $c \in C$ as follows:

$$S := \sigma_{c=Y}(D) \tag{6}$$

A query template is given by data analysts in advance. We define a set of data slices $\mathbb{S}$ and query template q as follows:

$$\mathbb{S} := \bigcup_{i=1}^{|C|} \{\sigma_{c_i=Y}(D) | Y \in values(c_i)\} \tag{7}$$

$$q(S) := {}_g G_{f(m)}(S) \tag{8}$$

where $values(c_i)$ is a set of unique values of dimension attribute $c_i \in C$, g is a dimension attribute for group-by operation, m is a measure attribute for aggregate function, and f is an aggregate function. G groups the records using g and aggregates the grouped values of m using f. Since query result $q(S)$ is a sequence of pair <unique value of g, aggregated value of m>, we introduce function t: sequence<K,V> $\rightarrow$ point($V[N]$) so that we map a query result to a data point in the N-dimentional space for computing LOF value.

Definition 1. *The problem here is to identify the top-n data slices in $\mathbb{S}$ that generate views with the largest LOF values for given query template q, defined as follows:*

$$\underset{S \in \mathbb{S}}{\text{argmax}}^{n} \ LOF(t(q(S)))$$

Example 3.1. *Remember the analysis example in Fig.1. In this case, ItemName is used as a dimension attribute, so $\mathbb{S}$ = { "Item A", "Item B", $\cdots$, "Item J" }. Since X axis and Y axis indicate months and normalized sales sum, respectively, q(S) is expressed as:*

$$q(S) = {}_{month} G_{sum(sales)}(S)$$

3.2 D4C Framework

We develop D4C on top of a relational database, a framework for automatically identifying top-n local outliers among OLAP query results. Fig. 2 depicts the D4C architecture. D4C automatically generates and executes OLAP queries from a query template specified by users, and then identifies top-n local outliers among the OLAP query results by computing LOF value each query result. The data analysis by D4C consists of three components:

Query Generator Given that users specify query template q, the query generator lists up various data slices $\mathbb{S}$, instantiate OLAP queries ($\{q(S) \mid S \in \mathbb{S}\}$) by combining each data slice S with the query template q, and then executes those queries.

Outlier Detector The outlier detector computes the LOF value for each query result and identifies top-n outliers based on the LOF values.

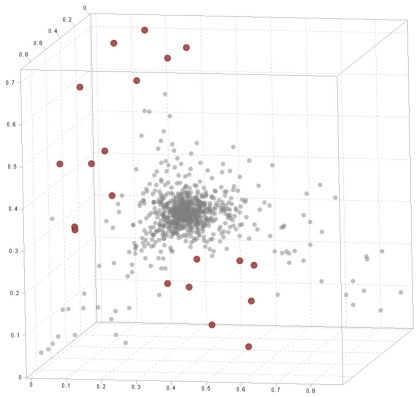

Figure 3: A visualized result in a 3-dimensional Euclidean space. Top 20 data slices with large LOF values are colored in red.

View Generator The view generator generates views of the query results identified as top-n outliers and then displays them on the system.

Query Generator

Users can generate various types of OLAP queries by changing query templates. They specify a query template, $q = {}_{g}G_{f(m)}$, (dimension attribute g, measure attribute m, and aggregate function f) and a set of dimension attribute, C, for extracting data slices.

Example 3.2. *Remember again the example in Fig. 1. OLAP queries are* $\{q(S) \mid S \in \mathbb{S}\}$. *One of the OLAP queries is expressed by an SQL statement as follows:*

```
SELECT Month, SUM(Sales)
FROM D
WHERE ItemName = 'Item A'
GROUP BY Month;
```

Outlier Detector

For each generated OLAP query for data slice S, D4C retrieves the query result from the database. This component identifies top-n local outliers by computing LOF value for each query result, $q(S)$ where $S \in \mathbb{S}$. The local outlier detection is performed for data points in a N dimensional Euclidean space (N represents the number of the values of the group-by attribute). That is, this component transforms each query result with N aggregated values to a data point in N dimensional Euclidean space, and then computes the LOF value for each data point. For instance, the number of dimensions of Euclidean space N is 12 in the case where the group-by attribute is "sales month" ("January", "February", $\cdots$, "December"). Fig. 3 depicts top-20 local outliers in the three dimensional Euclidean space for data slices whose number of values of the group-by attribute is three. From this figure, we observe that the data points with high LOF values are not away from the average position of the all data points, but are deviated from its nearest neighbors.

