Paper

Alarm Correlation in Mobile Telecommunications Networks based on *k*-means Cluster Analysis Method

Artur Maździarz

Systems Research Institute, Polish Academy of Science, Warsaw, Poland

https://doi.org/10.26636/jtit.2018.124518

Abstract— Event correlation and root cause analysis play a fundamental role in the process of troubleshooting all technical faults and malfunctions. An in-depth, complicated multiprotocol analysis can be greatly supported or even replaced by a troubleshooting methodology based on data analysis approaches. The mobile telecommunications domain has been experiencing rapid development recently. Introduction of new technologies and services, as well as multivendor environment distributed across the same geographical area create a lot of challenges in network operation routines. Maintenance tasks have been recently becoming more and more complicated, time consuming and require big data analyses to be performed. Most network maintenance activities are completed manually by experts using raw network management information available in the network management system via multiple applications and direct database queries. With these circumstances considered, identification of network failures is a very difficult, if not an impossible task. This explains why effective yet simple tools and methods providing network operators with carefully selected, essential information are needed. Hence, in this paper efficient approximated alarm correlation algorithm based on the *k*-means cluster analysis method is proposed.

Keywords— alarm correlation, alarm patterns, cluster analysis, mobile telecommunication network, root cause analysis.

1. Introduction

The history of mobile telecommunication started in the late 1970s, when analog telephony standards were introduced to cover basic voice calls. The entire family of these analog systems is referred to as 1G. In the 1990s, the digital age of mobile communication began along with the introduction of the so-called 2G technology. Technology development, driven by moving towards mobile data transfers with ever higher speeds, resulted in the introduction of 2.5G (GPRS), 3G and 4G/LTE standards. Currently, the telecommunication community is working on the development and introduction of the 5G standard, which is supposed to be ready for use by 2020 [1].

The generic diagram a mobile telecommunication network is presented in Fig. 1.

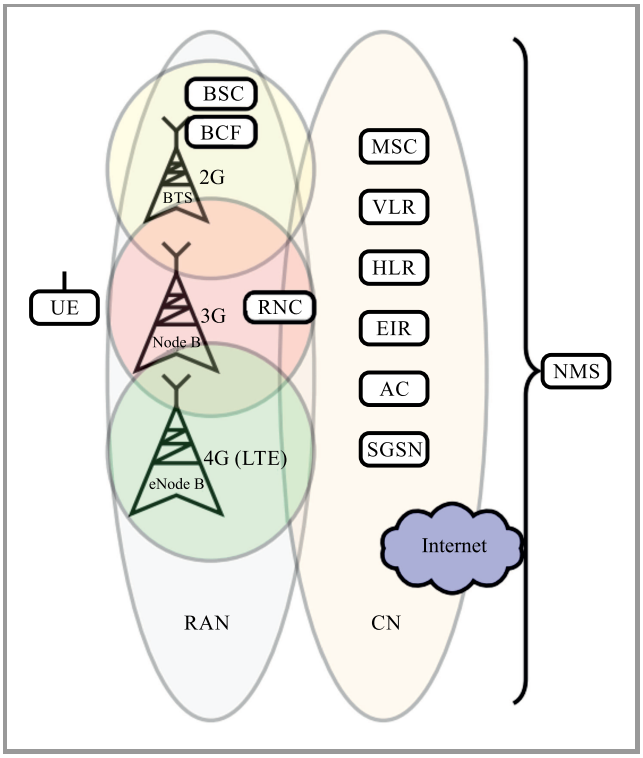

Fig. 1. Typical architecture of a mobile telecommunication network.

A mobile telecommunication network consists of two major functional subsystems: the radio access network (RAN) and the core network (CN).

RAN is responsible for managing radio resources, including strategies and algorithms for controlling power, channel allocation and data rate. It allows the user terminal equipment (UE) to access network services. The RAN consists of the following elements, depending on the technology used: 2G base station controller (BSC), 3G radio network controller (RNC), base station control function (BCF), 2G base transceiver station (BTS), 3G base transceiver station (NodeB), enhanced node B, 4G base transceiver station (eNodeB) and transceiver (TRX).

