



(51) International Patent Classification:

G06F 17/30 (2006.01) G06N 5/02 (2006.01)

(21) International Application Number:

PCT/EP2017/058671

(22) International Filing Date:

11 April 2017 (11.04.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: **NOKIA SOLUTIONS AND NETWORKS OY** [FI/FI]; Karaportti 3, 02610 ESPOO (FI).

(72) Inventors: **ZINGER, Nir**; Habrosh 63, 42860 Moshav Burgata (IL). **SHABTAY, Lior**; 8/45 Yohanan Bader st., 52102 Ramat Gan (IL). **DATTNER, Itai**; Harimon 19, 3082500 Ein Ayala (IL). **LAVIE, Eli**; Ha Halil 10/5, 49002 Petah Tikva (IL).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

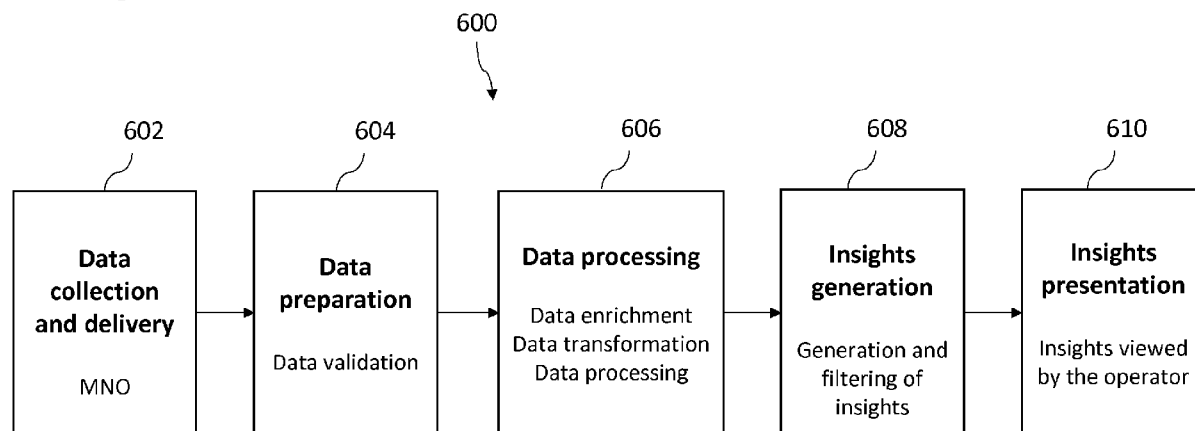
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: A COMPUTER IMPLEMENTED DATA PROCESSING METHOD

Figure 6



(57) Abstract: A computer implemented method comprising: analysing first data to determine itemsets that occur frequently within the first data, said first data comprising a plurality of sets of second data comprising items, said itemsets comprising two or more of said sets of second data; processing said itemsets to determine frequently occurring itemsets and to provide a set of filtered relatively frequently occurring itemsets; and using said set of filtered relatively frequent itemsets to provide an output.

A COMPUTER IMPLEMENTED DATA PROCESSING METHOD

FIELD

Some embodiments relate to method and apparatus for analysing data. Some
embodiments relate to a method and apparatus for processing network data, for
5 example to determine a cause of a problem in that network.

BACKGROUND

The advances in Big Data technologies allow mobile network operators (MNOs) to
keep records in a resolution of per transport-layer and application-layer session, as
10 well as per each event of each voice and data call at the network level. This
information typically adds up to millions of records per hour.

There may be a number of potential data sources. For example: Network Elements
(Base stations, RNC (radio network controller), SGSN (Service GPRS (general
packet radio service) support node, GGSN (Gateway GPRS Support Node) , etc.);
15 Probes (e.g. which collect network, transport, and application layer data); OSS
(Operations Support System); BSS (Business Support System); Traffica (Nokia) –
Real-time collection of network events; CRM (Customer Relationship Management)
system; and various influencer sources such as Internet sites, Weather reports,
Census, etc.

20 Problems in processing big data are not limited to the context of mobile networks
and can be found in a variety of different scenarios.

It is a technical challenge to extract meaning from this vast quantity of data. Some
embodiments may address this technical challenge.

25 SUMMARY

According to a first aspect, there is provided a computer implemented method
comprising: analysing first data to determine itemsets that occur frequently within the
first data, said first data comprising a plurality of sets of second data comprising

items, said itemsets comprising two or more of said sets of second data; processing said itemsets to determine frequently occurring itemsets and to provide a set of filtered relatively frequently occurring itemsets; and using said set of filtered relatively frequent itemsets to provide an output.

5

The processing may comprise removing one or more relatively frequently occurring itemsets from said set of relatively frequently occurring itemsets.

The processing may comprise applying a statistical function to said itemsets.

10

The applying a statistical function to said itemsets may comprise assigning a respective score to a respective itemset.

The processing may comprise determining a statistical function with respect to a key performance indicator associated with a plurality of sets of second data over all sets of second data associated with said key performance indicator with respect to a respective itemset.

15

The determining the statistical function may comprises one or more of determining an average and a standard-deviation of the key performance indicator over all sets of second data with respect to a respective itemset.

20

The processing may comprise determining if a respective itemset is a core itemset, said core itemset being one where there is not subset of that itemset having a score value which is greater than or equal to a score value of the itemset multiplied by a significance ratio.

25

The processing may comprise determining if a respective itemset is a tightened itemset, if for each core superset of the respective itemset, associated support information of respective itemset less the associated support information of the superset is greater than a threshold amount.

30

The processing may comprise using support information for a itemset, said support information for a respective itemset providing information about the number of sets of said second data comprising said respective itemset.

- 5 The processing may comprise determining if a respective itemset is a tightened area itemset, where statistics of a respective itemset is determined without sets of second data of a superset group of data, wherein respective itemset is determined to be a tightened area subset if said respective itemset is also a core itemset and a tightened itemset, according to the said statistics of the respective itemset.

10

The processing may be controlled by one or more parameters, said one or more parameters comprising an improvement-ratio, a significance-ratio, and an improvement/significance-change.

- 15 The providing an output may be based on at least one of one or more specified key performance indicators, items with a specified item of interest, and itemsets, using an analysis technique, wherein the analysis technique is at least one of pattern recognition and statistical analysis.

- 20 The method may comprise receiving a query, the query comprising one or more of key performance indicators, information about one or more data sources, and at least one item of interest, said using of said set of filtered relatively frequent itemsets to provide said output being dependent on said query

- 25 The method may comprise obtaining first source data, from a first source and second source data from at least one second source, said first source data and said second source data comprising said first data, wherein at least one of said sets of second data comprises first source data and second source data

- 30 The first data may comprise communication data and said output comprises information indicating one or more network parameters related to network performance.

The communication data may comprise one or more of subscriber data; application data; user device data; network operator data and internet data.

5 According to another aspect, there is provided a computer apparatus, said computer apparatus comprising at least one processor and at least one memory including computer code for one or more programs, the at least one memory and the computer code configured, with the at least one processor, to cause the apparatus at least to analyse first data to determine itemsets that occur frequently within the first data, said first data comprising a plurality of sets of second data comprising items, said
10 itemsets comprising two or more of said sets of second data; process said itemsets to determine frequently occurring itemsets and to provide a set of filtered relatively frequently occurring itemsets; and use said set of filtered relatively frequent itemsets to provide an output.

15 A database may be provided to be used in conjunction with the computer apparatus, said database storing said first data. Alternatively or additionally, the first data may be stored in the at least one memory of the computer apparatus.

20 The at least one memory and the computer code may be configured, with the at least one processor, to remove one or more relatively frequently occurring itemsets from said set of relatively frequently occurring itemsets.

The at least one memory and the computer code may be configured, with the at least one processor, to apply a statistical function to said itemsets.

25

The at least one memory and the computer code may be configured, with the at least one processor, to assign a respective score to a respective itemset.

30 The at least one memory and the computer code may be configured, with the at least one processor, to determine a statistical function with respect to a key performance indicator associated with a plurality of sets of second data over all sets of second data associated with said key performance indicator.

The at least one memory and the computer code may be configured, with the at least one processor, to determine at least one of an average and a standard-deviation of the key performance indicator over all sets of second data with respect to a respective itemset.

5

The at least one memory and the computer code may be configured, with the at least one processor, to determine if a respective itemset is a core itemset, said core itemset being one where there is not subset of that itemset having a score value which is greater than or equal to a score value of the itemset multiplied by a significance ratio.

10

The at least one memory and the computer code may be configured, with the at least one processor, to determine if a respective itemset is a tightened itemset, if for each core superset of the respective itemset, associated support information of respective itemset less the associated support information of the superset is greater than a threshold amount.

15

The at least one memory and the computer code may be configured, with the at least one processor, to use support information for a itemset, said support information for a respective itemset providing information about the number of sets of said second data comprising said respective itemset.

20

The at least one memory and the computer code may be configured, with the at least one processor, to determine if a respective itemset is a tightened area itemset, where statistics of a respective itemset is determined without sets of second data of a superset group of data, wherein respective itemset is determined to be a tightened area subset if said respective itemset is also a core itemset and a tightened itemset.

25

The at least one memory and the computer code may be configured, with the at least one processor, to cause the processing to be controlled by one or more parameters, said one or more parameters comprising an improvement-ratio, a significance-ratio, and an improvement/significance-change.

30

The at least one memory and the computer code may be configured, with the at least one processor, to cause the output to be based on at least one of one or more specified key performance indicators, items with a specified item of interest, and
5 itemsets, using an analysis technique, wherein the analysis technique is at least one of pattern recognition and statistical analysis.

The at least one memory and the computer code may be configured, with the at least one processor, to receive a query, the query comprising one or more of key
10 performance indicators, information about one or more data sources, and at least one item of interest, said using of said set of filtered relatively frequent itemsets to provide said output being dependent on said query

The at least one memory and the computer code may be configured, with the at least
15 one processor, to obtain first source data, from a first source and second source data from at least one second source, said first source data and said second source data comprising said first data, wherein at least one of said sets of second data comprises first source data and second source data

20 The first data may comprise communication data and said output comprises information indicating one or more network parameters related to network performance.

The communication data may comprise one or more of subscriber data; application
25 data; user device data; network operator data and internet data

According to another aspect, there may be provided an apparatus configured in use to provide the any of the previous methods.

30 The apparatus may comprise a computer device, a server, a bank of servers or the like.

A computer program comprising program code means adapted to perform the herein described methods may also be provided. In accordance with further embodiments apparatus and/or computer program product that can be embodied on a computer readable medium for providing at least one of the above methods is provided.

5

It should be appreciated that any feature of any aspect may be combined with any other feature of any other aspect.

BRIEF DESCRIPTION OF DRAWINGS

10 For a better understanding of some embodiments, reference will now be made by way of example only to the accompanying drawings in which:

Figure 1 schematically shows a database containing a number of transactions;

Figure 2 shows an example of division of the data space into classes;

15 Figure 3 schematically shows a illustrates some example dimensions that may affect subscriber experience;

Figure 4 shows an example of frequent-itemset mining and association-rule mining;

Figure 5 shows an input/output diagram for insight generation according to an embodiment;

Figure 6 shows a high level flowchart of an embodiment;

20 Figures 7 shows, by way of example, analysis of the deviation of an error-rate average amongst different transaction-population groups;

Figure 8 show an example for a determining core-itemsets from a list of frequent-itemsets;

25 Figure 9 show an example for a determining tightened-itemsets from a list of frequent-itemsets;

Figure 10 shows an example relating to filtering populations of transactions according to deviation of statistical functions;

Figure 11 shows an example flowchart relating to embodiments;

Figures 12a and 12b show a detailed example flowchart relating to embodiments;
and

Figure 13 shows an example of an apparatus of some embodiments.

5

DETAILED DESCRIPTION OF SOME EMBODIMENTS

Some embodiments may provide methods relating to the analysis and mining of the data which allow for example mobile network operators to gain new information and/or insights.

10 Quality of Experience (QoE) is the performance of a network from end-user point of view. It is a subjective measure of the end-to-end service performance from the user perspective and is an indicator of how well the overall network (including the operator network, the internet, the end user's own device and application, etc.) meets the user's needs

15 Quality of Service (QoS) is a measure of performance from a network point of view. It focuses on network performance and its capability to deliver the service according to the specified service level. QoS metrics include objective network layer parameters such as bandwidth, packet loss, delay, and delay variation.

