Capturing Data Topology Using Graph-based Association Mining

Khalid Kahloot, Peter Ekler
Department of Automation and Applied Informatics (AUT)
Budapest University of Technology and Economics (BME), Hungary
{khalid.kahloot, peter.ekler}@aut.bme.hu

Abstract— **A dataset can underline a statistical plausibility and implausible characteristics. A graph can model the inter-relationship between the set variables in a dataset. On the other hand, the association mining produces causal structures for a transactional dataset in various kinds. Therefore, a better data representation can be attained by merging both of the two powerful tools together. Knowledge within a dataset is captured as a topology by combining an algorithm of association rule mining with a complex graph theory. In this paper, we present a modified graph-based version of Apriori algorithm for association mining, in which the probabilities of frequencies are represented using a graph data structure. A computational approach is reflected in the graph and all rules are composed of nodes, which are interconnected by in-degree and off-degree edges. The algorithm is using Apriori statistical rule mining to compose weighted nodes and weighted directed edges graph. The computational approach is necessary to be able to unravel complex relationships between co-occurred values due to multi-hop graph connectivity and navigability. The modified algorithm is tested based on heterogeneously composed traffic datasets.**

Keywords-Graph-based data representation; topology capturing; Apriori rule mining; Association Analysis.

I. INTRODUCTION

The graph is a multiple purpose data structure that can be navigated, clustered, shortened, and visualized. In addition, the graph can be easily transformed into other data structures. A graph can model data for any phenomenon, which enclose actors and interactions in-between. For instance, social activities can be viewed as a graph, where nodes are the people and weighted edges are for actions from an originator to a recipient. Various examples can be given for the graph modeling such as a disease infection as a study of social epidemiology, the virus spread through a LAN in network security studies, and a geographical sensor deployment for studying IoT Ad-hoc collaboration.

Data association, on the other hand, aims to discover the probability of the co-occurrence of features in a dataset. The relationships between co-occurring features are expressed as association rules. A dataset can be analyzed into numerically weight relations between variables. To break it down, a relation can be formed in two stages. First, find a statistical pattern upon a dataset and for all variables. This step will specify which features are associated with others. Secondly, calculate the numerical weights of these associations based on frequency or another statistical method. This step will build a matrix of weights cross features.

In order to build data association, a large dataset is required. The association would not be confidential and recommendable for composing rules in a certain context in a domain with a large number of features. The main objective for composing such rules is minimizing the support thresholds in a similar way as the unsupervised learning. Therefore, in order to find associations involving rare patterns, the algorithm must run with very low minimum support thresholds. However, doing so could potentially increase the number of enumerated variant datasets, especially in cases with a large number of features. This could increase the execution time significantly.

We model an aggregated real dataset as weighted multi-dimensional directed graphs to allow the discovery of correlations between heterogeneous data types. We can retain important spatial structures by using the Apriori association mining with a graph, which extracts each node degree and then using it with support and confidence as parameters.

In order to capture topology, filtering algorithms can collaborate in the process of discovering a neighborhood of variables. Several recommendation algorithms can be used as model-based techniques as long as they can learn in unsupervised way. For example, feature reduction (PCA), Self-organizing Map (SOM), and Apriori association rules mining are commonly used for feature extraction and selection. SOM and PCA are often used to reduce dimensionality, but are not necessarily the best methods as they are linear and parametric methods. The set of output variables cannot be explained or labeled. Moreover, these two algorithms are sensitive to missing values. It is recommended to feed them data after being cleaned and standardized. On the other hand, the Apriori association mining can handle text and nominal data in addition to numerical data because it is a counting method, unlike the SOM and PCA which are arithmetic computational.

In this paper, we develop a graph structure and introduce new procedures to reduce or avoid the significant costs as mentioned above in the SOM and PCA. We name the algorithm Graph-based Association Mining (GAM). In other words, we have modified the Apriori algorithm for rule mining to work with a topological weighted multi-dimensional directed graph. Apriori algorithm generates a support graph based on a support threshold. Thereafter, the algorithm uses this graph by utilizing the Kachurovskii's theorem, which states that a monotonic confidence graph can be used to dimensionalize a graph. This procedure position nodes into dimensions and magnifies their weights correspondingly.