The CN is mainly responsible for high-level traffic aggregation, routing, call control/switching, user authentication

and charging. Some of the CN subsystems are: 2G, 3G mobile switching center (MSC), 2G, 3G visitor location register (VLR), 2G, 3G home location register (VLR), 2G, 3G authentication center (AC), 2G, 3G equipment identity register (EIR), 2G, 5G service GPRS support node (SGSN) [1]–[3].
The entire network is managed by the network management system, the so-called NMS, which provides several network management functionalities. One of the primary functions of the NMS is fault management. It is a term used in the network management domain, focusing on processes related to diagnosing and fixing network faults.
In the paper, we propose the approximated network faults diagnosing methodology based on the cluster analysis *k*-means algorithm.
The paper is organized as follows: In Section 2 we briefly introduce the Network Fault Management domain. Section 3 introduces novel alarm correlation methodology based on k-means clustering approach. Section 4 illustrates experiments and results achieved. Finally concluding remarks are given in Section 5.

2. Preliminaries and Problem Statement

The fault management domain of the network is characterized by a few definitions and notations that are central to this paper [4].

- **event** is an exceptional condition occurring in the operation of hardware or software within the network managed; an *instantaneous occurrence* at a time,
- **event correlation** is the process of establishing relationships between network events,
- **root causes**, are events that can cause other events but are not caused by other events; they are associated with an *abnormal state* of network infrastructure,
- **error** is a discrepancy between an observed or computed value or condition and a true value or condition, assumed to be correct,
- **failure** or **fault** is considered to be a kind of an error,
- **symptoms** are external manifestations of failures (errors) which are observed as alarms.

Fault diagnosis usually involves three processes: fault detection, fault localization (also known as fault isolation or root cause analysis) and testing the possible hypotheses [4].
Fault detection is the process of collecting information related to malfunctions of the network's components (network elements) in the form of alarms [4].
Fault localization or root cause analysis (RCA) is the process of identifying the causes of faults. It comprises several stages of correlating events (including alarms) which occurred over a certain period of time, and requires technical knowledge about the system analyzed [4], [5].
Alarm correlation is the process of grouping alarms which refer to the same problem, in order to highlight those which indicate the possible root cause [6].
The advantages of automating RCA and alarm correlation routines are numerous. By automating the troubleshooting process, we shorten the time needed for identifying a potential source of the problem, which impacts the duration of downtimes and quality of service (QoS) figures for the network in question. Short troubleshooting times bring benefits in the form of satisfying the terms of customers' service level agreements (SLAs). In addition, less skilled personnel can be involved in network operation tasks, thus reducing network maintenance costs [7].
There are several root cause analysis techniques described in the literature. We can divide them into three major categories: artificial intelligence techniques, model traversing and the so-called fault propagation model techniques [4]. All techniques are based either on predefined expert system knowledge, network static information or network dynamic information. The static knowledge comes from the topology and system structure. The dynamic network information is connected with the functional behavior of the network [4], [8], [9]. The methodology proposed in this paper helps discover relations between alarms generated by the network, thus contributing to analysis of static and dynamic network characteristics in the RCA process.
The amount of data to be analyzed and the limited analysis lead time pose a major challenge while troubleshooting faults in such a complex system like a telecommunication network. These two factors play a key role in fast problem resolution and minimize consequences for end users. The volume of troubleshooting data processed during propagation of faults in a large network can easily exceed several dozens of alarms per second. For those faults that impact the usability of the network by considerable amounts of end users, the resolution time is crucial and has a big financial impact on the service provider. To cope with the problem referred to above, the data correlation methodology should be characterized by fast processing, as well as by easy interpretation and reliable quantification of the results.
Medium size mobile telecommunication networks consist of several thousand of network elements, including RAN and CN subsystems. With all functional dependencies between network elements taken into consideration, the entire network is very complex. There are a lot of network elements, and each of them can potentially generate alarms. As per the fault management objective, network alarms collected by NMS should be correlated and the potential root cause of the problem should be identified within a short time. Analysis of alarm symptoms which leads to discovering the root cause of the problem is covered by the alarm correlation and root cause analysis processes. This paper focuses on the alarm correlation process which works on the alarm data sets. Each raw alarm data record contains several alarm attributes:

- **time of alarm**, this attribute contains the date and time with the precision of one second,

- **alarm number**, a unique number which identifies the fault. Usually the alarm numbers are divided into ranges representing a specific subsystem, network element type and alarm type,
- **alarm type** can be specified as communication, or for example equipment type,
- **alarm description** inside the alarm frame is a very short, compact description of the fault that usually contains a brief description (a few words) of what has happened,
- **alarm severity** specifies the importance of the fault and describes the alarm class. It can take one of the following logical values: critical, major, medium, minor or warning,
- **name of the object** is the object identification label which clearly identifies the network element which has generated the alarm event,