Key Quality Indicators (KQIs) describe QoE perceived by the user. Key Performance
20 Indicators (KPIs) describe QoS delivered by the network.

QoE is related to QoS. For example, sufficient QoS is a precondition for QoE, and some QoS KPIs, like those which describe bandwidth, loss, and delay directly influence the QoE for different applications. However, this influence is not necessarily through a strict formula or relationship, as other factors are also often
25 involved.

The previously presented practice of binning and thresholding does not always provide the complete picture. The subjective nature of quality of experience makes it difficult to automatically detect and/or determine the scope of the population of

transactions and subscribers which suffer from lower quality of experience, and obtain insights about the drivers leading to the reduced quality of experience.

Some embodiments may provide methods of analysis and mining of this information in order to gain information that was beforehand unknown to the MNO.

- 5 Determining the “driving forces” behind the phenomena seen by inspecting a database of user-plane sessions collected from the network may be advantageous in rectifying problems with the network. Determining the driving forces may include accurately scoping the population that is effected by each phenomenon. For example, when rules reveal faults in a network, each such fault is a ‘driving force’
10 which is common to a potentially large number of specific session degradations or faults, which are recorded in the database. Accurately scoping the population of sessions which are affected by each fault assists domain-experts in analysing and fixing the fault.

- Some embodiments, in the context of big data, relate to data mining. Some
15 embodiments may use frequent-itemset mining and association-rule mining. Frequent-itemset mining and association-rule mining may be used for performing Market Basket Analysis (MBA).

- In some embodiments, performing frequent-itemset and association-rule mining may comprise providing a database in which each record in the database is a transaction.
20 By way of example, the transaction may represent a basket of a single purchase, in the case of MBA. In this example, the transaction is a list of items. For example, in MBA these are the items in the basket represented by that transaction.

- A transaction may contain an itemset (IS). An itemset is a set of items, for example, a transaction (Milk, Bread, Sugar) contains the itemsets (Bread, Sugar), (bread),
25 (Milk, Sugar), (Milk, Bread, Sugar), etc.

In the context of a mobile network, the itemset may include appropriate items, for example operating system, type of address, device type and any other suitable items.

- The support of an itemset refers to the fraction or percentage of the transactions in
30 the database which contain that itemset, otherwise known as the itemset population.

Frequent itemsets are itemsets which have support which is greater than a pre-defined value, this value may be referred to as min-support (minimum support).

An itemset **S** is considered as a superset of an itemset **C** if all items in **C** appear also in **S**. In this case **C** is considered as the subset of **S**. The support of the subset **C** is
 5 always equal or greater than the support of its superset **S**.

Two previously presented algorithms for mining all frequent itemsets (i.e. those that pass the min-support threshold constraint) from a database of transactions are Apriori and FP-growth (Frequent Pattern Growth).

10 Association rules uncover relationships between seemingly unrelated data in a database. An example of an association rule would be "If a transaction contains a bread, then it is 80% likely to also contain milk."

Reference is made to Figure 1, which schematically shows a database 100 containing a number of transactions 106, 108, 110 and itemsets 102, 104. For example, association rule $[A, B] \Rightarrow y$ represents the relationship between the itemset
 15 populations of two itemsets: $[A, B]$ 104 and $[A, B, y]$ 102. In this example, $[A, B]$ 104 is considered as the 'left-hand side' of the rule, and $[y]$ is considered as the 'right-hand side'. The support of the rule is the size of the population of $[A, B, y]$ 102 relative to the size of the number of transactions in the database 100, which may be represented as: $|[A, B, y]|/|DB|$.

20 The confidence of the rule is the proportion of the members of $[A, B]$ 104 for which $[y]$ also holds: $|[A, B, y]|/|[A, B]|$. A rule **S** is considered as a subset/superset of a rule **C** if: (1) both rules have an identical right-hand side, and (2) the left-hand side of **S** contains a subset/superset of the items in the left-hand side of **C**.

In some embodiments, all association rules which comply with a min-support
 25 constraint may be found by determining all frequent-itemsets and dividing the frequent-itemsets in different ways into right-hand and left-hand parts. Each resulting rule is then checked, and those that do not pass the min-support or minimum confidence constraints are dropped.

Classification learning, otherwise known as classification training, uses a pre-
 30 classified set of data and models the data by creating a division of the data space

into 'classes'. The set of data comprises records containing different attributes/fields, each of which can take different values, with one class attribute which classifies each record to a specific 'class'.

Reference is made to Figure 2, which shows an example of a division of the data space 200 into three classes 202, 204, 206 according to a class attribute. By way of example the class attribute is represented by a shape in Figure 2. That is to say, circles represent class 202, squares represent class 204, and triangles represent class 206. The lines represent the outcome of a classification process, which models the mapping of the records space 200 into different classes 202, 204, and 206.

- 10 In some embodiments, classification of records in a database is achieved using association-rules mining by creating a modified database. The method of creating a modified database may comprise creating an item representing each value of each field. For example, if a field f gets the value v in a specific row, the item f_v is inserted to the respective transaction "basket". In some embodiments, each class
- 15 value may also be represented as an item.

By creating an item representing each value of each field, the database is modified to become a database of transactions as defined above, and as such can now be mined to find association rules.

- In some embodiments, only association rules containing one of the items
- 20 representing a class attribute or class-value are of interest, and the others may be removed.

- Some advantages that may be provided by using association rules for classification modelling are that (a) association rules are easy to interpret and understand (b) algorithms for association-rules mining are exhaustive and find all the rules which
- 25 comply to the specified bounds.

In some embodiments, scoping or localization of network problems based on communication-activity transactions is provided using association-rule based classification. Communication-activity transactions may include at least one of calls or sessions. Each record may contain details about one or more of the recorded

communication-activity, involved network-elements, technical-details, error-codes, performance-indicators, or quality-indicators, etc.

In some embodiments, the communication-activity records may be classified according to one or more parameters, for example, a classification may include at least one of whether and why the activity was successful or not, the activities performance, or the activities quality. This classification may be used, for example, to detect and/or determine the scope of failures in the network. Detecting failures allows them to be fixed.

High complexity of internet services causes MNOs to have limited understanding of the subscriber's experience. This generates a high customer care work load, and places significant burden on the engineering and operations teams.

Subscriber experience is a complex issue. Understanding subscriber experience and the forces driving it is a complex issue due to the large number of dimensions that can impact the subscriber experience. One or more dimensions may affect a user's experience. A dimension may be a variable with one or more values that can potentially effect the subscribers' experience. A dimension may take many values resulting in a huge number of potential combinations. Some of these dimensions may be outside of the direct control of the MNO.

Figure 3 illustrates some example dimensions or parameters that may affect subscriber experience. The subscriber experience may be affected by dimensions relating to at least one of the subscriber 302, the app 304, the mobile device 306, the MNO 308, and the internet 310.

Dimensions that affect subscriber experience relating to the subscriber 302 may include, for example, one or more of the subscriber plan, and the usage pattern.

Dimensions that affect subscriber experience relating to the app 304 may include, for example, one or more of the application type, application efficiency, application protocol, code type, and the requested resolution. More specifically, code type, and the requested resolution may be used for specific applications, e.g. in video streaming.

Dimensions that affect subscriber experience relating to the mobile device 306 may include, for example, one or more of the device's processing power, available memory, screen resolution, and configuration.

5 Dimensions that affect subscriber experience relating to the MNO 308 may include, for example, one or more of the cell characteristics, aggregation, transport, core radio access technology (RAT), communications service provider (CSP) throughput, latency, loss, and communications service provider quality of service.

10 Dimensions that affect subscriber experience relating to the internet 310 may include, for example, one or more of latency, loss, throughput, server load, and performance.

In some embodiments, the population of transactions and subscribers which suffer from lower quality of experience are automatically detected and the scope of the problem may be determined. The population of transactions and subscribers which suffer from lower quality of experience may be further analysed based on associated
15 data to determine insights about the drivers leading to their reduced experience.

In embodiments, part or all of the data described in relation to Figure 3 can be used in order to create a communication-activity database. A communication-activity database may comprise a transaction per event and/or session at a specific layer. A transaction may be represented, for example, by a row in a database.

20 A communication-activity database may comprise, for example, a transaction per call, or per reported network-layer event. Alternatively or additionally, a communication-activity database may comprise a transaction per application-layer session. The application-layer session may be one or more of a Domain Name System (DNS) session, a HyperText Transport Protocol (HTTP) session, a File
25 Transfer Protocol (FTP) session or the like.

A reported network-layer event may comprise, for example, one or more of the following attributes per call-party: Radio Network Controller (RNC) id, Cell-Id, Servicing GPRS Support Node (SGSN) id, Technology used by the call-party, Day-of-week, and a time stamp. Alternatively or additionally, a reported network-layer
30 event may comprise, for example, one or more of the following attributes per call:

Start-time, Duration, and a code representing the cause of the end of the call or the like.

An application-layer session may, for example provide call attributes. Call attributes may include, for example, the cell-id, the Access Point Name (APN). An application-layer session may, for example provide application level attributes, for example, the
5 protocol, host-name, Day-of-week, or time stamp. The application-layer session may, alternatively or additionally provide performance reflecting attributes, for example, one or more of retries-count, latency, packet-loss, and application end-cause.

Given a set of communication-activity records which record the network activity
10 during a specific time-period, some embodiments may identify and/or determine the scope of the factors which have a significant impact on the overall/average QoS or QoE of large groups of subscribers.

Some embodiments may use data mining techniques and advanced analytics to automate insight generation. Insight generation may comprise finding the patterns in
15 the data base and/or determine the scope of the impacting factors. Insights generation may comprise finding the patterns in a modified database, such as a communication-activity database. Insights may describe, for example, systematic issues that impact large number of subscribers. For example, a systematic issue may be that roamers from a specific network using a specific device type on a
20 specific radio technology experience performance issues.

By automatically determining systematic issues, operators may focus on and resolve systematic issues. Systematic issues, such as low throughput, may result in low data service experience. Automatic detection of systematic issues allows issues to be fixed faster, and resolving issues that impact subscriber's data service experience
25 may improve the network performance. It should be appreciated that the same technique may alternatively be used to determine good performance to identify parameters which may be changed for users which experience poor performance.

Reference is made to figure 4, wherein frequent-itemset mining and association-rule mining may be used in order to find insights about network conditions and parts
30 which lead to a low QoS. By way of example, an analogy is drawn between a supermarket user 410 and a network user 420. User 410 completes transactions, a

transaction may be represented by a row in the data, wherein each row contains a set of purchased items such as bread 411, eggs 412, and milk 413. A product is a purchased item in the basket. Insight generation may be performed on a database containing transactions of a plurality of supermarket users. Insight generation may provide insights 414, for example, 91% of the customers that bought bread and eggs, also bought milk. By way of example, the network user 420 completes transactions, a transaction may be represented by a row in the data, wherein each row contains a communication-activity record. Each communication activity may contain a set of one or more virtual "products". Each product may represent, for example, a network issue such as high delay 423. Each product may be, for example, a value of a specific field, such as device type such as device type y 422, or RNC X 423. Insight generation may be performed on a database containing transactions of a plurality of network users. Insight generation may provide insights 424, for example, 91% of the subscribers who were browsing using device type Y through RNC X, experienced high delay.

A frequent-itemset mining and association-rule mining based method allows a large database with multiple dimensions to be analysed quickly. A limitation of frequent itemsets, and association/classification rules is that they only work with discrete items, and enumerated features that can be converted into items. Association-rules mining may be used to provide classifications. For example, if some of the fields are numeric, they may first be mapped into enumerated fields using a binning or discretization process. More specifically, an enumerated field or feature is derived from a numeric feature by defining which value ranges map to which target enumerated value. An example numeric value can be a key performance indicator (KPI) of a network element.