View Generator

This component generates views that visualize the query results with top-n highest LOF values in line/column charts. For reference, we also visualize the nearest neighbors of the outliers so that how the outliers are deviated from their nearest neighbors. In these views, the horizontal axis denotes the values of the group-by attribute and the vertical axis denotes the aggregated values by the aggregate function for the aggregate attribute. Each aggregated value is normalized as the ratio, so that the trend of each data slice can be equally compared on the same scale.

4 DATA ANALYSIS

In this section, we report the analysis results obtained by applying D4C to the sales data provided from one of the largest fashion EC site in Japan. We observe that D4C automatically finds interesting results that are no easily obtained by traditional analysis tools which requires manual labors. We also discuss how the results can help making decisions on sales strategies.

4.1 Dataset and query templates

Fashion EC dataset: We used a dataset of transactions made at a fashion EC, which is provided through Joint Association Study Group of Management Science. The data of the fashion EC contains 1, 111, 365 records obtained from April 2015 to March 2016 and its size is 5, 384 MB. Each record contains the attributes of the purchased item, the purchased user type, purchased date, discount rate, and the questionnaire result. The item type includes its sales price, category, color, brand, provider (shop), and size. The user type includes his/her sex, age, and living place (prefecture and region).

Query templates: Table 1 shows five query templates used in the experiments. The **dimensions [#]** column indicates the cardinality of **Group-by attribute** column, which is the number of unique values of **Group-by attribute**. The **data slices [#]** column indicates the cardinality of the attributes used for extracting **Data slice**. We set the number of nearest neighbors k of LOF at 10 by following the tips described in [2].

4.2 Result of Analysis

We show the analysis results and describe interesting observations that may help making decisions on sales strategies. Figs. 4, 5, 6, 7, and 8 depict the identified local outliers with their nearest neighbors. They are generated from the five query templates in Table 1. In each figure, the identified local outlier is colored by red, ten data slices in the neighbor of the identified local outlier are colored by orange, and the average of all data slices is colored by blue.

Q1: Which Category is the most Local Outlier in the average sales grouped-by Prefecture? The result is depicted in Fig. 4. We observe that "trash box" category is identified as the most local outlier among other categories. The figure shows that there are exceptional values in several prefectures. For example, "trash box" sales both in Osaka and Fukushima (16th and 4th values from the right on X axis in Fig. 4, respectively) are higher than their nearest neighbors, while in Kumamoto it is lower than its nearest neighbors. This result is explained by the fact of the annual emission of garbage per person reported by the Ministry of the Environment[1]. Osaka was ranked at the top in the emission of garbage in 2014. Fukushima was always ranked in top-5 from 2014 to 2017. In contrast, Kumamoto was ranked in very low level. Thus, the outlier detection reveals some hidden knowledge from the given dataset. Another observation we found is that the sales of "trash box" in Gunma and Niigata are relatively low, although these prefectures were ranked in top-10 of the emission of garbage per person. This investigation implies that there is a

[1] https://www.env.go.jp/recycle/waste_tech/ippan

Table 1: Query templates

Query Pattern	Group-by attribute	Aggregate function / attribute	Dimensions [#]	data slices [#]
Q1	Prefecture	AVG / Sales Price	47	226
Q2	Prefecture	COUNT / Order	47	226
Q3	Purchased date	SUM / Sales Price	12	226
Q4	Age	SUM / Sales Price	6	226
Q5	Region	COUNT / Order	8	226