In the case of a fault of a specific network element, the alarm rate can reach several dozen alarms per second. Usually, failures related to one network element cause other network components to send relevant alarms as well. Additional troubleshooting difficulty in a complex system like a mobile telecommunication network stems from the number of network elements, as well as from their geographical distribution. In the attached example a set of BTSes connected to the BSC via BCFs is considered. The transmission problem related to the connection between the BSC and BCFs generates several alarms from BCFs and BTSes. The example shows how one problem triggers a string of alarms for all related network elements. If outage of critical network elements occurs, the network management system is flooded by large quantities of alarms. In these conditions, the operator has very limited time to diagnose what and where has happened. This is the reason behind the need to develop fast and simple methods to deal with big amount of symptom-describing data (alarms). It is worth mentioning that apart from the fast alarm correlation methodology, the additional goal is to work on reducing the amount of data (alarms) which are being analyzed. This is achieved by identifying repeatable alarm patterns which can be analyzed as one atomic entity to simplify the correlation process and to reduce amount of data to be processed.
In the following section a methodology is proposed which addresses most of the abovementioned challenges involving the correlation of alarms in mobile telecommunication networks.

3. Proposed Methodology Approach using Cluster Analysis in RCA

There are several RCA methods proposed in the literature which relate to the subject of correlating alarm symptoms. In general, the methods are complicated and difficult to be implemented in practice. Therefore, this approach to alarm correlation is fast and practical.
As mentioned in Section 2, each alarm has six major attributes: occurrence time, number, description, type, severity, name of the alarming object (network element). All attributes can be used in the RCA process. The most important alarm attribute, which plays a fundamental role in troubleshooting, is alarm occurrence time. It is the main factor used for alarm correlation in this proposal.
In the approach presented, the alarm correlation methodology focuses on discovering, within the alarm data set, those which occurred within a short period of time. Hence, in this paper we will use the cluster analysis domain, assuming alarm occurrence as the clustering attribute. Practice shows that alarms which represent causal sequences of events may be grouped into clusters with limited time intervals. The alarm clusters identified constitute an alarm correlation hypothesis, which should be further analyzed by domain experts. Apart from the correlation of alarms, the goal is to find the root cause of the sequence of clustered alarms. Practice shows that the first alarm in the cluster (based on the occurrence time) is usually the root cause. It may happen that multiple incidents occur within the same time interval. In such a case, it is always the expert's role to evaluate the alarm clusters and to validate the alarm correlation hypothesis proposed.
The nature of the alarm flow reflects certain physicality of the incident within the network. The alarms which are related are either collected at the same time or are generated by network elements with a certain delay. In practice, it has been observed that correlated alarms can occur within intervals of 1–2 s. In the light of the above, it is essential to establish a fast methodology for discovering, within alarm data sets, alarms clusters characterized by the difference between alarm occurrence of approximately 2 seconds. Hence, we define the correlation criterion as the interval between the occurrence of alarms within the cluster.
Figure 2 illustrates two alarm cluster examples. The first cluster includes three alarm events $\{e_1, e_2, e_3\}$ that occurred at the same time, the second cluster consists of three alarm events $\{e_4, e_5, e_6\}$ which occurred sequentially, with a one second delay.

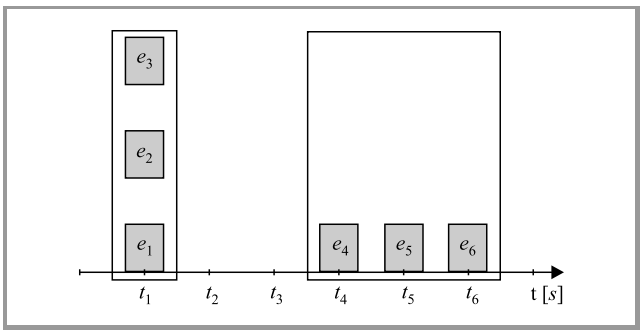

Fig. 2. Visualization of alarm correlation.

The cluster analysis domain offers techniques satisfying the objective of the method that consists in discovering clusters

of alarms. For the analysis, we selected the k-means clustering method for a filtered set of network elements known as a topology chain.

Each topology chain consists of network elements which share the same parent as the root of the topology. Typically, the roots are the main components of the network architecture and contain several child objects to perform the individual function. According to an alternative definition, the root object is the object which does not possess a parent, it is the first object in the hierarchy of a given type. An example of a topology chains is presented in Fig. 3.