For example, a database of transactions which describe network events or sessions may contain a field that is a KPI, which describes a specific aspect of that network event, such as the percentage of packets which were lost during that session. The database may contain features which describe different details about each session, or each transaction, for example, one or more of APN, RAT_type, device_type, target web_domain, RNC Name, City, Antenna type, and cell_load. Furthermore, the database may contain one or more key performance indicator (KPI) which describe

the performance of the sessions, such as data-rate, TCP-retransmission rate or the like.

It may be desirable to understand and reveal failures or phenomena in the network which effect the values of this KPI for specific groups of events. A classification

5 target may be obtained by mapping the KPI according to acceptable and problematic values for the network. A value range of the KPI which is considered acceptable and/or a value range which is considered problematic for the network may be defined. Solutions for automatic mining of association and classification rules may then be performed.

10 Reference is made to figure 5, which shows an input/output diagram for insight generation 500 according to an embodiment. An insight 510 may comprise an output of the processor(s) processing the data. An insight 510 may, for example, identify a problem in the network. An insight may be used to fix the problem, for example, by controlling an aspect of the network. In that case, the insight may be one or more
15 control outputs. The output may be provided to a user interface and/or to a control apparatus. An automated insight generation unit 508 is provided with a QoE impactor selection 502, and communication-activity data. The communication-activity data may for example be user-plane (application layer) data 504, and control plane (network layer) data 506. Based on the user-plane data 504, control plane data 506
20 and the QoE impactor selection 205, the automated insight generation unit 508 automatically finds and/or determines the scope of meaningful issues that impact subscriber's experience. The automated insight generation unit 508 generates insights that allow operators to understand what leads to low subscriber's experience.

25 Insights 510 may include, for example, meaningful issues that are common to subscribers getting lower speed test results or meaningful issues that are common to cells with higher dropped calls ratio. Insights may automatically find drivers for one or more of the following phenomena using TCP Drivers analysis: TCP low throughput (e.g. per app), TCP high retransmission, and TCP high Latency. Furthermore,
30 Insights may automatically find drivers for one or more of the following phenomena using DNS Drivers analysis: DNS Errors, DNS no reply, and DNS high latency.

In some embodiments data mining techniques may be used in combination with advanced analytics.

Reference is now made to figure 6 which describes an embodiment. A

communication-activity database may be created 602 by monitoring the network at

5 IP based interfaces. The IP based interfaces may be the GN and GI interfaces. The GN interface may provide an IP based interface between SGSN and other SGSNs and (internal) GGSNs. The GI interface may provide IP based interface between the GGSN and a public data network (PDN). The data is optionally validated 604 to confirm the created database is free of errors. The data is enriched 606 using data
10 about the topology and configuration of the network. Insights are then generated from this data 608. The method may find insights which are based on patterns leading to statistical deviation of KPI or KQI values. Insights may enable identification of populations which suffer from reduced QoE compared to other subscribers or sessions. The insights may pinpoint combinations of different domain-
15 feature values which drive QoE-impactor statistics to less desirable values. These insights may then be presented to the operator 610 for further investigation and solving.

This differs from the classical classification-based methods which serve well when looking for QoS breaches, but are not adapted to provide QoE improvement

20 regardless of specific well-defined QoS thresholds. This may allow the MNO to save operational expenditure and increase QoE. Operational expenditure may be reduced by increasing engineering and network efficiency. There is a technical improvement to one or more parts in the network. For example performance in the network may be improved.

25 Embodiments may improve QoE due to preventative care as systematic issues may be repaired before the customer calls customer care. There is a technical improvement in that identification of a problem and its solution may be made quickly without requiring the problem first to be identified by an end user.

Furthermore, issues may be revealed that are not detected by traditional network
30 monitoring methods.

Traditional network monitoring methods require a manual and repetitive drill down by experts to generate insights. The traditional methods require the use of thresholds. This requires the correct setting a threshold manually and if set incorrectly will not identify a problem. The manual nature of insight generation leads to missed or
5 incorrectly scoped issues. Furthermore, pre-tailored reporting is necessary and is focused on specific scenarios. This technique requires pre-knowledge of one or more parameters that are causing a problem and that the problem itself is known.

In some embodiments, automatic insight generation is provided, which provides comprehensive identification and accurate scoping of issues. Furthermore, automatic
10 insight generation may provide a cognitive self-aware network. Some embodiments avoid the need for thresholds, such as mentioned previously. Some embodiments are able to provide identification of problems and/or insights for solutions to problems.

The processing of data in embodiments provides an effective use of computing
15 resources compared to traditional arrangements, in some embodiments.

In some embodiments, there is detecting and scoping population of transactions according to their classification confidence or other classification qualities. In some embodiments, alternatively or additionally, detecting and scoping population of transactions to statistical functions which apply to numerical fields is performed.

20 Detecting and scoping population of transactions to statistical functions which apply to numerical fields may provide insights relating to the driving forces causing different transaction populations to have different statistical properties.

Numerical fields may be, for example, KPIs (key performance indicators) and KQIs (Key Quality Indicators) of networks or any other type of systems. This is useful, for
25 example, to find drivers which effect the average of a KPI or a KQI. Furthermore, numerical fields may comprise standard-deviation, or a combination of average and standard-deviation.

Reference is now made to figure 7. Figure 7 shows analysis of the deviation of an error-rate amongst different transaction-population groups 700. Each group is
30 identified by an itemset, for example, Group1 = [A, B]. Assume that the average of the KPI over the entire database 708 is 3.5. It is clear from figure 7 that group 1 702

significantly deviates from this average for the worse, and therefore should be reported. That is to say, the average of itemset represented by group 1 702 exceeds the KPI average of the database and thus may be automatically reported to an operator. Neither group 2 704 nor group 3 706 exceed the average of the KPI of the database, suggesting that they are operating typically with relation to the given KPI. Optionally, a predetermined KPI threshold may be defined by the analytics client.

Mining may be performed using an algorithm for Frequent-Pattern mining. For example, mining may be performed using a frequent-pattern mining library-function, which implements the frequent pattern (FP) growth algorithm. A FP-growth

Parameter may be required to execute the frequent-pattern mining library-function. The FP-growth Parameter may be, for example, min-support that may be set according to operator requirements, what size of population is interesting, as well as according to scalability requirements.

In order to mine population of transactions according to deviation of statistical functions, a list of frequent itemsets is determined, and then the statistical function is calculated for each one.

The serial process as described may suffer from scalability issues (i.e. takes too much time). In some embodiments, the FP-growth algorithm may be amended to calculate the statistical function as part of building the itemsets. Calculating the statistical function as part of building the itemsets may reduce processing time.

Depending on the target KPI and the statistical function, different statistical-function results might be of interest. For example, high average values are of interest when the KPI is the errors percentage in some system, while low average values are of interest when the KPI is the throughput of a specific session represented by the database transactions or records. A scoring function may be used which, for example, assigns a score to each itemset according to the result of applying the statistical function over the population it represents. A scoring function may be used to generalise the comparison of itemsets. A threshold may be used to decide whether each itemset is interesting or not. This threshold can be manually preconfigured or dynamically set as part of the process, for example, the threshold may be set to be the score of the entire database, or the score of an itemset. The list

of frequent itemsets may be large, a scoring function may be used in order to identify and filter-out redundant, less interesting and misleading itemsets.

In some embodiments, a user may specify a minimum improvement-ratio constraint (**R**). Itemsets whose statistical-function result is not an improvement-ratio of at least **R** more than its subsets which appear in the mined itemset list, may be filtered from the itemset list.

A filter may determine whether an itemset is a core itemset. For example, given a significance ratio $R < 1$, an itemset C is a core itemset, if there exists no subset S_b of C such that $\text{score}(S_b) \geq \text{score}(C) * R$. That is to say, if there is no subset S_b of the itemset C having a score which is greater than or equal to the score of the itemset multiplied by the significance ratio R , then C is a core itemset.

Reference is now made to figure 8, which provides an example of a determining core-itemsets from a list of frequent-itemsets 800. In this example, the score is the group-average. Each IS may be numbered for reference, e.g. IS1, IS2, etc. The average of group IS1 is 5, in other words, $\text{Avg}(\text{IS1}) = 5$. Core-itemsets may be determined using the group average of the itemset. Determining the core-itemsets using the group average may be performed according to the group-average deviation. In this embodiment, by way of example, a larger value is considered to be worse (as would be the case, for example, for Error rate KPI), and the significance ratio for each itemset is assigned a value of 0.8. That is to say $R = 0.8$. Figure 8 shows 4 frequent-itemsets: [A,B] 802; [A,C] 806; [A,B,C] 808; and [A,B,D] 804. The average (AVG) for each group, and the significance ratio, may determine if the itemset is a core itemset. For example:

IS1: [A,B] : average = 5 - trivially a core itemset (no subset)

IS2: [A,C] : average = 5 - trivially a core itemset (no subset)

IS3: [A,B,C] : average = 7 – IS3 is a core itemset since $\text{Avg}(\text{IS3}) * R = 7 * 0.8 = 5.6 > 5 = \text{Avg}(\text{IS1}) = \text{Avg}(\text{IS2})$

IS4: [A,B,D] : average = 5.5 – IS4 is not a core itemset, since $\text{Avg}(\text{IS4}) * R = 5.5 * 0.8 = 4.4$ which is less than $\text{Avg}(\text{IS1})$

The core-itemsets filter allows reducing the amount of rules. The itemset-list resulting from the core-itemsets filter may still contain redundancy. An itemset-list with redundant itemsets may drive misleading rules. The filters ‘tightened itemsets’, and ‘tightened-area itemsets’ may be used to reduce redundancy, reducing the risk of misleading rules.

An itemset **C** is considered a tightened itemset, if for each superset **Sp** of **C** such that **Sp** is a core itemset (if such exists), the following holds: $(\text{support}(\mathbf{C}) - \text{support}(\mathbf{Sp})) > \mathbf{X}$. That is to say, if for each and every core superset of **C**, the support of **C** minus the support of the superset is more than **X**, the itemset **C** is a tightened itemset.

X may be, for example, the minimum support value. The support of an itemset refers to the fraction or percentage of the transactions in the database which contain that itemset.

Reference is now made to figure 9, which contains 3 core itemsets 900: [A,B] 902; [A,C] 906; and [A,B,C] 904. The following is a tightened-itemset (group-average deviation) example. In this embodiment, by way of example, a larger value is considered to be worse (as would be the case, for example, for Error rate KPI), and that **X** = 0.05. Below is the frequent-itemset list:

IS1: [A,B] : average = 5, support = 0.25
 IS2: [A,C] : average = 4, support = 0.2
 IS3: [A,B,C] : average = 7, support = 0.19

IS2 is not a tightened itemset since $\text{support}(\text{IS2}) - \text{support}(\text{IS3}) = 0.2 - 0.19 = 0.01$, and the result is less than **X**, $0.01 < 0.05$, IS3 is a superset of IS2, wherein IS3 is a core itemset. Thus IS2 is not a tightened itemset.

In addition to being redundant, non-tightened itemsets might also be misleading. For example, [A, C] may not determine the driving forces. That is to say, the phenomenon may not be a result of the combination of A and C without B, and the rule [A, C] -> average(KPI)=5 is only a side effect of [A, B, C] -> average(KPI)=7.

Reference is made to figure 10. Two groups are shown 1000, group A 1010 and group B 1020. Group A 1010 shows a small number of transactions, having a high KPI. Group B 1020 shows a large number of transactions with a KPI below that of group A, but above the average KPI 1030. When attempting to automatically detect population of transactions according to deviation of statistical functions, the risk of providing misleading rules is larger than with automatic discovery association rules

and classification rules. This is because smaller groups of transactions with considerably different KPI values can have a larger effect on the statistics of larger groups containing them. For example, it may be that the only reason for the itemset of Group B 1020 passing the above-defined filters, is the contribution to the KPI-statistics of the transactions belonging to Group A 1010, wherein Group A 1010 is a subset of group B 1020. In such a case, the rule of Group B 1020 is actually misleading. The 'tightened-area itemsets' filter may be used to overcome this issue. Using the tightened-area itemsets, the statistics of the subset itemset are calculated (e.g. the one representing Group B), not including the transactions of the superset group (e.g. that of Group A). With the newly-calculated statistics, the subset itemset is checked against the 'core' filter and the min-score requirement, if the subset now fails to meet the above criteria (core itemsets, tightened itemsets), the subset is filtered out.