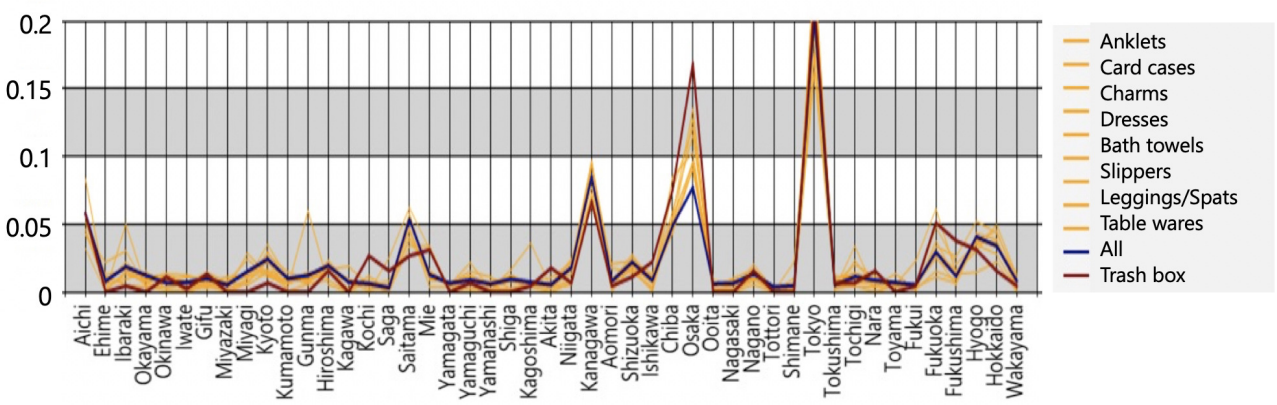

Figure 4: local outlier "trash box" in the average sales grouped by 47 prefectures

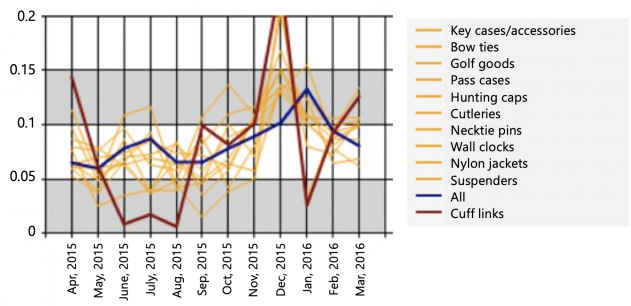

Figure 5: local outlier "cuff links" in the sum of the sales grouped by months

potential demand of "trash box" for customers in those prefectures: we can leverage it to improve the profit of the sales. Also, it would increase the profit if we sell a pair of "trash box" and its nearest neighbors, such as bath towels and slippers, because their sale treads are close each other.

Q2: Which Category is the most Local Outlier in the Sum of the Sales grouped-by Month? The result is depicted in Fig. 5. This figure shows the yearly sales trend of men's accessories (yellow lines), such as suspenders and necktie pins. "cuff links" (red line) is identified as the most local outlier, which shows a different trend from other categories. In particular, the sales from June to August are remarkably low and they are high in December, March, and April. The reason is that, since it is summer from June to August in Japan and the temperature is high, we rarely wear long-sleeved shirts and thus we do not need cuff links. In other seasons, we wear them. Another observation is as follows. In Japan, we have a custom to send gifts that relate to suits (bow ties, necktie pins, suspenders, cuff links) to new business persons in March/April or give accessories (key cases/accessories) to lovers in Christmas season. Therefore, the profit of the sales can be increased by recommending users "cuff links" with long-sleeved shirts in winter and with thin long-sleeved shirts in summer.

Q3: Which Category is the most Local Outlier in the Order Count grouped-by Prefecture? The result is depicted in Fig. 6. We observe that "sun visors" is identified as the most local outlier, since it has an exceptional trend in several prefectures compared with its nearest neighbors. For example, the sales of "sun visors" are extremely high in Ibaraki and Gunma, but it is low in Hokkaido. This result is explained by the fact of the annual sunshine hours reported by the Ministry of Land, Infrastructure, Transport and Tourism. Ibaraki and Gunma were ranked within top-5 in 47 prefectures in 2014. In Hokkaido, the annual sunshine hours is lower than average and the average air temperature is relatively low, so the customers in Hokkaido are not expected to use "sun visors". Therefore, it may be possible to increase the profit of the sales by recommending sunshade hats and sunlight control items to the customers, who live in the prefectures with long annual sunshine hours or high average air temperature. From another perspective, the result may follow the Golf population[2]. According to this external dataset, the Golf population in Ibaragi and Gunma rank at 1 and 7, respectively.