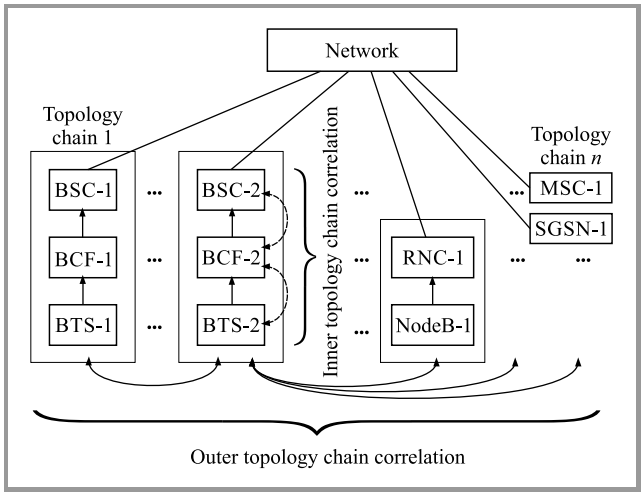

Fig. 3. Mobile telecommunication network topology and correlation view.

The term cluster analysis was used for the first time in 1954 in the context of analyzing anthropological data [10]. The k-means algorithm is recognized as the most important algorithm in the entire history of data mining [11], [12]. It represents the so-called combinatorial family of clustering algorithms. The cluster analysis, also known as classification without supervision, has two major characteristics. The clusters are unknown a priori and we do not dispose of the learning set. The goal of the analysis is to discover and group disjoint sets of data which are sharing similar characteristics (qualitative or quantitative features). In this type of analysis, the goal is to propose a data set split maximizing similarity features inside the sets and, at the same time, minimizing similarity between the disjoint sets. The same task can be translated into minimizing object dissimilarities inside the sets (clusters) and maximizing dissimilarities between sets. The cluster analysis process is based on the comparison of data set observations, resulting in generating groups of data which are more similar to each other within the group, than to objects from other groups (clusters). The popular methods of measuring dissimilarities, described in the literature concerned with cluster analysis, include the following: Euclidean distance, squared Euclidean distance, Minkowski distance, Mahalanobis distance, cosine distance and power distance [13]–[16]. In the approach presented, we analyze the time of occurrence of the alarms within the network, which is noted as: $X = \{x_1, \ldots, x_N\}$. The observations have labels assigned $i \in \{1, \ldots, N\}$. The squared Euclidean distance in the time domain is used as the correlation measure:

$$d(x_i, x_j) = \| x_i - x_j \|^2 . \tag{1}$$

The general principle of combinatorial clustering is based on the analysis of three characteristics: the total sum of dissimilarities between sample elements (T – total), the sum of dissimilarities between sample elements belonging to the same cluster (W – within), the sum of dissimilarities between sample elements belonging to different clusters (B – between).

The characteristics presented satisfy the following inquiry: $T = W + B$. For a given data set the value of T is constant and we target to minimize W or maximize B characteristics across all possible assignments of data set elements to the clusters [13], [16].

We denote dissimilarities between observations as $d(x_i, x_j)$ and we also define classificator $C(i)$, the function which based on the input maps the data to specific class, in our case the cluster. Classificator $C(i)$ returns cluster number ($k \in K$) for each observation i, j from the input data set.

Following the above notations, we can define W as [13], [16]:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} \| x_i - x_j \|^2 , \tag{2}$$

$$\overline{x}_k = \frac{1}{N_k} \sum_{C(j)=k} x_j , \tag{3}$$

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| x_i - \overline{x}_k \|^2 , \tag{4}$$

where: $\overline{x}_k$ is the mean vector associated with k-th cluster denoted as m_k and it is called centroid for cluster k, and N_k is the number of elements in cluster k.

Inquiry (4) serves as a basis for an entire family of algorithms referred to as k-means method algorithms.

The idea behind the k-means algorithm can be specified as follows [13]:

1. Propose clusters distribution determining means (centroids) of the clusters $\{m_1, \ldots, m_k\}$.
2. Assign the observations to the closest cluster based on its distance to the centroid.
3. Update the centroids based on the observations values assigned to the clusters.
4. Repeat steps 1–3 until centroids do not change and the observations do not change their assignments.