Mining statistical functions of frequent itemsets may provide a lot of important information about the behaviour and faults in the networks. In many cases this information cannot be revealed by mining association-rules. For example, when mining for degradation of service, there is no specific definition of a "fault" and as such, association rules cannot detect such issues. An itemset list which is small enough, non-redundant and not misleading can be analysed by domain experts to detect degradation of service.

Embodiments described herein demonstrate that itemset scoring is not limited to working with discrete items and enumerated features. Therefore binning is not required for numerical features. As such, numerical features may serve as the target score of itemsets.

Some embodiments, automatically detect and/or determine the scope of populations and driving forces effecting the data-rate and TCP-retransmission rate. The outputs of the execution may be one or more of a list of itemsets which identify the respective population and driving-force, the average target-KPI value for each itemset, the support (number of sessions) for each itemset, and the affected-subscribers number for each itemset.

Reference is made to figure 11, which relates to embodiments. In embodiments, data is collected 1100. Data collection may be performed by, or on behalf of, a single operator, or a group of operators. At least one operator may be an MNO operator. Data may be collected from one or more sources 1110, 1120, 1130, such as Traffic
5 1120, a layer7 probe 1110, the user plane, or the control plane. The collected data may be stored using The Hadoop Distributed File System (HDFS). HDFS is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to applications as required. Data preparation 1140 may comprise one or more of selecting relevant transaction data from the HDFS, selecting one or more
10 relevant data columns, and selecting transactions that occurred in a relevant time frame. The selected raw data 1150 may then be loaded into memory 1160, or local storage to enhance performance. The data may be enriched 1170, that is to say the database is modified. Enrichment of data may be performed 1170. Enrichment of data 1170 may involve adding information to a database based on another source of
15 information. Another source of information may include, for example, one or more of a database, or information repository. Enrichment of data may further comprise, for example, classification by device type, OS type, or Cell location. A data transform 1170 may be performed, for example, one or more of dimension value binning, and dimension selection may be performed for each job/assignment. In some
20 embodiments, features and dimensions may be used interchangeably. Following data enrichment, insights 1190 may be generated by running the described method 1180. An insight may, for example, be based on the average of the KPI suggest that subscribers browsing a specific website get lower throughput when the device manufacturer is a specific manufacturer. By way of example, an insight may be “By
25 average, subscribers browsing YouTube get lower throughput where the device manufacturer is company X and the Type Allocation Code (TAC) category is handheld.

Some embodiments may be applied, for example, to any set of data wherein a numerical KPI of interest can be applied. The set of data may be a set of
30 transactions, a set of sessions or other data. For example, the transactions may be the content of supermarket baskets, and the KPI may be the amount of time the customer spent in the store.

Figures 12a and 12b show a detailed example flowchart 1200 relating to embodiments. At step S1202, an analytics client formulates a query, the query including: one or more key performance indicators, a list of relevant data sources, and a request for transactions with a specific property. The list of relevant data sources may include at least one or probe-data or CRM data. The specific property may be, for example, that the transaction is of a specific protocol. The one or more key performance indicators may be per transaction key performance indicators.

At step S1204, the analytics client transmits the formulated query to an analytics server.

At step S1206, the analytics server extracts, from a source database of transactions, the list of relevant data sources from each transaction of the source database of transactions containing the specific property. In some embodiments, the analytics server may only query fields relevant to the specific property of the database. Each transaction in the source database may have one or more feature. One or more feature may comprise a value, wherein a value may be a numerical value.

At step S1208, the analytics server enriches the extracted data to form a modified database, the enrichment process may create one or more new per transaction features created by adding information to a database based on another source of information. In some embodiments, per transaction features may be per transaction data features.

At step S1210, the analytics server applies a data transform to the modified database, wherein the data transform is at least one of feature value binning, and feature selection. Feature value binning may refer to binning by feature value.

At step S1212, the analytics server applies a further data transform to the modified database, converting the per transaction data features into items representing the values of features in each transaction.

At step S1214, the analytics server analyses the modified database to determine itemsets that occur frequently within the modified database, and generates associated data, wherein associated data comprises a score and support for each

itemset. Analysis may, for example, include at least one of frequent-itemset mining and association-rule mining.

At step S1216, the analytics server filters the frequent itemsets to generate a set of filtered itemsets, filtering removes itemsets that are not desired based on one or more of a scoring function, support, and itemset analysis. The scoring function may comprise one or more of the improvement-ratio, or the significance ratio. Itemset analysis may comprise one or more of the core set filter, the tightened itemset filter, and the tightened-area itemset filter.

At step S1218, the analytics server determines at least one insight based on at least one of the one or more key performance indicators, transactions with a specific property and the filtered itemsets using an analysis technique such as pattern recognition or statistical analysis.

Finally, at step S1220, the analytics server transmits the determined insights to the analytics client.

Figure 13 shows an example of an analytics server 1330, for example, to be coupled to and/or for communicating with an analytics client. The analytics client may be provided at the analytics server. The analytics server 1330 can be arranged to provide an output, information processing, and/or communication operations. An analytics server can be configured to provide control functions in association with generation, communications, and interpretation of information repositories. The analytics server 1330 comprises at least one memory 1331, at least one data processing unit 1332, 1333 and an input/output interface 1334. Via the interface the analytics server can be coupled to the analytics client. The analytics server 1330 can be configured to execute an appropriate software code to provide the output, information processing, and/or communication operations.

Some embodiments may be provided by two or more servers and/or two or more computer devices.

It should be understood that each block of the flowchart, of the Figures and any combination thereof may be implemented by various means or their combinations, such as hardware, software, firmware, one or more processors and/or circuitry.

It is also noted herein that while the above describes example embodiments, there are several variations and modifications which may be made to the disclosed solution without departing from the scope of the present invention.

In general, the various embodiments may be implemented in hardware or special
5 purpose circuits, software, logic or any combination thereof. Some aspects of the invention may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device, although the invention is not limited thereto. While various aspects of the invention may be illustrated and described as
10 block diagrams, flow charts, or using some other pictorial representation, it is well understood that these blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

15 The embodiments may be implemented by computer software executable by a data processor, or by hardware, or by a combination of software and hardware. Computer software or program, also called program product, including software routines, applets and/or macros, may be stored in any apparatus-readable data storage medium and they comprise program instructions to perform particular tasks. A
20 computer program product may comprise one or more computer-executable components which, when the program is run, are configured to carry out embodiments. The one or more computer-executable components may be at least one software code or portions of it.

Further in this regard it should be noted that any blocks of the logic flow as in the
25 Figures may represent program steps, or interconnected logic circuits, blocks and functions, or a combination of program steps and logic circuits, blocks and functions. The software may be stored on such physical media as memory chips, or memory blocks implemented within the processor, magnetic media such as hard disk or floppy disks, and optical media such as for example DVD and the data variants
30 thereof, CD. The physical media is a non-transitory media.

The memory may be of any type suitable to the local technical environment and may be implemented using any suitable data storage technology, such as semiconductor based memory devices, magnetic memory devices and systems, optical memory devices and systems, fixed memory and removable memory. The data processors
5 may be of any type suitable to the local technical environment, and may comprise one or more of general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs), application specific integrated circuits (ASIC), FPGA, gate level circuits and processors based on multi core processor architecture, as non-limiting examples.

10 Embodiments of the inventions may be practiced in various components such as integrated circuit modules. The design of integrated circuits is by and large a highly automated process. Complex and powerful software tools are available for converting a logic level design into a semiconductor circuit design ready to be etched and formed on a semiconductor substrate.

15 The foregoing description has provided by way of non-limiting examples a full and informative description of the exemplary embodiment of this invention. However, various modifications and adaptations may become apparent to those skilled in the relevant arts in view of the foregoing description, when read in conjunction with the accompanying drawings and the appended claims. However, all such and similar
20 modifications of the teachings of this invention will still fall within the scope of this invention as defined in the appended claims. Indeed there is a further embodiment comprising a combination of one or more embodiments with any of the other embodiments previously discussed.

CLAIMS

1. A computer implemented method comprising:
analysing first data to determine itemsets that occur frequently within the first data, said first data comprising a plurality of sets of second data comprising items,
5 said itemsets comprising two or more of said sets of second data;
processing said itemsets to determine frequently occurring itemsets and to provide a set of filtered relatively frequently occurring itemsets; and
using said set of filtered relatively frequent itemsets to provide an output.
- 10 2. A method according to claim 1, wherein said processing comprises removing one or more relatively frequently occurring itemsets from said set of relatively frequently occurring itemsets.
3. A method as claimed in claim 1 or 2, wherein said processing comprises
15 applying a statistical function to said itemsets.
4. A method as claimed in claim 3, wherein said applying a statistical function to said itemsets comprises assigning a respective score to a respective itemset.
- 20 5. A method according to any preceding claim, wherein processing comprises determining a statistical function with respect to a key performance indicator associated with a plurality of sets of second data over all sets of second data associated with said key performance indicator with respect to a respective itemset.
- 25 6. A method according to claim 5, wherein determining the statistical function comprises one or more of determining an average and a standard-deviation of the key performance indicator over all sets of second data with respect to a respective itemset.
- 30 7. A method as claimed in any preceding claim, wherein said processing comprises determining if a respective itemset is a core itemset, said core itemset

being one where there is not subset of that itemset having a score value which is greater than or equal to a score value of the itemset multiplied by a significance ratio.

8. A method as claimed in any preceding claim, wherein said processing
5 comprises determining if a respective itemset is a tightened itemset, if for each core
superset of the respective itemset, associated support information of respective
itemset less the associated support information of the superset is greater than a
threshold amount.
9. A method as claimed in any preceding claim, wherein said processing
10 comprises using support information for a itemset, said support information for a
respective itemset providing information about the number of sets of said second
data comprising said respective itemset.
10. A method as claimed in claim 7, 8 and 9, wherein said processing comprises
15 determining if a respective itemset is a tightened area itemset, where statistics of a
respective itemset is determined without sets of second data of a superset group of
data, wherein respective itemset is determined to be a tightened area subset if said
respective itemset is also a core itemset and a tightened itemset, according to the
said statistics of the respective itemset.
- 20
11. A method as claimed in any preceding claim, wherein said processing is
controlled by one or more parameters, said one or more parameters comprising an
improvement-ratio, a significance-ratio, and an improvement/significance-change.
- 25
12. A method according to any preceding claim, wherein providing an output is
based on at least one of one or more specified key performance indicators, items
with a specified item of interest, and itemsets, using an analysis technique, wherein
the analysis technique is at least one of pattern recognition and statistical analysis.
- 30
13. A method, according to any preceding claim, comprising:

receiving a query, the query comprising one or more of key performance indicators, information about one or more data sources, and at least one item of interest, said using of said set of filtered relatively frequent itemsets to provide said output being dependent on said query

5

14. A method as claimed in any preceding claim comprising obtaining first source data, from a first source and second source data from at least one second source, said first source data and said second source data comprising said first data, wherein at least one of said sets of second data comprises first source data and
10 second source data.

15. A method according to any preceding claim, wherein said first data comprises communication data and said output comprises information indicating one or more network parameters related to network performance.

15

16. A method as claimed in claim 15, wherein said communication data comprises one or more of subscriber data; application data; user device data; network operator data and internet data.

20

17. A computer program comprising computer executable code which when run on at least one processor is configured to cause the method of any one of the preceding claims to be performed.

25

18. An apparatus configured in use to cause the method of any one of the preceding claims to be performed.

19. A computer apparatus, said computer apparatus comprising:

at least one processor and at least one memory including computer code for one or more programs, the at least one memory and the computer code configured,
30 with the at least one processor, to cause the apparatus at least to:

analyse first data to determine itemsets that occur frequently within the first data, said first data comprising a plurality of sets of second data

comprising items, said itemsets comprising two or more of said sets of second data;

process said itemsets to determine frequently occurring itemsets and to provide a set of filtered relatively frequently occurring itemsets; and

5 use said set of filtered relatively frequent itemsets to provide an output.