Q4: Which Category is the most Local Outlier in the Sum of the Sales of each Age range of women? The result is depicted in Fig. 7. We observe that "necktie pins" is identified as the most local outlier. It is interesting to find that all its nearest

[2]https://todo-ran.com/t/kiji/19677

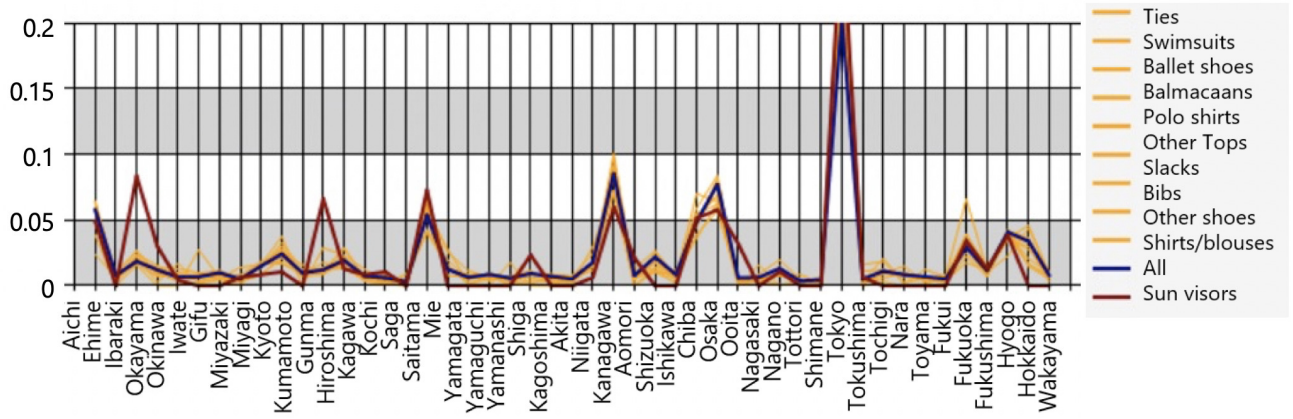

Figure 6: local outlier "sun visors" in the order count grouped by 47 prefectures

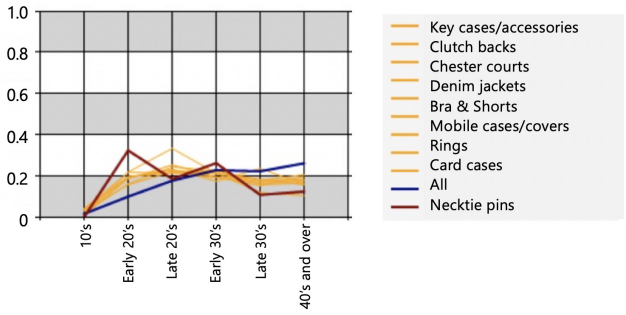

Figure 7: local outlier "necktie pins" in the sum of the sales grouped by age ranges of women

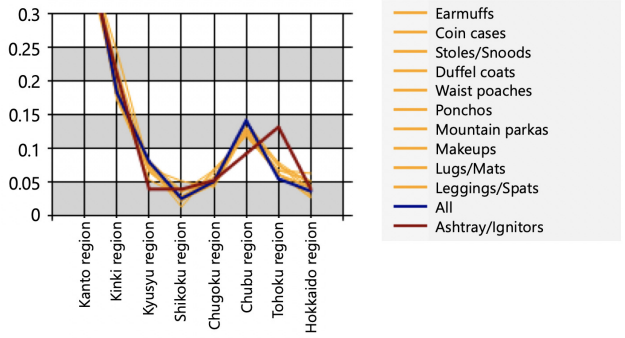

Figure 8: local outlier "ashtray/ignitor" in the order count grouped by eight regions

neighbors are the types of women's items, however, only "necktie pins" is the type of men's items. The sales of "necktie pins" is exceptional, in particular, it is high for the age range of the early 20's women and low in the late 30's or later. We conjecture that the 20's women present "necktie pins" to their boyfriends in the same age range, because there are many men who start jobs in their 20's. Therefore, there is a possibility that the sales can be increased by recommending young women "necktie pins" just before major anniversaries, Christmas day, or birthday.