The steps referred to above accomplish the following optimization task which can be seen as a variance minimization task [17]:

$$\min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| x_i - m_k \|^2 . \tag{5}$$

An important note for this method is that we have to specify the number of clusters K in advance, and that the number of clusters we predefine should be lower than the number of elements in the sample N ($K < N$) [13], [16].
In this paper we focus on practical applications of the *k*-means method. As the *k*-means method requires specifying the number of clusters for the analysis, we perform the analysis by iterating the number of clusters K from 1 up to the value of $\overline{K} < N$. The proposed correlation methodology is based on applying the *k*-means iterative algorithm to pre-filtered data sets which represent the so-called topology chains and can be described by the following inquiry:

$$\sum_{\substack{topology\\chain}} \left(\max_{K/c\leq 3} \left(\min_{C,\{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| x_i - m_k \|^2 \right) \right). \quad (6)$$

In the proposed approach, we introduced an additional parameter which is used as the clustering criterion. It is the average Euclidean squared distance between the observations in cluster c. The coefficient c is expressed by the relation of within-cluster sum of squared distances between the observations (the average squared distance between the observations within the cluster in the time domain) to the number of observations in the cluster (*cluster_size*):

$$c = \frac{average\_squared\_distance\_within\_cluster}{cluster\_size}.$$

From the alarm correlation point of view, the squared distance up to 3 s (distance of 1.73 s) is a reasonable value for general fault management in mobile telecommunication networks.
The alarm correlation methodology proposed in this paper can be summarized by the following steps:

1. Pre-processing – decomposing alarm data sets into smaller parts, following the root object filtering criteria (generation of topology chains),
2. Applying the *k*-means iterative algorithm, along with the time correlation criteria for each of the filtered topology chains from step 1 (this part requires multiple execution, due to *k*-means algorithm's stability issue),
3. Formulating the RCA hypothesis list based on results of step 2,
4. RCA analysis performed by domain experts.

In the experiment, we used the `R` package and the *k*-means function implemented in this environment. The *k*-means function in `R` offers several algorithms like Lloyd, Forgy, and MacQueen [18]–[22]. Lloyd's, MacQueen's and Forgey's (for continues cases) algorithms follow an intuitive, definition-based approach by repeatedly computing and assigning the observations to the closest center (centroid) [23]. By default, the `R` package uses the *k*-means algorithm implementation proposed by Hartigan and Wong.

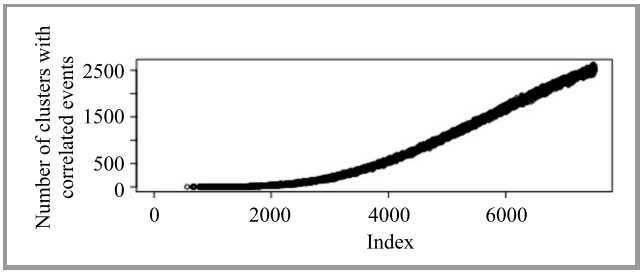

Fig. 4. Number of clusters satisfying the correlation objective, identified for the entire sample, with no topology filtering.

Experiments show that the number of clusters discovered by iterating the k-means method is growing non-linearly until we reach the K clusters split. It is illustrated by the results presented in Fig. 4. In addition, the processing of the algorithm is time consuming (computational complexity $O(n^3k)$, where n is the size of the data sample, k is the number of clusters) and results in a processing time of several hours for a data set containing several thousand alarms.
Due to above constraints and in consideration of the role of topology filtering in the RCA analysis, we have proposed an additional pre-processing step, which makes the methodology more efficient and acceptable from the point of view of the processing time.
The additional step consists in dividing the data set into subsets containing alarms belonging to one topology chain (following one topology root network element). The topology pre-filtering step introduces a very useful property of the *k*-means iterative methodology. It introduces a global maximum to the function between the number of clusters satisfying the correlation objective and the total number of clusters generated. This property, shown in Fig. 5, is observed for the first time and was not described in any paper in the past as per the author's knowledge.
This approach also addresses technical specificities of the correlation which shows that the majority of correlated events originate from the same topological chain. This type of correlation is called inner topology correlation. It is also possible to execute an outer topology correlation by com-

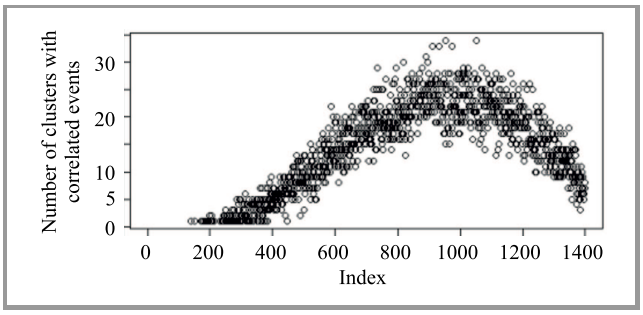

Fig. 5. Number of clusters satisfying the correlation objective (number of observations per split, where $c \leq 3$), as the function of the number of splits K for sample 1. The global maximum of 34 clusters satisfying the correlation objective was achieved for the value of 954 and 1051 total clusters generated for a sample containing 1400 alarms.

paring the centroids and alarms associated with them with inner topology correlation sets.