Figure 1

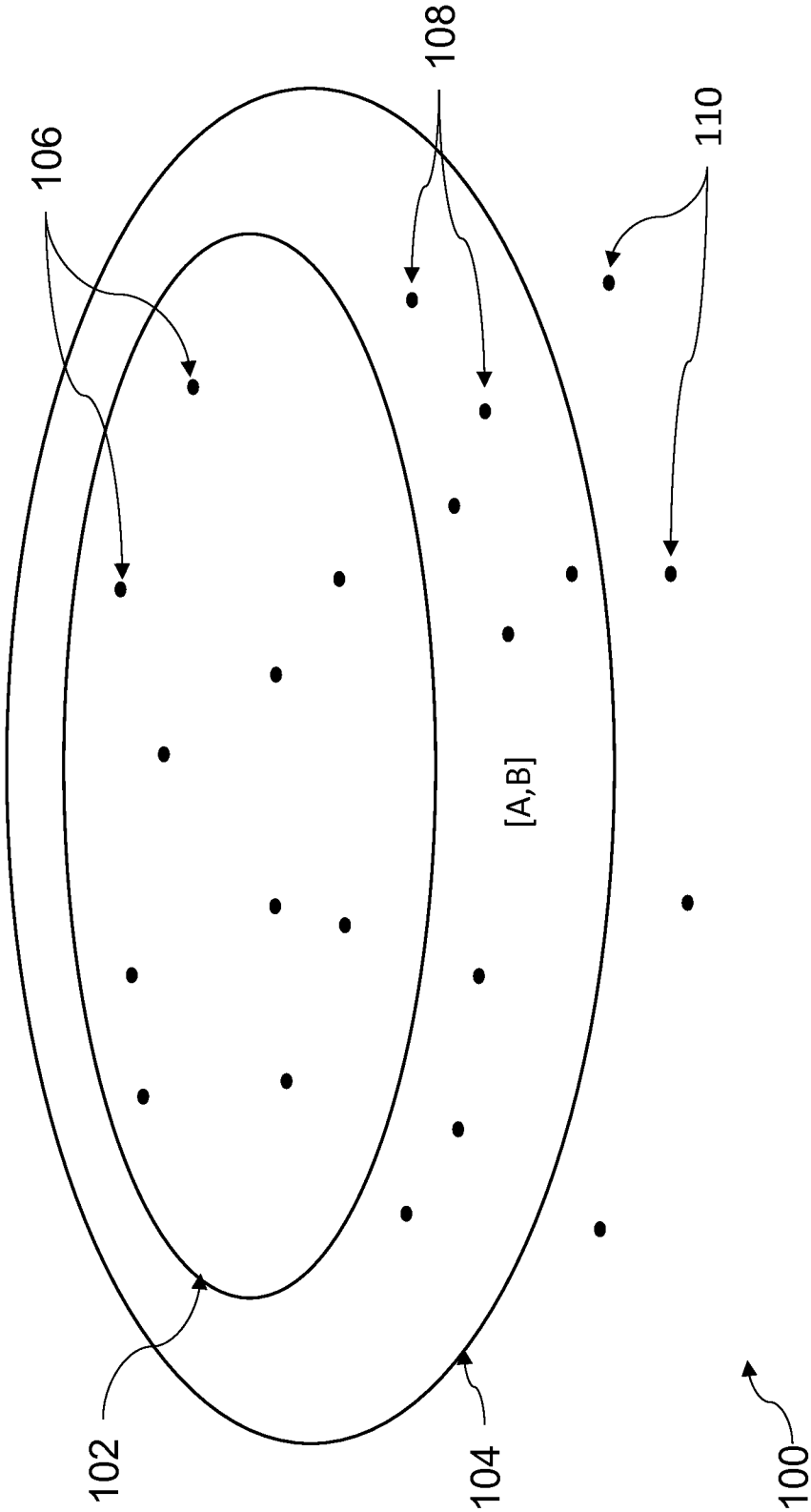


Figure 2

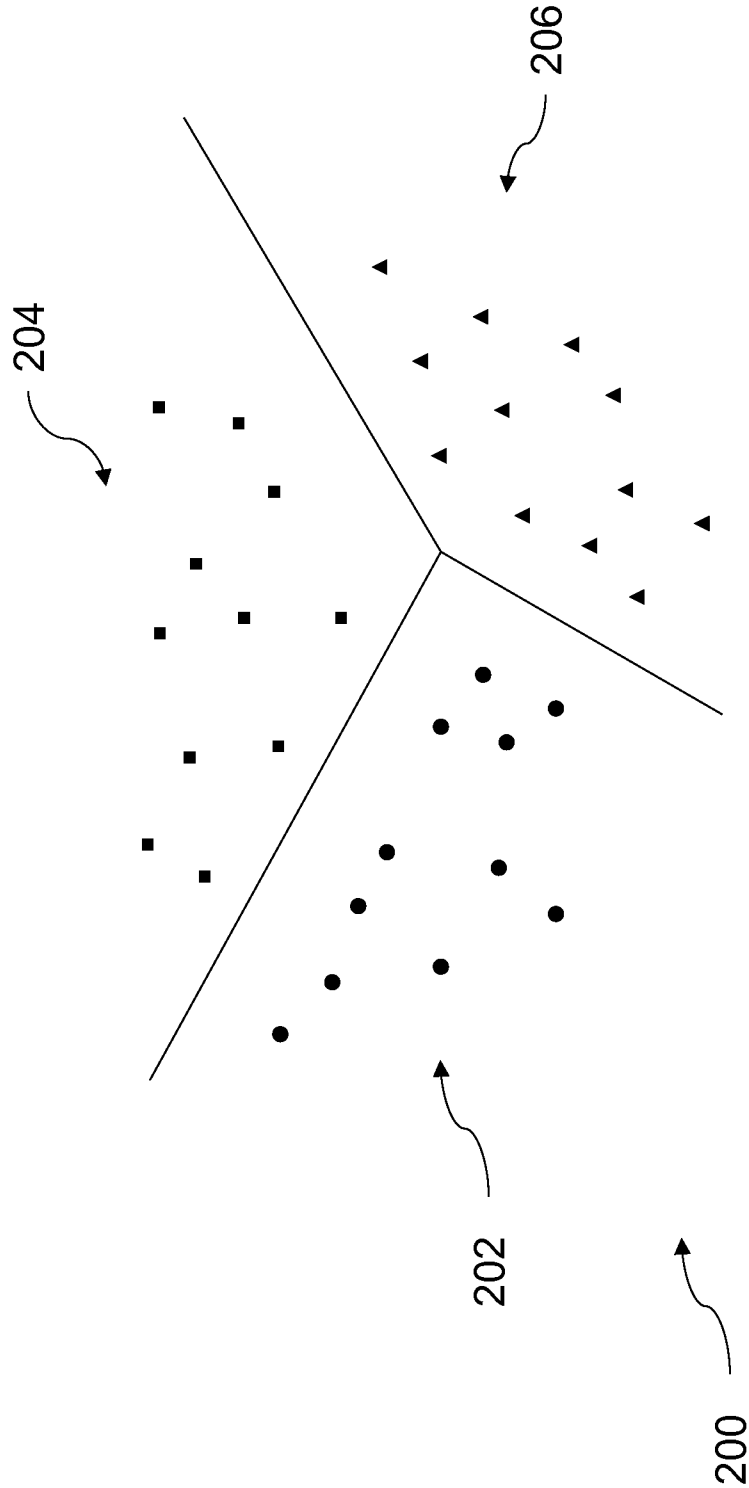


Figure 3

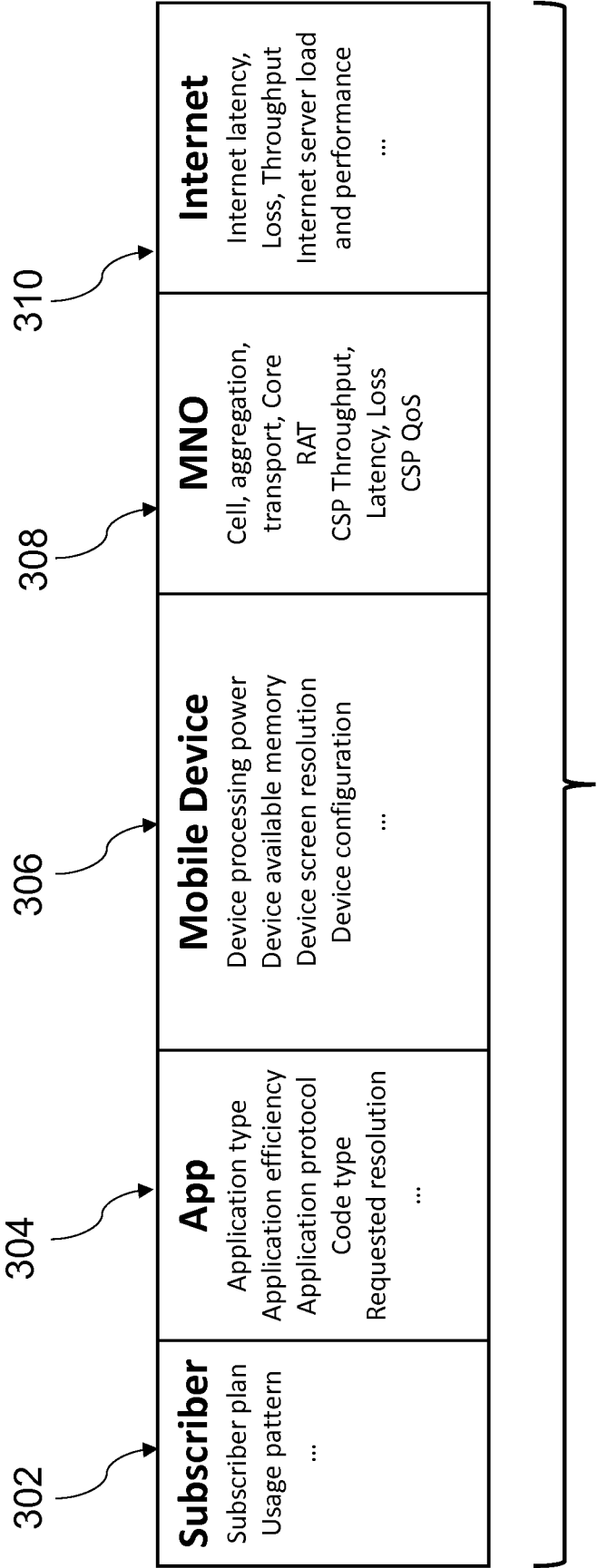
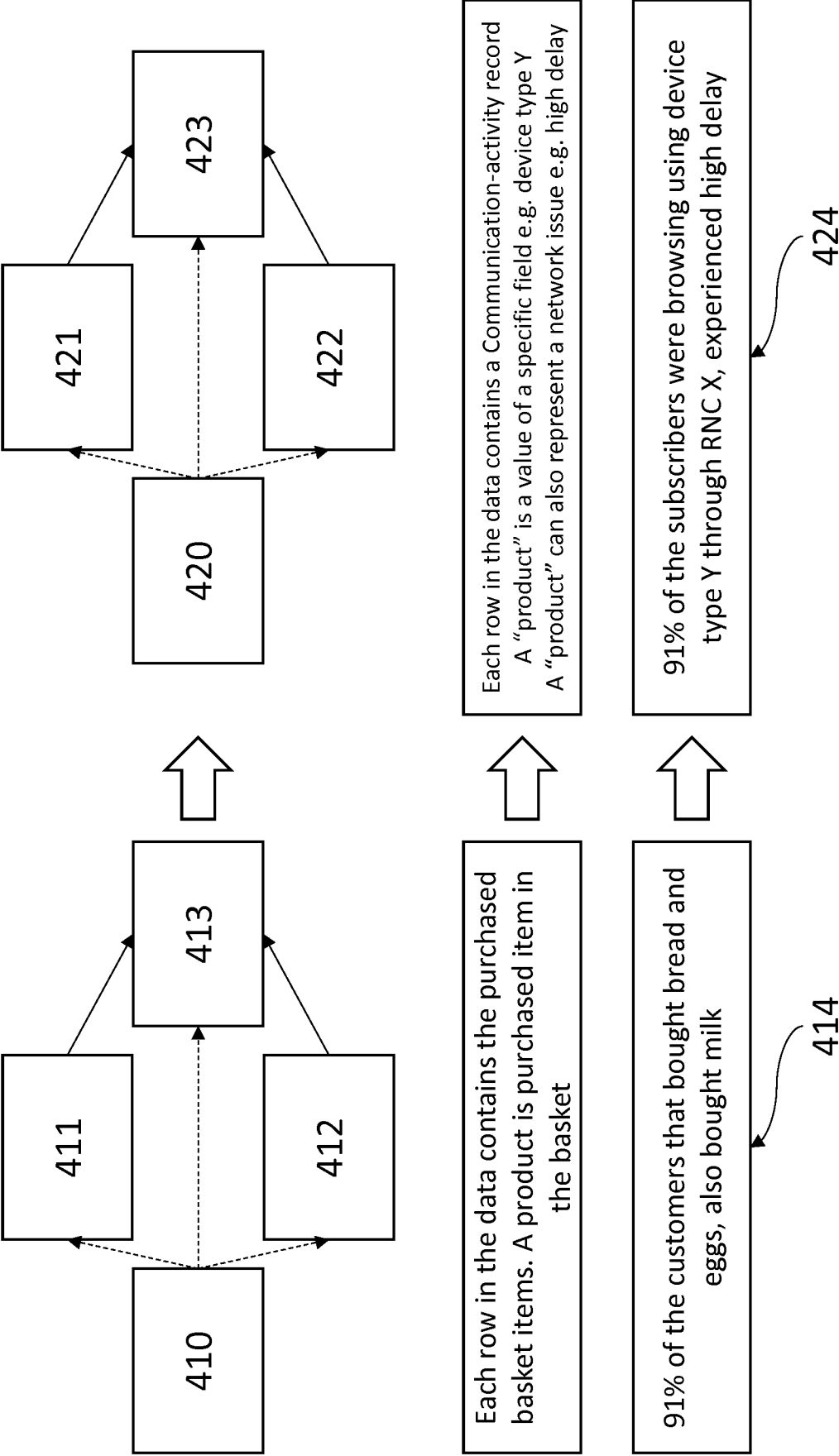