The layout of this paper starts with related work in Section II. The formal definition of a graph, properties, procedures, is illustrated in Section III. A performance comparison between variations of Apriori algorithm is described and a practical example of application over a dataset with a discussion is presented in Section IV. Finally, a highlight of purpose, applications, and future work are stated in the conclusion in Section VI.

II. Related Work

In the medical field, many types of research have considered the social effect in causality of spreading diseases or phenomena. The dataset of features is considered as a network to represent the environmental and medical confounders causes of a certain disease, which in some cases need to be adequately controlled. Researchers draw a cautious observation in health studies that conclude an attributive correlation between friends and disease spread as social network effect.

Many studies addressed obesity phenomena and the effects of social networks on its spread in countries such as England and USA or among a certain age such as elderly and children. More details can be found in [9]-[13] respectively. For instance, El-Sayed et al. [8] presented an application of simulation models for causal inference in epidemiology. They assessed whether interventions targeting highly networked individuals could help to reduce population obesity. By using network-based interventions, they recommended a useful anti-obesity strategy.

Cohen et al. in [14] used an empirical estimation to examine the network effect using common methods. They test the hypothesis against unlikely social transmission of acne and headaches. The health of the group is described in one equation with estimating social network effects within reference groups. First, that friendship selection is non-random, which leads to a correlation between the error term and friend's health. Secondly, the confounding factors affect all members of the reference group.

Other studies focus on building relationships among dataset. The behavioral data sheds a considerable light on the amount of unknown and hidden relationships. Such data is not prone to saliency cognitive filters. Studies like [5]-[7] agree that it is not enough to consider the self-reported edges and behavioral dataset especially with such as a self-reported ones with those inferred by a factor analysis of behavioral data. Moreover, these studies urge caution in combining different kinds of data. A network with multiple types of edges can obscure important nuances that should be leveraged through parallel analyses rather than flattened into a single monolithic network.

Eagle et al. [5] provide an objective way to identify, to wrestle with, and to mitigate the cognitive biases of human's expression, which often muddy the waters for scientific understanding of human phenomena. They constructed networks representing reported friends, who are communicating on phone on Saturday night, and are traveling. Nodes reflect the two groups of colleagues at the first-year of business school and the Media Laboratory students working together in the same building on campus.

Interesting research topics are concerned of graph-based visualization for the association rules, which is more suitable for visual analysis and comparison in aggregated perspective on the most important rules. Graph-based techniques can be found in [1]-[3]. Hahsler et al. in [1] introduced a new interactive Graph-based visualization method with itemsets as vertices, which allows to intuitively explore and interpret highly complex scenarios. Hahsler et al. in [1] utilized the framework for visualizing provided by Ertek et al. in [2]. As for the latter, they approached through a Market Basket Analysis (MBA) case study where the data mining results were visually explored for a supermarket dataset. Likewise, Rainsford et al. in [3] define a temporal interval data for a temporal interval algorithm of association mining. To visualize temporal relationships, a circular graph has been adapted as a set of associations that allows underlying patterns in the associations to be identified.

III. Modified Apriori Algorithm

This section defines a graph as a data structure and sets its properties, in addition, it explains the procedures operated by this graph. Moreover, a modified version of Apriori is illustrated to represent data as a support graph, thereafter, to reform the support graph into dimensional confidence graph.

A. A Definition for the weighted Multi-Dimensional Directed Graph

The weighted multi-dimensional directed graph is a directed graph with self-loops and parallel weighted edges, but edges are positioned in dimensions. In other words, multi-edges are multiple edges between two nodes and each edge hold a weight as data attributes to represent the capacity of that edge. Nodes are also holding a weight as data attributes for representing the magnitude of that node. Let the definition of the graph be G; a weighted multi-dimensional directed graph with size of k is defined as follows;

$$G = \left(N(G), E(G), \psi_G\right), \tag{1}$$

$$N(G) = \{N_1, N_2, \ldots., N_k\}, \tag{2}$$

$$E(G) = \left\{e_{ij}, e_{ik}, e_{jl}, \ldots., e_{lk}\right\}, \tag{3}$$

$$\psi_G(e_{ij}) = \{(d_1, \psi_{ij(1)}), (d_2, \psi_{ij(2)}), \ldots., (d_\lambda, \psi_{ij(\lambda)}), \ldots., (d_n, \psi_{ij(n)})\} \quad (4)$$