An example of the distribution of topology chains in a mobile telecommunication network is presented in Fig. 3. This methodology of correlation enables us to identify the method of partitioning the data set which maximizes the number of clusters for given correlation criteria, within a reasonable time. It is an optimal clustering solution which we seek for each data set, as we expand the maximum number of clusters for a given data set. The method can be used selectively for a chosen topology chain, or as the concatenation of several or all topology chains.

4. Results and Examples

The presented examples of partitioning performed on real alarm data samples are from a live mobile operator network. The data set which has been selected for simulations had 1440813 alarms divided into several sample sets. The data set which is analyzed in the example contains 7517 alarms. For the inner topology chain analysis we selected one BSC (BSC-1) which filters 1600 alarms belonging to that topology chain. The data originates from a heterogeneous, live mobile network containing 2G, 3G and 4G network elements, and was collected between July 2014 and May 2015. The data sample selected for analysis contains 28 BSCs and 27 RNCs.

As mentioned earlier, the *k*-means algorithm results depend on the initial conditions. This means that the starting centroids selected for the analysis, as well as the convergence process of each iteration result in a different number of clusters computed by the algorithm. Situations are also experienced where, for given number of clusters, the algorithm does not converge in within a specified limit of iterations or, where solutions are trapped in the local extremum. The above factors mean that each iteration run finishes with a different amount of detected clusters, as well as with a different amount of clusters matching the events correlation criteria specified: $\frac{average\_squared\_distance\_within\_cluster}{cluster\_size} \leq 3$.

Regardless of the specificities referred to above, the approximated iterated k-means algorithm proposed herein selects major clusters from the data set and the results are satisfactory. It can be seen that the main clusters, especially those with several events, are discovered by each iteration of the algorithm.

Figure 4 presents the algorithm's output for the entire data set containing 7517 alarms, without topological prefiltering. The test took 4 hours to perform in this case. We can see that iteration of the *k*-means algorithm for non-filtered data generates a number of clusters growing in a non-linear trend. Figure 5 presents output of the *k*-means iterative algorithm which was run on a pre-filtered data subset representing alarms belonging to the BSC-1 network element topology chain.

From the RCA perspective, each cluster which satisfies the correlation objective ($c \leq 3$) represents a cause of the first alarm or of several alarms from the cluster identified. The algorithm generates only a filtered correlation hypothesis, which has to be verified by an expert before assuming repairs of the network [5]. The experiments confirmed effectiveness of the methodology in question. In all clusters which satisfy the correlation criteria ($c \leq 3$), the troubleshooting hypothesis has been verified very quickly. It is worth mentioning that during the root cause hypothesis verification stage, topology is the factor that should always be taken into account. The proposed methodology takes into consideration topological aspects of the troubleshooting process by analyzing topology chain correlations. Thus, by default, we take into consideration the topological relation between the network elements generating alarms.

4.1. Alarm Inner Topology Correlation Example Discussion

For the example and discussion we selected one of 34 clusters identified by the *k*-means algorithm iterations with the centroid value of 82482.667 for inner topology correlations related to the BSC-1 topology chain. It represents one of

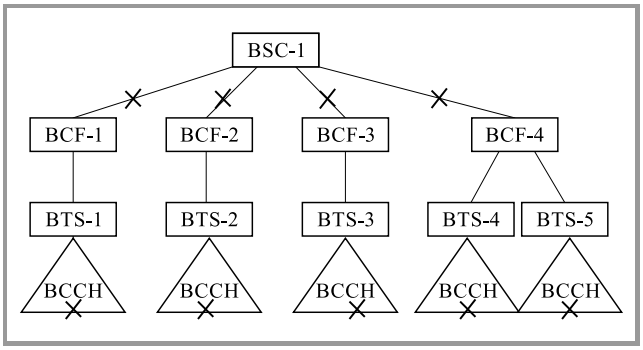

Fig. 6. Alarm correlation example.