Figure 4



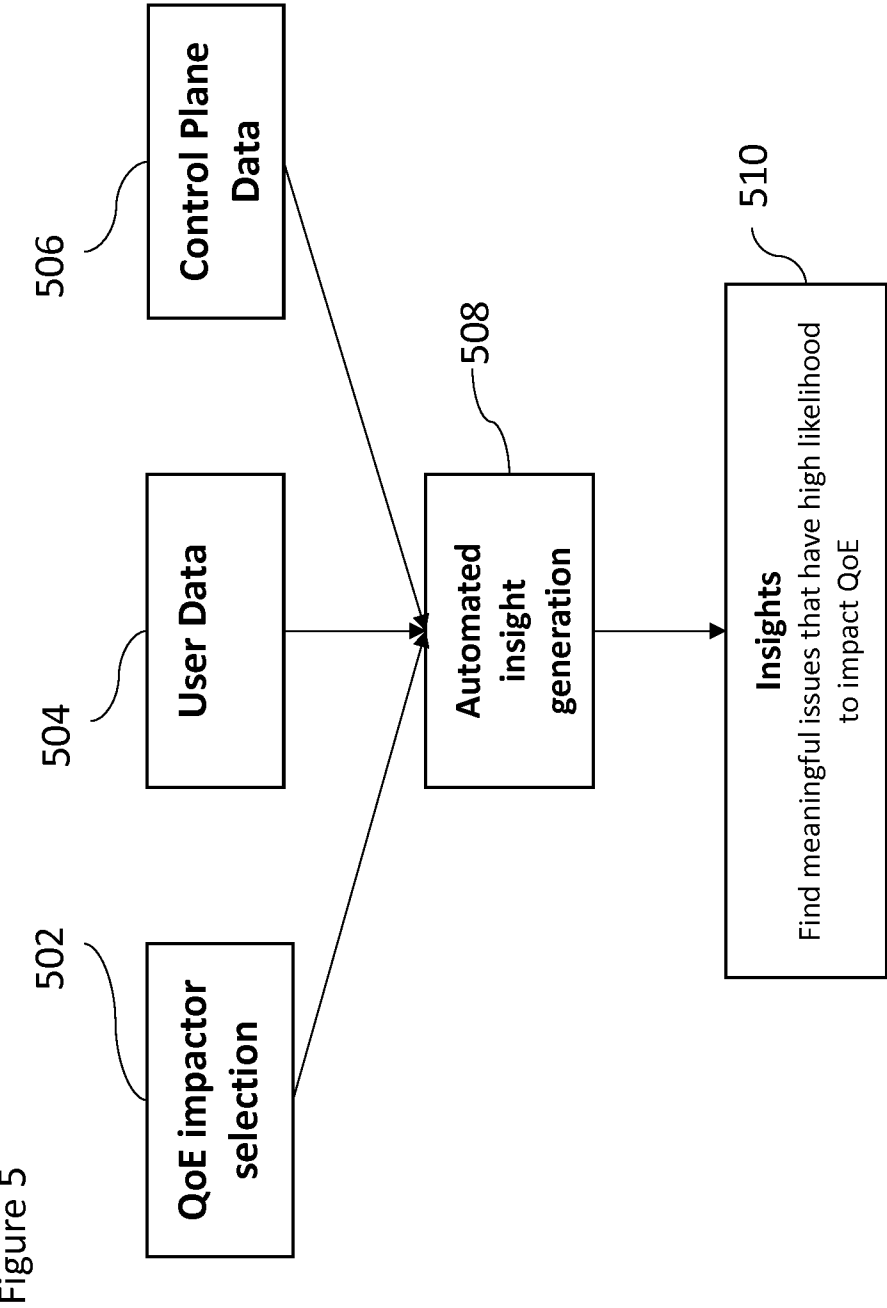


Figure 6

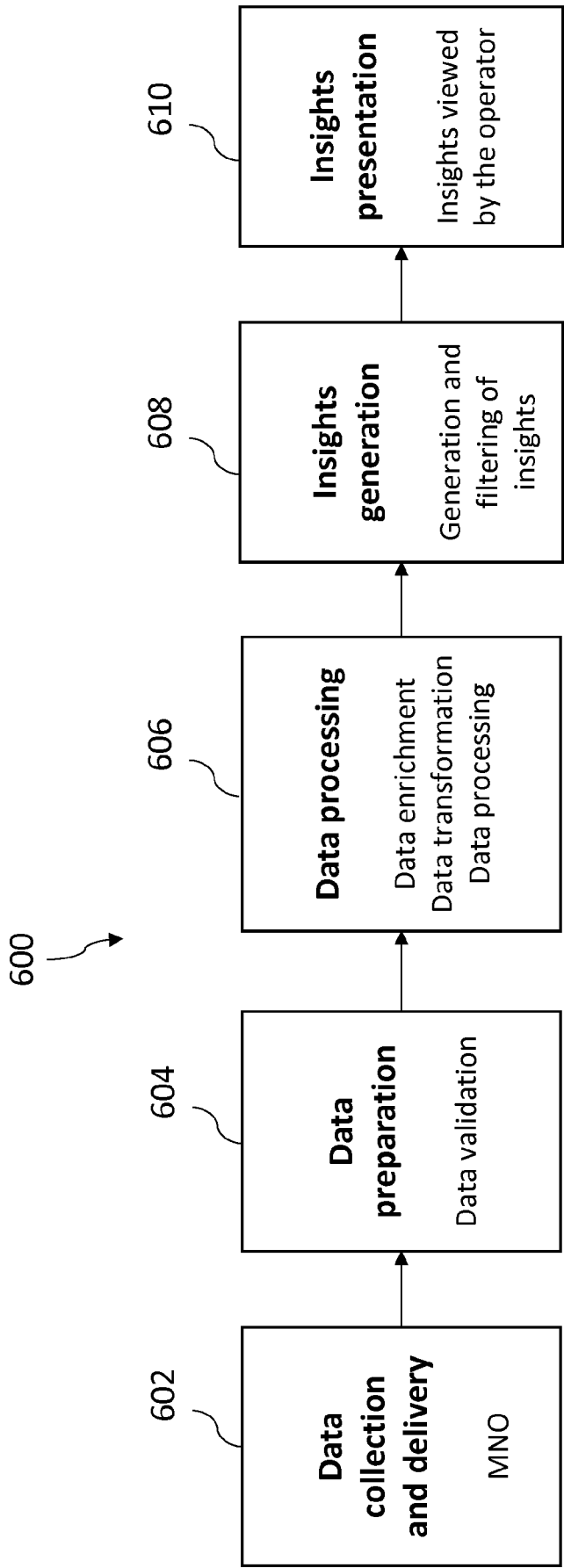


Figure 7

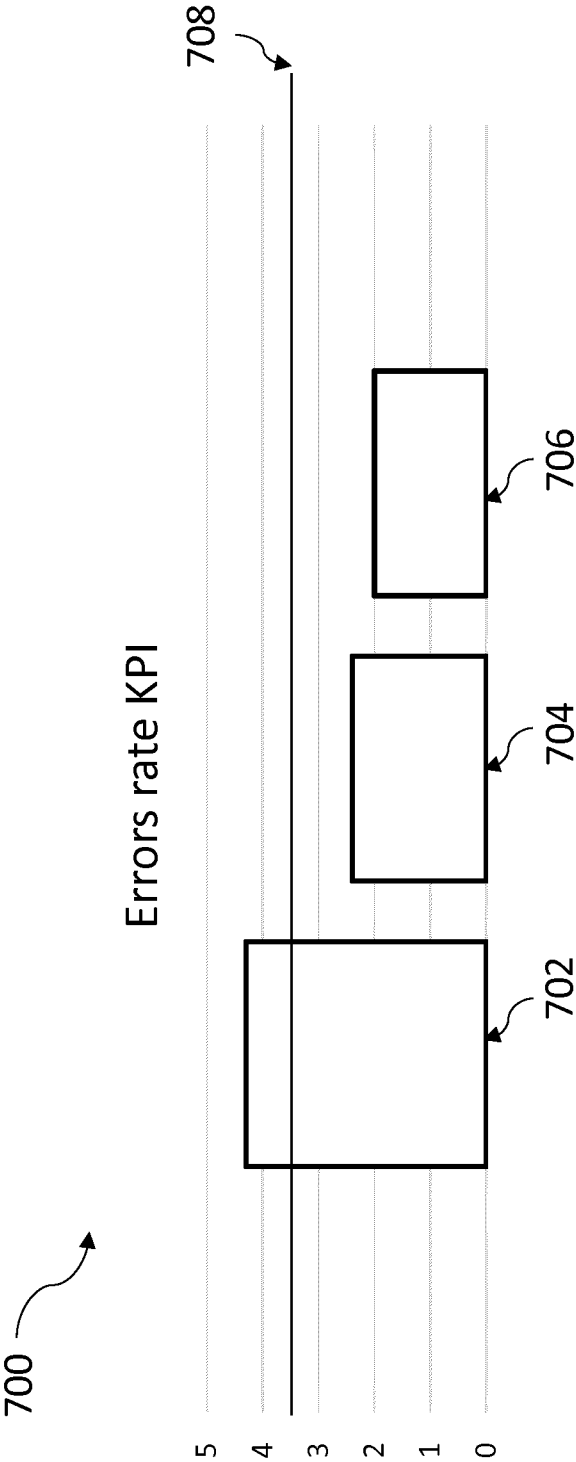
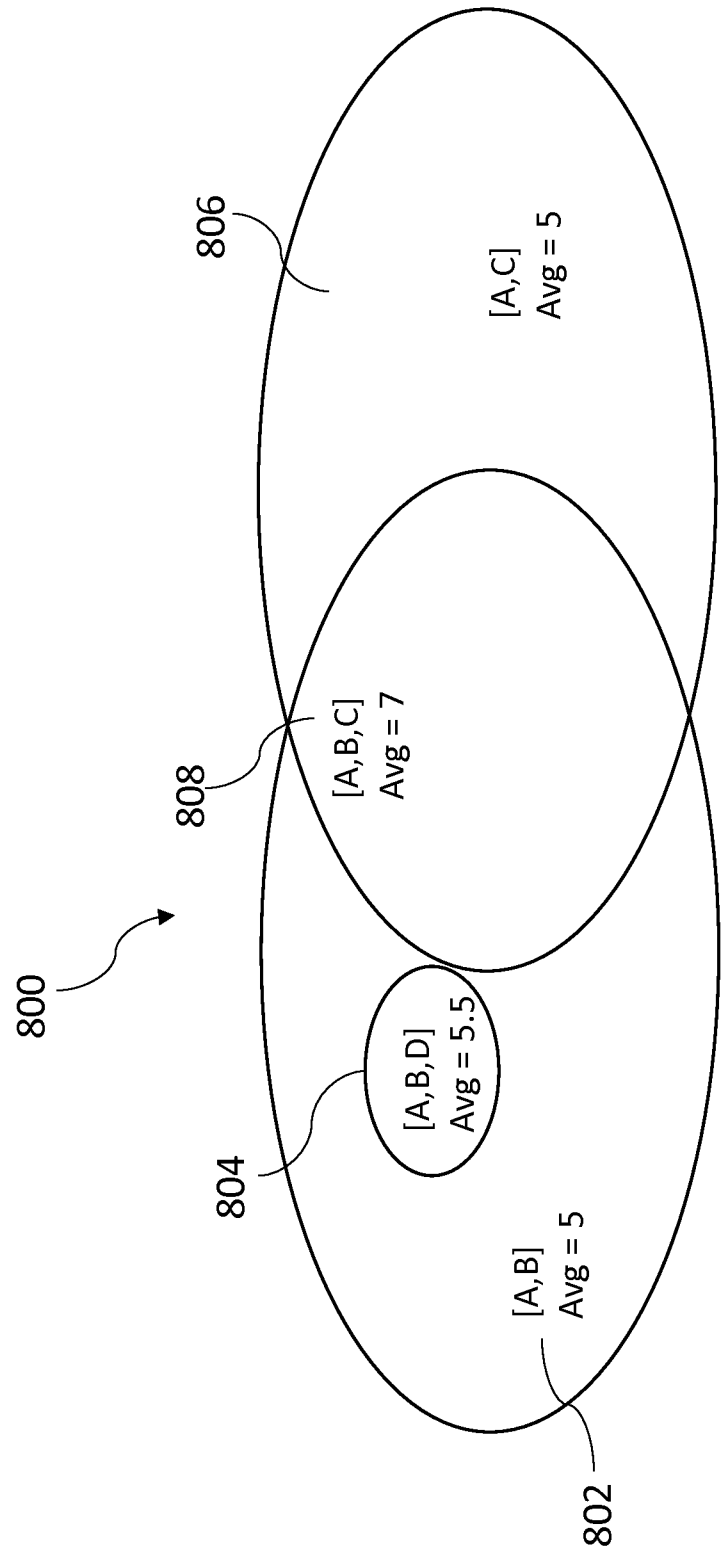