$$d_\lambda(E) = \{(e_{ij}, \psi_{ij(\lambda)}) \;\forall e_{ij} \subseteq E(G) \;\Big|\; d(e_{ij})\, d_\lambda\}, \quad (5)$$

$$d(G) = \{d_1(E), d_2(E), \ldots., d_\lambda(E), \ldots., d_n(E)\}, \quad (6)$$

In (2), N_i is a node in the G and ω_i denotes the magnitude of the node N_i. In (4), e_{ij} is a directed edge between $N_i \; to \; N_j$. In (4), $\psi_{ij(\lambda)}$ is the weight of the edge for representing the capacity and $\psi_G(e_{ij})$ is the set of all directed edges between $N_i \; to \; N_j$. While N_i & e_{ij} are denoted as identifiers but ω_i & $\psi_{ij(\lambda)}$ are numerical values, edges are directed as, $e_{ij} \neq e_{ji} \equiv (d_\lambda, \psi_{ij(\lambda)}) \neq (d_\lambda, \psi_{ji(\lambda)})$.

In (5), d_λ is the λ_{th} dimension for the edge, which is a ranking factor, and $d_\lambda(E)$ is the set of all edges, which positioned in the $d(\lambda)$ dimension for any node in the graph G. In (6), $d(G)$ is the dimensional representation of the graph G, in which $d_\lambda(E)$ is the set of the defined above.

The construction of the graph G starts by supplying N_i, ω_i the procedure $add\_node()$ to represent a unique identifier and a magnitude respectively. In turn, the procedure $add\_node()$ guarantees no duplicate identifiers for the nodes, however, in case of inserting the same identifier twice, the procedure $add\_node()$ will sum up the magnitudes. By comparison, the procedure $add\_edge()$ connects between two identifiers of two nodes N_i & N_j to create a directed edge e_{ij} with a weight $\psi_{ij(\lambda)}$ positioned in the dimension $d(\lambda)$. Nonetheless, another procedure is needed to construct paths between a set of nodes, which is procedure $add\_path()$. By the supply a set of node $\{N_1, N_2, \ldots., N_k\}$ and weight $\psi_{1k(\lambda)}$ for this directed path, in turn, the procedure $add\_path()$ will create directed edges $\{e_{i1}, e_{i2}, \ldots., e_{k\,j}\}$ but in this case the weight will be divided equally for the edges, i.e. $\frac{\psi_{ik(\lambda)}}{k-1} \;\forall e_{ij}$.

Suppose that an algorithm can generate relationships between values into graph nodes and edges as described above. A procedure $combine()$ is introduced to construct one graph G by combining the two graphs. As shown in Fig. 1, graph G_α and graph G_β have node N_i occur in both, also graph G_α and graph G_γ have nodes N_i, N_j and N_k occur in both. The procedure $combine()$ combines the graphs together without losing the edges and sums up the magnitude of co-occurred nodes as is shown in Fig. 2.

B. Modified Apriori Algorithm

Apriori algorithm considers the data as items in a collection of baskets and statistically generates rules for consequent frequencies of items. The modified version of Apriori considers the dataset as a set of features with a variant set of values. Then, it calculates weights that express probabilistic relationships between all-to-all cross features in the dataset. For example, for a feature f_1, an association a_1 is derived from a dataset containing f_1, f_2, $\mathrm{f}_3, \ldots., f_\mathrm{d}$.