Table 1
2G alarm correlation example

Time of event occurrence	Alarm type	Network element
82481	BTS O&M link failure	BSC-1/BCF-2
82481	BTS O&M link failure	BSC-1/BCF-4
82481	BTS O&M link failure	BSC-1/BCF-3
82481	BTS O&M link failure	BSC-1/BCF-1
82484	BCCH missing	BSC-1/BCF-4/BTS-4
82484	BCCH missing	BSC-1/BCF-4/BTS-5
82484	BCCH missing	BSC-1/BCF-1/BTS-1
82484	BCCH missing	BSC-1/BCF-3/BTS-3
82484	BCCH missing	BSC-1/BCF-2/BTS-2

the clusters for a global maximum split of 954 clusters for the data set sample. The fault illustration is presented in Fig. 6. Table 1 describes alarms used in the example. The transmission type of alarm in the network element (BCF) is causing problems with the radio broadcasting channel in another network element (BTS). The rectangle below the BTS shown in the picture symbolizes the radio network sector which is broadcast by the BTS. Inside the sector there is one radio channel component (BCCH), which plays a signaling role for the sector. The broadcast common channel (BCCH) is handling signaling communication in the sector and allows UE to log in to the network. Due to BCCH missing, there is no traffic in this sector. The problem is affecting all sectors.

5. Summary and Conclusions

In the experiment conducted, we have been analyzing several dozens of data samples with alarms from a real life mobile telecommunication network. The *k*-means iterative clustering methodology for data pre-filtered topology-wise is a very effective approach enabling to discover alarm correlation clusters (potential root cause analysis hypothesis). We have proposed an approximated alarm correlation algorithm which employs the *k*-means method for the topology chain data set by iterating the number of clusters from 0 up to K $(K < N)$. In the first stage, we propose to execute so-called inner topology chain correlation, which may be followed by an outer topology chain correlations analysis. The inner topology chain correlation iterations are characterized by reaching global maximums for the function of cluster numbers satisfying the correlation criteria:$(\frac{average\_squared\_distance\_within\_cluster}{cluster\_size} \leq 3)$ to the total number of clusters generated. This feature implies the possibility to limit the number of *k*-means function iterations to the value linked with the described maximum, which will additionally reduce the execution time. It has been observed that a vast majority of the correlated alarms originate from the inner topology chain correlation analysis, and that they play a fundamental role in selecting the event correlation hypothesis. The tests confirmed that the computation time of inner topology correlations is very reasonable in terms of practical alarm correlation. Partitioning operations for samples containing between 1200 and 2000 alarms took 10–15 s maximum.

In addition, from the overall RCA process perspective, the centroids identified indicate the moments in time which the troubleshooting engineer should pay special attention to. It has been also proven that data clustering significantly reduces the size and the quantity of the data analyzed, which makes the analysis process (network problem troubleshooting) much faster and more efficient. As far as final conclusions concerning the root cause of the faults are concerned, we need to consider other alarm attributes as well. These include: severity, number, description, type, network element type and name. There is one more practical conclusion related to the experiment. The correlation method can be used to create the so-called suppression alarm rules in the NMS. The suppression rules can be discovered after offline analysis of correlated alarms from the network and they reduce number of alarms being analyzed. For example, all alarms labeled as "BTS O&M link failure" and "BCCH missing" from the case presented in Fig. 6, identified within the same network element, can be suppressed by 1 alarm labeled "Traffic outage". This approach is similar to the pattern recognition concept, where patterns in data set analyzed are recognized and where predefined data subsets are used for further analysis and classification of data [24], [25].

References

[1] M. Lopa and J. Vora, "Evolution of mobile generation technology: 1G to 5G and review of upcoming wireless technology 5G", *Int. J. of Modern Trends in Engineer. and Research*, vol. 2, no. 10, pp. 281–290, 2015.

[2] K. Singh, S. Thakur, and S. Singh, "Comparison of 3G and LTE with other generation", *Int. J. of Comput. Applic.*, vol. 121. no. 6, pp. 42-47, 2015.

[3] A. Kumar, J. Sengupta, and Y. Liu, "3GPP LTE: the future of mobile broadband", *Wirel. Person. Commun.*, vol. 62, pp. 671–686, 2012 (doi: 10.1007/s11277-010-0088-3).

[4] M. Steinder and A. S. Sethi, "A survey of fault localization techniques in computer networks", *Science of Comput. Program.* vol. 53, pp. 165–194, 2004 (doi: 10.1016/j.scico.2014.01.010).

[5] S. K. Bhaumik, "Root cause analysis in engineering failures", *Transact. of the Ind. Instit. of Metals*, vol. 63, no. 2–3, pp. 297–299, 2010 (doi: 10.1007/s12666-010-0040-y).

[6] A. Bouillard, A. Junier, and B. Ronot "Alarms correlation in telecommunication networks", INRIA, pp. 17, 2013, [Online]. Available: https://hal.inria.fr/hal-00838969 (accessed on May 7, 2018).