Figure 8



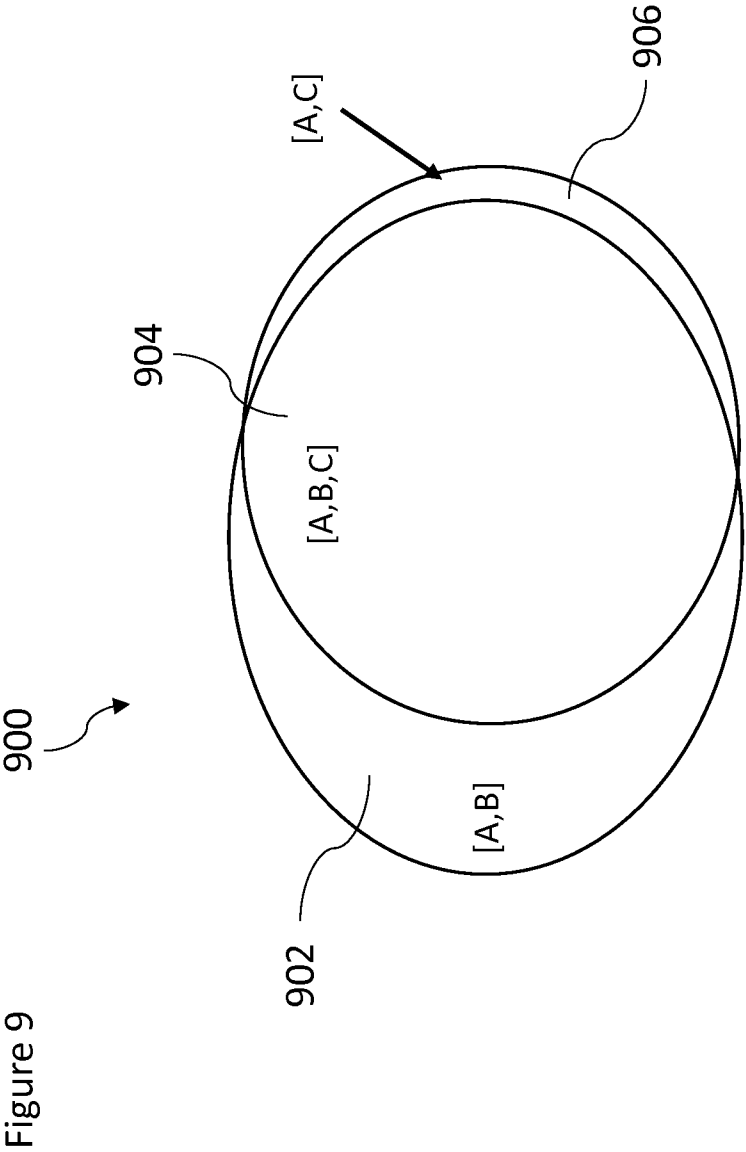


Figure 10

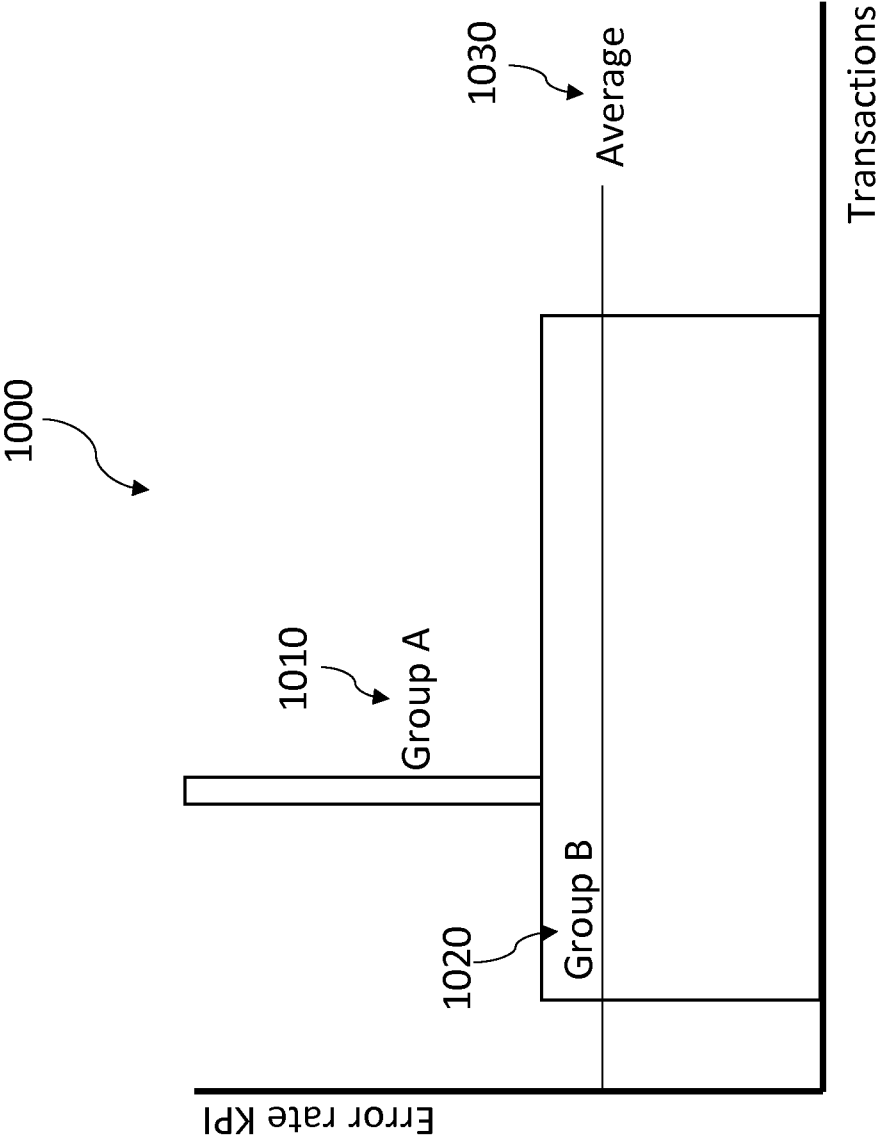
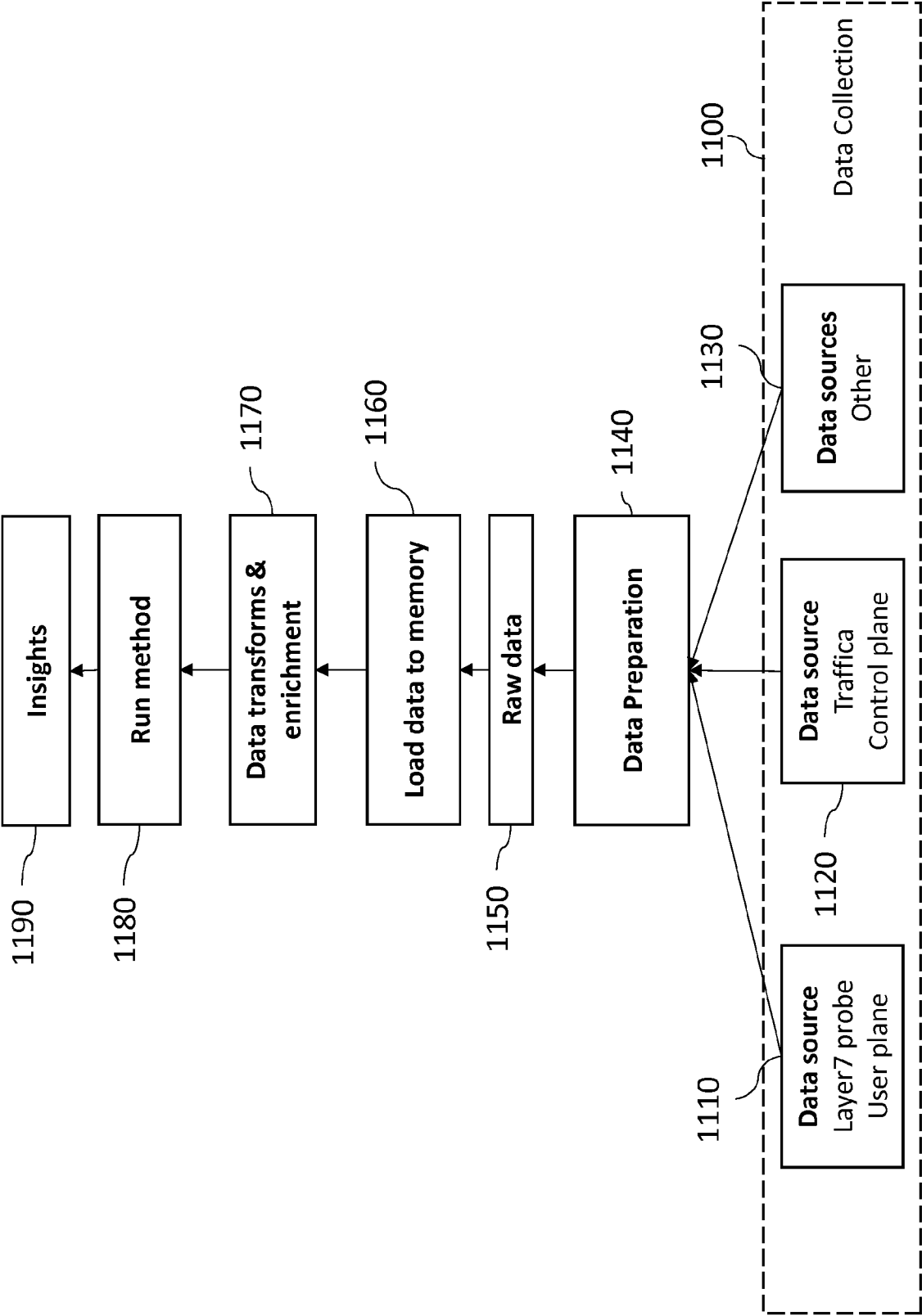