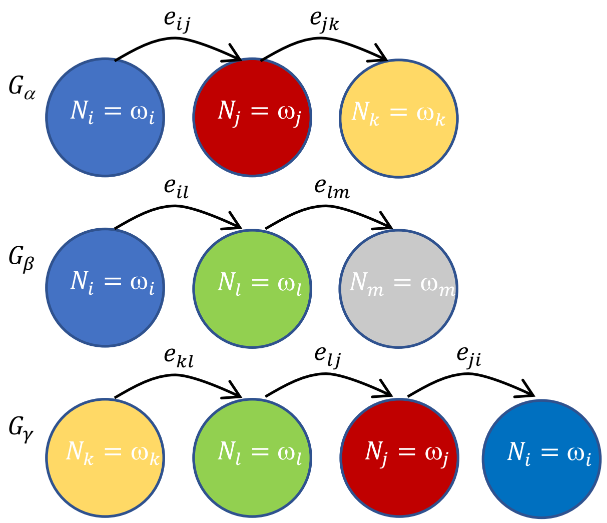

Figure 1. G_α, G_β and G_γ are Graphs with nodes coocured in all

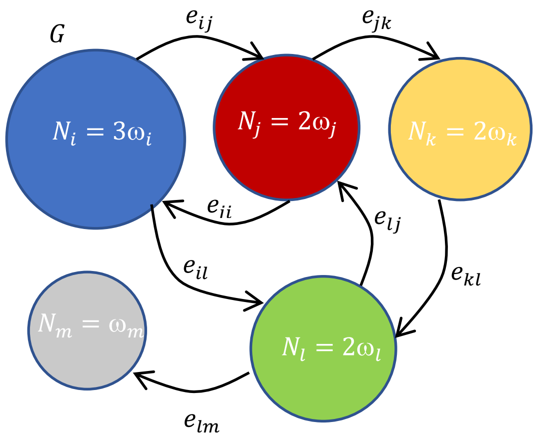

Figure 2. Graph G combined out of G_α, G_β and G_γ

This association states how often the feature f_1 changes co-concurrency to the other features of the dataset. Like decision tree rules, the algorithm derives the association from a target feature by maximizing the split and minimizing the error. The set of associations $\{a_1, a_2, \ldots., a_d\}$ is mapped to a set of graphs $\{G_1, G_2, \ldots., G_d\}$. As a matter of fact, the modified version of Apriori utilizes the graph data structures immediately by the first step. As presented below, two $algorithms$ should be executed conclusively. First, the algorithm $generate\_support\_graph($ uses parameter $minSupport$, which is a threshold of metric $Support$. Second, the algorithm $dimensionalize()$uses parameter $minConfidence$, which is a threshold of metric $Confidence$. The formal definitions of these metrics are:

$$Support: s(N_i \rightarrow N_j) = \frac{\sigma(N_i \cup N_j)}{n} \quad (6)$$

$$Confidence: c(N_i \rightarrow N_j) = \frac{\sigma(N_i \cup N_j)}{\sigma(N_i)} \quad (7)$$

B.1. Graph Generation using Apriori Association

Let $F = \{f_1, f_2, f_3, \ldots, f_d\}$ be the set of all variables and $V = \{v_1, v_2, v_3, \ldots, v_n\}$ be the set of all values in the in a dataset D . The objective of the algorithm $generate\_support\_graph()$ is to build a topological structure out of this dataset using association analysis. The topological structure is a weighted multi-dimensional directed graph, which contains features as nodes and directed edges out of each. An important property of an edge is its weight, which refers to the a statistically likelihood of occurrence through the dataset in a certain order. The algorithm is using the support count as a weighting scale. Mathematically, the support count, σ(N), for an ordered subset of features F can be defined as follows:

$$\sigma(N) = \sum_{i=1}^{n} |\{v_i | f \subseteq v_i, v_i \in V\}| \quad (8)$$

An edge is an implication expression of a directed navigation from a node to another. For instance, expressing an edge between two nodes would be; $v_{l(f_i)} \rightarrow v_{k(f_j)}$ as an association between certain the value v_l for a feature f_i co-occurred by a certain value v_k for feature f_j. Simply, it can be named as nodes $N_i \rightarrow N_j$. However, the disjoint nodes are expressed as $N_i \cap N_j = \emptyset$. The weight of an edge can be measured in term of support that determines how often a rule is applicable to a given dataset, while confidence determines how frequently a certain value $v_{l(f_i)}$ co-occurred by a certain value $v_{k(f_j)}$.