Q5: Which Category is the most Local Outlier in the Count of the Orders of each Region? The result is depicted in Fig. 8. We observe that "ashtrays/ignitor" is identified as the most local outlier. This figure also shows that the Tohoku region is the only source of the outlierness. We conjecture that the reason is that smoking rate is higher in northern part of Japan, including Tohoku region, as reported in "Overview of National Life Basic Survey" [3] by the Ministry of Health, Labor and Welfare.

5 RELATED WORK

In this section, we review the related work to our work. We describe visualization and analysis tools, exploratory data analysis techniques, and LOF techniques.

As for the data visualization, Polaris [19] is a system that integrates basic database queries with visualization by using Table algebra. Tableau is a commercial data visualization tool for data analysts, which is developed based on Polaris. These visualization tools automatically select an optimal visualization settings for a dataset, however they require manual selection of all attributes for analysis. Google Fusion Tables and DEVise [12] are tools that automate processes of collecting, integrating, and visualizing data from multiple data sources. Google Fusion Table gathers various data from the Web, and then creates a table by integrating the data, and then visualizes analysis results. DEVise is a data search system that enables users to view and share visualized analysis results of huge datasets composed of multiple data sources.

There are many exploratory data analysis techniques, such as Sarawagi [16], Tang et al. [21], MuVE [5] SEEDB [15, 23, 24], Mizuno et al. [13], and Zenvisage [17, 18]. Sarawagi [16] proposed a method that searches for a specific single cell in a multi-dimensional data cube. Tang et al. [21] proposed a systematic framework that searches for top-n analysis results based on both multiple utility functions and select operation in order to automatically extract multiple insights without any user inputs. MyVE [5] quantifies each view of data slice by using multiple utility functions in order to enable group-by in numerical dimensions, which are not supported by SEEDB, and then identifies OLAP queries with high usefulness. Although SEEDB [15, 23, 24],

[3] https://www.mhlw.go.jp/toukei/saikin/hw/c-hoken/03/hyo2.html

Mizuno et al. [13], and Zenvisage [17, 18] are different from each other in terms of analysis workflow, all of them evaluate the outlierness of data slices according to the aspect of the global outlier. Our framework, D4C, automatically searches for exceptional data slices in a similar way as the systems [13, 17, 18] but it employs LOF for detecting unexpected trends.

LOF is a major technique of anomaly detection, and thus there are many derivatives techniques [3]. LOCI [14] speeds up computing LOF values by introducing a new outlier measure by using the standard deviation of the local density of k nearest neighbors. LoOP [10] introduces a probabilistic concept to LOF so that it makes consistent to quantitative interpretation of local outliers even in the distance space having a multimodal distribution. Knorr's method [9] divides the distance space into hypergrids for reducing computation cost with respect to the number of data linearly.

6 CONCLUSION

We described D4C, which automatically identifies top-n data slices that generate local outlier results of automatically generated OLAP queries. D4C is built on top of RDBMS. Through our Fashion EC data analysis, we identified local outliers by using D4C and we explained how to use the analysis results in practice. We also showed that how the analysis results help making decisions on sales strategies.

There are two types of future work. First, we extend our system to omit the user input, a query template. By computing p-value [4, 11] from LOF value, we allows comparing local outliers obtained by multiple query templates. Second, we introduce a semantic hierarchy between attributes, so that we can drill down and roll up the analysis results.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP16K00154.

REFERENCES

[1] Christopher Ahlberg. 1996. Spotfire: An Information Exploration Environment. *ACM SIGMOD Record* 25, 4 (December 1996), 25–29.

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: Identifying Density-based Local Outliers. *ACM SIGMOD Record* 29, 2 (May 2000), 93–104.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR)* 41, 3 (July 2009), 15:1–15:58.

[4] Debabrata Dash, Jun Rao, Nimrod Megiddo, Anastasia Ailamaki, and Guy Lohman. 2008. Dynamic Faceted Search for Discovery-driven Analysis. (2008), 3–12. https://doi.org/10.1145/1458082.1458087

[5] Humaira Ehsan, Mohamed A. Sharaf, and Panos K. Chrysanthis. 2016. MuVE: Efficient Multi-Objective View Recommendation for Visual Data Exploration. *Proceedings of the ICDE* (2016). https://doi.org/10.1109/ICDE.2016.7498285

[6] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc.