Figure 11



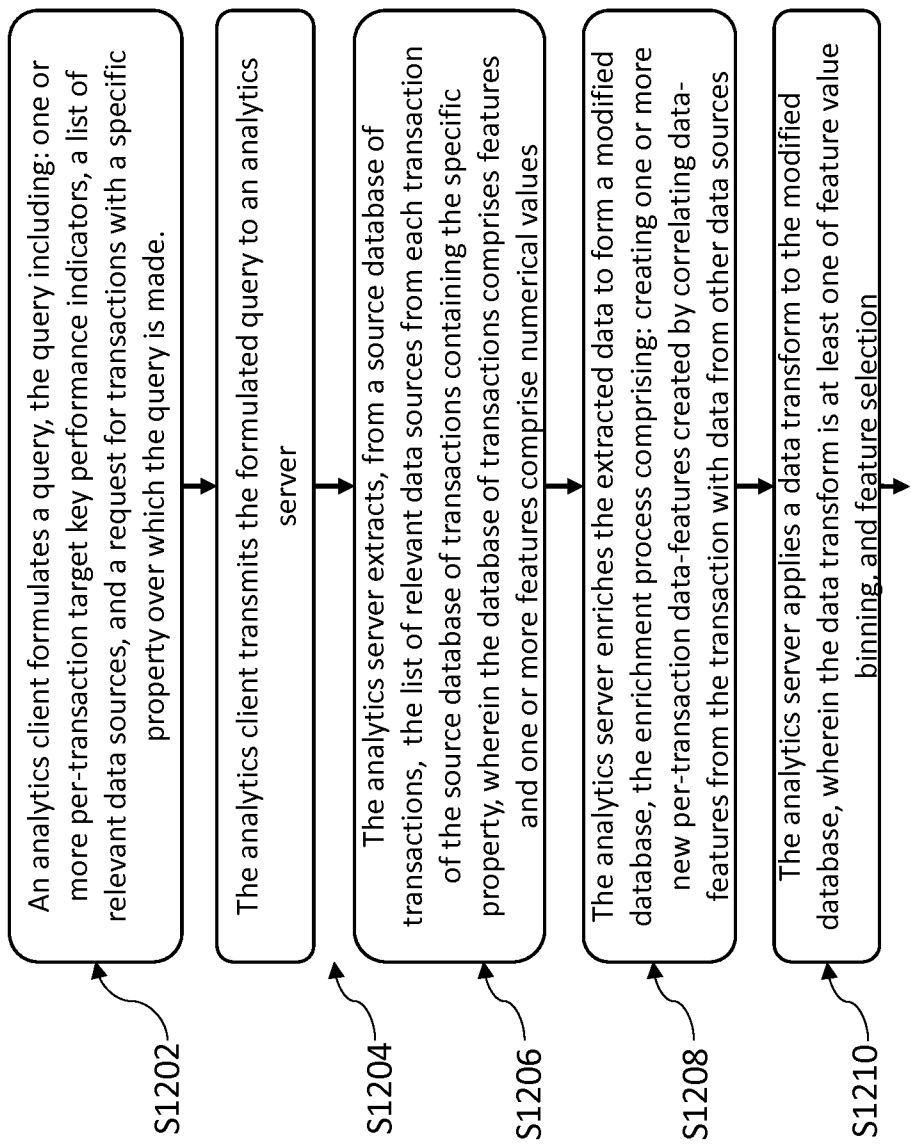


Figure 12a

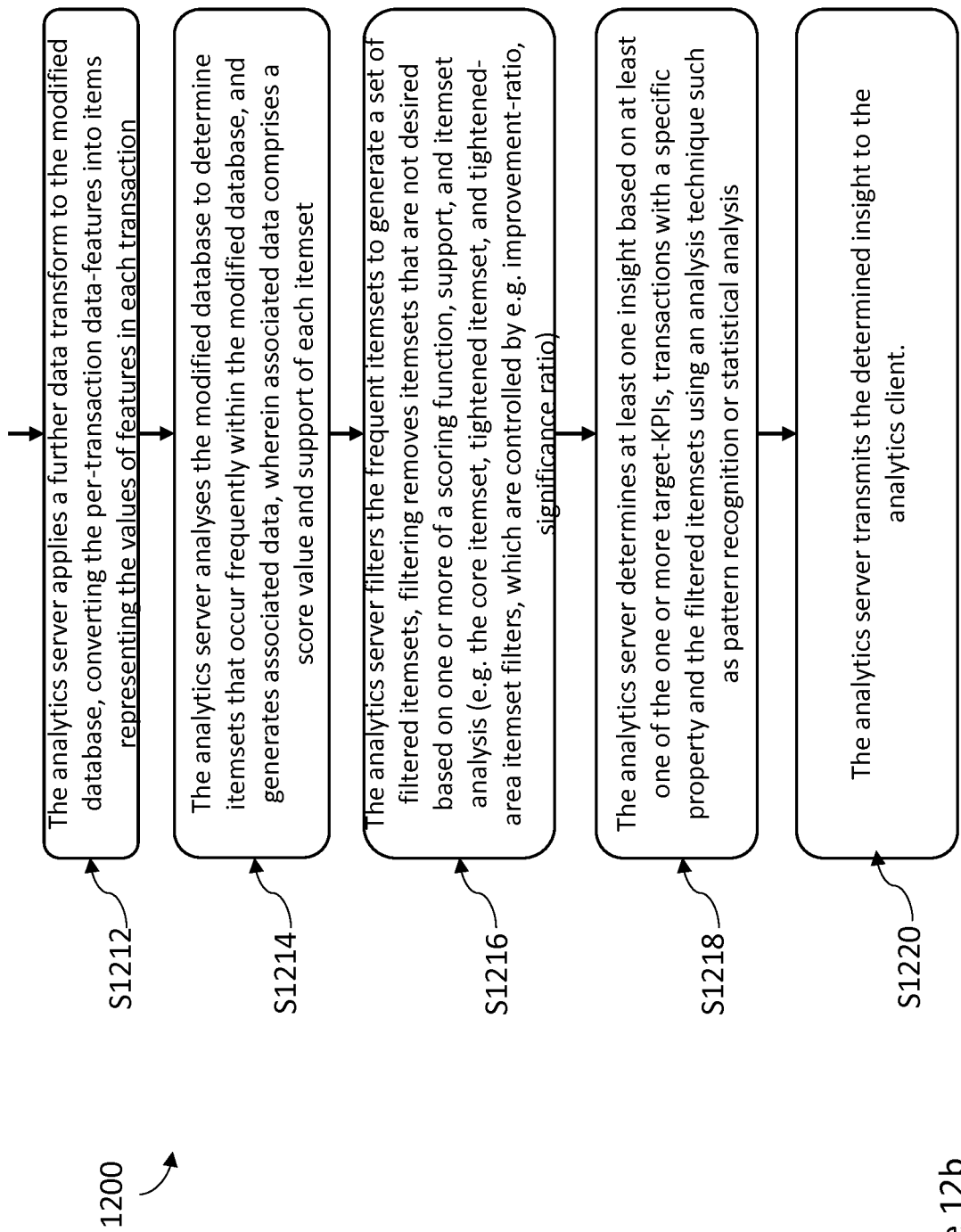


Figure 12b

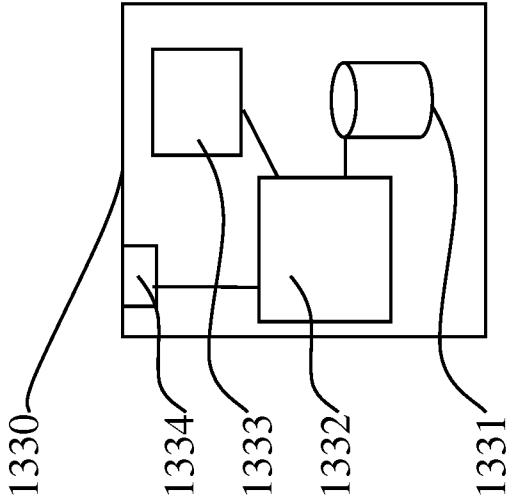


Figure 13

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/058671

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F17/30 G06N5/02
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Kim Ervasti: "Correlation discovery from network monitoring data in a big data cluster", 31 December 2014 (2014-12-31), pages 1-9, XP055374625, Retrieved from the Internet: URL: https://www.semanticscholar.org/paper/Correlation-discovery-from-network-monitoring-data-Ervasti/ba6e3227d6e7cf3ad67d70b416c8f70f1600a860 [retrieved on 2017-05-19] abstract page 2, left-hand column, last paragraph page 2, right-hand column, line 1, paragraph 2 - line 5 page 4, left-hand column, line 2, last paragraph - line 4 page 4, left-hand column, line 14 - line -/--	1-19



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

6 June 2017

Date of mailing of the international search report

14/06/2017

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Rameseder, Jonathan

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/058671

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	17 page 4, right-hand column, line 11 - line 25 page 5, left-hand column, line 1 - line 3 page 2, left-hand column, line 15 - line 19 page 5, left-hand column, line 17, paragraph 2 - line 20 table 1 figure 3 ----- L Anonymous: "Overview page: Correlation discovery from network monitoring data in a big data cluster", Semantic Scholar, 31 December 2014 (2014-12-31), XP055375154, Retrieved from the Internet: URL:https://www.semanticscholar.org/paper/Correlation-discovery-from-network-monitoring-data-Ervasti/ba6e3227d6e7cf3ad67d70b416c8f70f1600a860 [retrieved on 2017-05-22] the whole document ----- X LIU JUN ET AL: "Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop", IEEE NETWORK, IEEE SERVICE CENTER, NEW YORK, NY, US, vol. 28, no. 4, 1 July 2014 (2014-07-01), pages 32-39, XP011554326, ISSN: 0890-8044, DOI: 10.1109/MNET.2014.6863129 [retrieved on 2014-07-23] the whole document ----- X YANGYANG WU ET AL: "Mining Alarm Database of Telecommunication Network for Alarm Association Rules", DEPENDABLE COMPUTING, 2005. PROCEEDINGS. 11TH PACIFIC RIM INTERNATIONAL SYMPOSIUM ON CHANGSHA, HUNAN, CHINA 12-14 DEC. 2005, PISCATAWAY, NJ, USA, IEEE, 12 December 2005 (2005-12-12), pages 281-286, XP010902839, DOI: 10.1109/PRDC.2005.40 ISBN: 978-0-7695-2492-4 the whole document ----- -/--	1-19 1-19 1-19

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2017/058671

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WU JIAN ET AL: "A dynamic mining algorithm of association rules for alarm correlation in communication networks", COMMUNICATION SYSTEMS SOFTWARE AND MIDDLEWARE AND WORKSHOPS, 2008. COMSWARE 2008. 3RD INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 6 January 2008 (2008-01-06), pages 799-802, XP031279784, ISBN: 978-1-4244-1796-4 the whole document -----	1-19
X	CN 105 827 422 A (CHINA MOBILE GROUP SHANGHAI CO) 3 August 2016 (2016-08-03) the whole document -----	1-19
A	US 2013/204830 A1 (FRANKE DAVID [US]) 8 August 2013 (2013-08-08) the whole document -----	1-19
A	US 6 278 997 B1 (AGRAWAL RAKESH [US] ET AL) 21 August 2001 (2001-08-21) the whole document -----	1-19

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2017/058671

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
CN 105827422	A	03-08-2016	NONE
US 2013204830	A1	08-08-2013	US 8401986 B1 19-03-2013
		US 2013204830 A1	08-08-2013
US 6278997	B1	21-08-2001	NONE