Algorithm 1. Graph Generation by support

```
$G_t = \{G_1, G_2, \ldots, G_d\}$
$N_k = \{i \mid i \epsilon I \wedge \sigma(\{i\}) \geq N \times minSup\}$
$E_k = Aproiri(N_k, N_k)$ //All Edges
while $N_k \neq \Phi$ do
    $N_{k-1} = subset(N_k, minSup)$
    $E_{k-1} = Aproiri(N_k, N_{k-1})$
    $G_i = add\_path(E_{k-1})$
    $i = i + 1$
end while
output $G_s = combine(G_t)$
```

As shown in *Algorithm 1*, the objective of the algorithm $generate\_support\_graph()$ is to eliminate the weakest edges by using procedure Apriori support association to weight the edges. Moreover, the algorithm aims to reduce the number of comparisons by getting advantage of graph data structure and graph algorithms. Let N_k denote the set of nodes and E_k denotes the set of edges. Initially, the procedure Apriori support generates temporary set of graphs G_t to represent the set of features then determines the support of each nodes. Iteratively, the algorithm generates new edges and updates the weight of the already existing edges and uses procedure $combine()$ to generate the support graph G_s.

B.2. Graph Dimensionalize using Confidence Graph

The algorithm dimensionalize() is a procedure to reform the support graph G_s by directed edges generated by Apriori confidence graph G_c, which is partitioned to satisfies the confidence threshold. Unlike the support measure, confidence does not have any monotone property and generates only one edge e_{ij} for each subset. In other words, the edges generated by this procedure can only be true entirely ordered that is the reason to divided the weight $\psi_{ik(\lambda)}$ equally over between edges $e_{i1}, e_{i2}, \ldots, e_{kj}$. According to Kachurovskii's theorem, the monotonic confidence procedure can generate a topological vector space X; that is in a graph G of $X \rightarrow X^*$ is composed of monotonic analytical function such as procedure Apriori confidence.

Algorithm 2. Graph Dimensionalize

```
$k = |G_s|$ // size of graph Gs
$E_{ij} = \{(S_i, S_j, e_{ij}) \mid e_{ij} \subset E(G_s) \wedge e_{ij} \epsilon \lambda\}$
for each $e_{ij}$ in $E_{ij}$ do
    $\psi_{ij(\lambda)} = conf(e_{ij})$
    if $\psi_{ij(\lambda)} > minConf$ then
        // graph of subset(i)
        $G_i = (N(S_i), E(S_i), 1)$
        // graph of subset(j)
        $G_j = (N(S_j), E(S_j), 1)$
        $G_{ij} = combine(G_i, G_j, \psi_{ij(\lambda)})$
$G_c = combine(G_{ij} \forall \lambda)$
output $G_c$
```

As shown in Algorithm 2, the procedure Apriori confidence is used by the algorithm $dimensionalize()$ to position the edges in a dimension λ based on the confidence of each edge which extends and dimensionalizes into a confidence graph G_c. Given a supporting graph G_s, the procedure finds all the edges having $Confidence \geq minConf$, where $minConf$ is the corresponding confidence threshold. Likewise, the support graph generation, a brute-force approach is applied for mining confidence association rules. However, now it works over the given support graph G_s. This approach is much optimized and less expensive because the algorithm has to compare the edges of the graph G_s in a logarithmic time instead of exponentially comparing like in the traditional Apriori algorithm. More specifically, for a confidence graph G_c , which was extracted from a support graph G_s that contains d nodes, the possible edges are:

$$\mathrm{E(G_s)} = 3^{\log(d)} - 2^{\log(d)} + 1 \qquad (9)$$

Such approach can be less expensive because it requires $O(N\ log\ (d)\ \lambda)$ comparisons, where N is the number of nodes i.e., $N(G_s) = \{N_1, N_2, \ldots.., N_k\}$ number of nodes' support graph $\mathrm{G_s}$ to represent the set of features in $F = \{f_1,\ f_2,\ f_3, \ldots..,f_d\}$, where $k = 2^d - 1$ is the number of edges in the graph $\mathrm{G_s}$, and λ is the maximum number of co-occurrences, which represents the dimension.

IV. Results and Discussion

Apriori algorithm has enormous number variations which modified the data structure to outperform the original algorithm. The data structure of GAM is quite the difference of other variations of Apriori algorithm. Although GAM does not address the performance issues, it is importance to compare the performance of GAM with famous well-known variant implementations. A survey and comparison are presented in [15]. All tests were carried out on ten public "benchmark" databases, which can be downloaded from [16]. First, we compared GAM used for storing filtered transactions against a sorted list, a red-black tree (B-tree) and a trie.