[7] A. Samba, "A Network management framework for emerging telecommunications networks", in *Modeling and Simulation Tools for Emerging Telecommunication Networks. Needs, Trends, Challenges and Solutions*, A. N. Ince and E. Topuz, Eds. New York: Springer, 2006.

[8] P. Hong and P. Sen, "Incorporating non-deterministic reasoning in managing heterogeneous network faults", in *Proc. 2nd IFIP/IEEE Int. Symp. on Integrated Network Manag.*, Washington, DC, USA, 1991.

[9] M. T. Sutter and P. E. Zeldin, "Designing expert systems for real time diagnosis of self-correcting networks", *IEEE Network*, vol. 2, no. 5, pp. 43–51, 1998 (doi: 10.1109/65.17979) .

[10] A. Jain, "Data clustering: 50 years beyond k-means", *Pattern Recog. Let.*, vol. 31, no. 8, pp. 651–666, 2010 (doi: 10.1016/j.patrec.2009.09.011).

[11] H. Steinhaus, "Sur la division des corp materiels en parties", *Bullet. of the Polish Acad. of Sciences*, vol. 4, no. 12, pp. 801–804, 1956 [in French].

[12] X. Wu, *et. al.* "Top 10 algorithms in data mining", *Knowl. and Infor. Sys.*, vol. 14, no. 1, pp. 1–37, 2007 (doi: 10.1007/s10115-007-0114-2).

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* New York: Springer, Series in Statistics, 2001.

[14] S. T. Wierzchoń and M. A. Kłopotek, *Algorithms of Cluster Analysis*, Warsaw: Wyd. IPI PAN, 2015.

[15] A.-L. Jousselme and P. Maupin, "Distances in evidence theory: Comprehensive survey and generalizations", *Int. J. of Approx. Reason.*, vol. 53, no. 2 pp. 118–145, 2012 (doi: 10.1016/j.ijar.2011.07.00C).

[16] J. Koronacki and J. Ćwik, *Statystyczne systemy uczące się*. Warszawa: Exit, 2008 [in Polish].

[17] L. Morissette and S. Chariter, "The k-means clustering technique: General considerations and implementation in Mathematica", *Tutor. in Quantit. Methods for Psychol.*, vol. 9, no. 1, pp. 15–24, 2013 (doi: 10.20982/tqmp.09.1.p015).

[18] J. MacQueen, "Some methods for classifications and analysis of multivariate observations", in *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, L. M. Le Cam and J. Neyman, Eds. Berkeley, CA, USA: University of California Press, pp. 281–297, 1967.

[19] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm", *J. of the Royal Statis. Soc. Series C (Appl. Statistics)*, vol. 28, pp. 100–108, 1979, (doi: 10.2307/2346830).

[20] E. W. Forgy, "Cluster analysis of multivariate data: efficiency vs. interpretability of classifications", *Biometrics*, vol. 21, pp. 768–769, 1965.

[21] S. P. Lloyd, "Least squares quantization in PCM", *Transact. on Infor. Theory*, vol. 28, no. 2, pp. 128–137, 1982 (doi: 10.1109/TIT.1982.1056489).

[22] R documentation help of kmeans function [Online]. Available: https://www.rdocumentation.org/packages/stats/versions/3.5.0/topics/kmeans (accessed on May 7, 2018).

[23] M. Telgarsky and A. Vattani, "Hartigan's Method: k-means clustering without Voronoi", in *Proc. 13th Int. Conf. on Artif. Intel. and Statistics AISTATS*, Chia, Sardinia, Italy, 2010.

[24] G. H. Ball and D. J. Hall, *ISODATA: a novel method of data analysis and pattern classification*. Menlo Park, CA: Stanford Research Institute, 1965.

[25] *Pattern recognition*, A. Pinz, T. Pock, H. Bischof, and F. Leberl, Eds. New York: Springer, 2012 (doi: 10.1007/978-3-642-32717-9).

Artur Maździarz received his M.Sc. in Telecommunications Engineering from the Warsaw University of Technology, Faculty of Electronics and Information Technology, in 1999. Since 1999 he has been linked with the Nokia corporation, holding different engineering and management positions – all of them concerning Network Management Systems. Currently, he is a Ph.D. candidate at the Systems Research Institute of the Polish Academy of Science. His scientific research focuses on fault propagation models, event correlation and root cause analysis in mobile telecommunication networks.

E-mail: artur.mazdziarz@nokia.com
Systems Research Institute
Polish Academy of Science
Newelska 6
01-447 Warsaw, Poland