[7] Stratos Idreos, Olga Papaemmanouil, and Surajit Chaudhuri. 2015. Overview of Data Exploration Techniques. *Proceedings of the ACM SIGMOD* (2015). https://doi.org/10.1145/2723372.2731084

[8] Niranjan Kamat, Prasanth Jayachandran, Karthik Tunga, and Arnab Nandi. 2014. Distributed and interactive cube exploration. *Proceedings of the International Conference on Data Engineering* (2014), 472–483. https://doi.org/10.1109/ICDE.2014.6816674

[9] Edwin M. Knorr and Raymond T. Ng. 1998. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proceedings of the International Conference on Very Large Data Bases*. 392–403.

[10] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. 2009. LoOP: Local Outlier Probabilities. In *Proceedings of the ACM Conference on Information and Knowledge Management*. 1649–1652.

[11] Martin Krzywinski and Naomi Altman. 2013. Significance, P values and t-tests. *Nature Methods* 10 (oct 2013), 1041. https://doi.org/10.1038/nmeth.2698http://10.0.4.14/nmeth.2698

[12] Miron Livny, Raghu Ramakrishnan, Kevin Beyer, Guangshun Chen, Donko Donjerkovic, Shilpa Lawande, Jussi Myllymaki, and Kent Wenger. 1997. DEVise: integrated querying and visual exploration of large datasets. *ACM SIGMOD Record* 26, 2 (June 1997), 301–312.

[13] Yohei Mizuno, Yuya Sasaki, and Makoto Onizuka. 2017. Efficient Data Slice Search for Exceptional View Detection. In *International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data*.

[14] Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B. Gibbons, and Christos Faloutsos. 2003. LOCI: Fast Outlier Detection Using the Local Correlation Integral. In *Proceedings of the International Conference on Data Engineering*. 315–326.

[15] Aditya Parameswaran, Neoklis Polyzotis, and Hector Garcia-Molina. 2013. SeeDB: Visualizing Database Queries Efficiently. *Proceedings of the VLDB Endowment* 7, 4 (December 2013), 325–328.

[16] Sunita Sarawagi. 2000. User-Adaptive Exploration of Multidimensional Data. In *Proceedings of the International Conference on Very Large Data Bases*. 307–316.

[17] Tarique Siddiqui, Albert Kim, John Lee, Karrie Karahalios, and Aditya Parameswaran. 2016. Effortless Data Exploration with Zenvisage: An Expressive and Interactive Visual Analytics System. *Proceedings of the VLDB Endowment* 10, 4 (November 2016), 457–468.

[18] Tarique Siddiqui, John Lee, Albert Kim, Edward Xue, Chaoran Wang, Yuxuan Zou, Lijin Guo, Changfeng Liu, Xiaofo Yu, Karrie Karahalios, et al. 2017. Fast-Forwarding to Desired Visualizations with zenvisage. In *Biennial Conference on Innovative Data Systems Research*.

[19] Chris Stolte, Diane Tang, and Pat Hanrahan. 2002. Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (Jan. 2002), 52–65.

[20] Bo Tang, Shi Han, Man Lung, Yiu Rui, and Ding Dongmei. 2017. Extracting Top-K Insights from Multi-dimensional Data. *Proceedings of the ACM SIGMOD* (2017). https://doi.org/10.1145/3035918.3035922

[21] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. 2017. Extracting Top-K Insights from Multi-dimensional Data. In *Proceedings of the ACM International Conference on Management of Data*. 1509–1524.

[22] Manasi Vartak and Samuel Madden. 2014. SEEDB : Automatically Generating Query Visualizations. *Proceedings of the VLDB* 7, 13 (2014), 1581–1584. https://doi.org/10.14778/2733004.2733035

[23] Manasi Vartak, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2014. SEEDB: Automatically Generating Query Visualizations. *Proceedings of the VLDB Endowment* 7, 13 (August 2014), 1581–1584.

[24] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2015. SEEDB: efficient data-driven visualization recommendations to support visual analytics. *Proceedings of the VLDB Endowment* 8, 13 (September 2015), 2182–2193.