TABLE I. MEMORY NEEDED AS SORTING FREQUENCIES FOR THE T40I10D100K DATASET

min_freq	GAM	Sorted list	B-tree	trie
0.05	60.1	9.3	10.8	55.4
0.02	79.7	12.7	14.1	70.3
0.0073	96.3	19.5	20.3	80.3
0.006	100.8	21.3	21.5	88.4

As shown in Table I, when it comes to memory, complex data structures are in distress. The close competitor of GAM is the trie implementation and still overcomes the GAM especially with high frequencies. However, the added structures of nodes and edges are very important and we did not mean to optimize memory although the difference is quite acceptable.

The Dataset contains accidents over the years 2012 to 2014 of traffic flow in the city of London, UK. It has been compiled from the UK government sources and it is available online for analysis. Accident events are aggregated to a square grid and stacked vertically. The number of casualties colors each event. This map was developed by a professional pythonist called Dave Fisher-Hickey and it was published on Kaggle website [17]. The available data describe the Average Annual Daily Flow, which tracks how much traffic there was on all major roads in addition to accident data from police reports.

The dataset contains 26 features. Primarily, to put down a summary, the most important features in the dataset are coordinates, number of vehicles, number of causalities, light, weather conditions and more. The values presented are 31153 records. As for Apriori parameters, the support threshold should be small for representing as many features as possible. On the contrary, the confidence threshold should be large because it is used to group sets of values into dimensions, i.e., the graph should be extended an additional dimension only for high confidence frequencies on values. We chose $minSupport$ as 17% and $minConfidence$ as 68%.

TABLE II. GRAPH NODES AND WEIGHTED

Feature	***Value***	***Weight***
Light Conditions	Daylight	3,20225
Human Control	None within 50 meters	2,98344
Casualties	None	2,94931
Accident Severity	3	2,53462
Physical Facilities	No physical crossing within 50 meters	2,43446
Weather Conditions	Fine without high winds	2,408
Number of Casualties	1	2,3346
2nd Road Number	0	2,32879
Road Type	Single carriageway	2,26773
Road Surface Conditions	Dry	2,06181
Urban/Rural Area	1	1,98775
Number of Vehicles	2	1,799987
Junction Control	Give way	1,50366
Speed limit	30	1,39708
1st Road Class	3	1,39596
2nd Road Class	6	1,22609
2nd Road Class	-1	1,17361
Urban/Rural Area	2	1,01224
Number of Vehicles	1	0,90406
Road Surface Conditions	Wet/Damp	0,85696
1st_Road_Number	0	0,79245
Light Conditions	Darkness	0,5885

The modified algorithm GAM scanned through features and values to produce 22 (feature, value) pairs with weights as shown in Table II. For instance, feature "light conditions" was chosen twice with "daylight" and "darkness" with 3,20225 and 0,5885 respectively. The interpretation can be as accidents are more likely to happen in daylight 3 times more than in darkness but still, the light condition is a most significant feature. Another example is "speed limit" that was chosen to be "30" and that implies that 52% of the accidents happened on roads with speed limit of 30. Likewise, we can state that accidents occur in "Fine without high winds" for "Weather Conditions" with a probability of 61.8%.

To study the "Number of Casualties", the highest frequent value is 1. The node of "Number of Casualties=1" has 5937 out edges for neighbors:

- 2nd Road Number = 0,
- Urban/Rural Area = 1,
- 1st Road Class = 6,
- 1st Road Number = 0,
- Human Control = None within 50 meters,
- Number of Vehicles = 2,
- 2nd Road Class = 6,
- Light Conditions = Daylight: Streetlight present,
- 1st Road Class = 3,
- Road Surface Conditions = Dry,
- Accident Severity = 3,
- Junction Control = Giveaway or uncontrolled,
- Carriageway Hazards = None,
- Road Type = Single carriageway,

The values listed above are direct co-occurred values for an accident with "Number of Casualties = 1". As a summary of GAM results, the number of possible scenarios is 7836, which is also the number of in-edges into this node. The edges are causes of accidents with parameters of "Number of Casualties = 1" and 60.5% of them when "2nd Road Number = 0". The pair (Speed limit = 30, Number of Casualties = 1) has 404 edges in between. Edges were positioned in 946 dimensions for the pair (Number of Casualties = 1, Road Type = Single carriageway).

V. CONCLUSION

We have combined Apriori association mining and graph theory to provide the weighted directed graph of a topological representation of the data. We have provided a formal description for the graph and explained the procedures operated over such graphs like a combination of two graphs. We algorithmically searched through a dataset of features and values by using Apriori, but with the help of graph data structure. The construction of the graph was done under a support threshold. Then this graph was dimensionalized using confidence threshold. We have showed indirect relationships between features and values, which appeared by navigating the path between the corresponding nodes in the graph. As a future work, we are planning to analyze the graph by applying PageRank and Hits algorithms.

ACKNOWLEDGMENT

This work was supported by the National Research, Development, and Innovation Fund of Hungary in the frame of FIEK_16-1-2016-0007 (Higher Education and Industrial Cooperation Center) project and by the ÚNKP-17-4-IV New National Excellence Program of the Ministry of Human Capacities.

REFERENCES

[1] M. Hahsler, & R. Karpienko, "Visualizing association rules in hierarchical groups", Journal of Business Economics,

[2] Apr 2017, Vol. 87, n. 3, pp 317-335, isn=1861-8928.

[3] G. Ertek & A. Demiriz, "A framework for visualizing association mining results", 21th International Symposium Proceedings, Istanbul, Turkey, pp 593–602, November 2006.

[4] C. Rainsford, J. Roddick, "Visualization of temporal interval association rules", 2nd International Conference Shatin Proceedings, Hong Kong, China, pp 91-96, December 2000.

[5] Nathan Eagle, Aaron Clauset, Alex (Sandy) Pentland, and David Lazer "Reply to Adams: Multi-dimensional edge inference", PNAS vol. 107 (9) , 2010

[6] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, "Clustering With Multi-Layer Graphs: A Spectral Perspective," in IEEE Transactions on Signal Processing, vol. 60, no. 11, pp. 5820-5831, Nov. 2012.

[7] A. Y. Kibangou and C. Commault, "Observability in Connected Strongly Regular Graphs and Distance-Regular Graphs," in IEEE Transactions on Control of Network Systems, vol. 1, no. 4, pp. 360-369, Dec. 2014.

[8] A. El-Sayed, L. Seemann, P. Scarborough & S. Galea; Are Network-Based Interventions a Useful Antiobesity Strategy? An Application of Simulation Models for Causal Inference in Epidemiology. Am J Epidemiol, vol. 178 (2), pp 287-295, 2013

[9] R. Bender, K. Jöckel, C. Trautner, M. Spraul, M. Berger, "Effect of Age on Excess Mortality in Obesity". JAMA.1999;281(16):1498-1504. doi:10.1001/jama.281.16.1498

[10] Amy L. Louer, Denise N. Simon, Karen M. Switkowski, Sheryl L. Rifas-Shiman, Matthew W. Gillman, Emily Oken" Assessment of Child Anthropometry in a Large Epidemiologic Study" J Vis Exp. Vol. 120, pp 54895, 2017

[11] Y Claire Wang, Klim McPherson, Tim Marsh, Steven L Gortmaker, Martin Brown, Health and economic burden of the projected obesity trends in the USA and the UK, The Lancet, Volume 378, Issue 9793, Pages 815-825, 2011,

[12] Kvaavik E, Batty GD, Ursin G, Huxley R, Gale CR. Influence of Individual and Combined Health Behaviors on Total and Cause-Specific Mortality in Men and WomenThe United Kingdom Health and Lifestyle Survey. Arch Intern Med. pp 711-718, vol. 170(8), 2010

[13] Flegal KM, Carroll MD, Ogden CL, Curtin LR. Prevalence and Trends in Obesity Among US Adults, 1999-2008. JAMA. 2010;303(3):235-241. doi:10.1001/jama.2009.2014

[14] Cohen-Cole, Ethan and Fletcher, Jason M, "Detecting implausible social network effects in acne, height, and headaches: a longitudinal analysis", vol. 337, 2008

[15] F. Bodon, "A Survey on Frequent Itemset Mining, Technical Report", Budapest University of Technology and Economic, 2006

[16] B. Goethals, M. Zaki, "FIMI 2003: Workshop on Frequent Itemset Mining Implementations". CEUR Workshop Proceedings series, vol. 90 , 2003

[17] https://www.kaggle.com/c/classify-traffic-signs/data, accessed on Feb. 7th, 2017