Proceedings of the Federated Conference on Computer Science and Information Systems pp. 675–683
DOI: 10.15439/2019F308
ISSN 2300-5963 ACSIS, Vol. 18

An Approach to Customer Community Discovery

Jerzy Korczak
International University of Logistics and Transport
ul. Sołtysowicka 19B
51-168 Wrocław, Poland
Email: jerzy.korczak@ue.wroc.pl

Maciej Pondel
Wroclaw University of Economics,
ul. Komandorska 118-120
53-345 Wrocław, Poland
Email: maciej.pondel@ue.wroc.pl

Wiktor Sroka
Wroclaw University of Economics,
ul. Komandorska 118-120
53-345 Wrocław, Poland
Assentis Technologies AG,
Blegistrasse 1, 6343 Rotkreuz,
Switzerland
Email: wiktor.sroka@gmail.com

Abstract—**In the paper, a new multi-level hybrid method of community detection combining a density-based clustering with a label propagation method is proposed. Many algorithms have been applied to preprocess, visualize, cluster, and interpret the data describing customer behavior, among others DBSCAN, RFM, k-NN, UMAP, LPA. In the paper, two key algorithms have been detailed: DBSCAN and LPA. DBSCAN is a density-based clustering algorithm. However, managers usually find the clustering results too difficult to interpret and apply. To enhance the business value of clustering and create customer communities, the label propagation algorithm (LPA) has been proposed due to its quality and low computational complexity. The approach is validated on real life marketing database using advanced analytics platform Upsaily.**

I. Introduction

DETECTING communities is one of the usual and important problems in modern data analysis of decision support systems. Many approaches and algorithms of community discovery have been published in network literature [1]-[7]. A community can be considered as a densely connected group of nodes that is only loosely connected to the rest of the network [8]. An example of such a community in a large network is a set of customers in marketing information systems having similar profile or behavior [9].

In recent years, the efficient data mining of large volume and high dimensional data has become of utmost importance. Therefore, applying the most appropriate method of obtaining accurate and business-oriented partitions is crucial. In literature one can find many clustering algorithms, starting with classical k-means, through density-based, partitioning, self-organizing maps, graph-based, grid-based, to combinational and hybrid solutions. These algorithms are usually evaluated based on clustering measurements, showing that some clustering techniques are better for large datasets while others give good results finding clusters with arbitrary shapes. Nonetheless, there is no one algorithm that can achieve the best performance on all measurements for any given dataset [4][10][11][13] and also obtain the best results.

Therefore, in marketing analysis, discovering accurate and business focused partitions using a single algorithm in isolation becomes highly complex. There are many reasons for these difficulties: sensitivity to initial values, unknown quantity of expected clusters, non-spherical datasets, sensitivity to noise and outliers, varying densities of clusters, or difficulties of business interpretation.

To strengthen the business outcomes and reduce weaknesses of the single algorithm approaches, a new hybrid multi-level method of community discovery will be proposed. It combines density-based clustering with business-oriented label propagation method. Five basic algorithms have been integrated into this method: DBSCAN, RFM, k-NN, UMAP and LPA. The DBSCAN, which has already been used in many applications [10]-[13], is taken as the density-based algorithms. DBSCAN identifies clusters by measuring density as the number of observations in a designated area. If the density is greater than the density of observations belonging to other clusters, then the defined area is identified as a cluster. Usually, in business application, DBSCAN creates a lot of difficult to interpret clusters. To improve cluster quality and interpretation, a second algorithm is proposed that enriches the results of DBSCAN and is able to form communities. After analysis of various community detection methods, the label propagation algorithm (LPA) was selected due to its simplicity and low computational complexity. The LPA was proposed by Raghavan et al. [14] for detecting communities in large networks. The idea of label propagation is as follows: before beginning computation, some nodes of the network possess assigned labels. During process execution, the labels are propagated iteratively throughout the network according to the formula below.

$$g_j = \arg\max_g \sum_j A_{ij}\delta(g_j g) \quad (1)$$

where A_{ij} is an element of the adjacency matrix of the network, δ is equal to *1* when its arguments are the same, and *0* otherwise. There are many extensions of original label propagation algorithm [15], [8], [16]. In our approach, a weighted network is assumed, so formula (1) is rewritten as:

The project was partially financed by the Ministry of Science and Higher Education in Poland under the programme "Regional Initiative of Excellence" 2019 - 2022 project number 015/RID/2018/19.

$$g_j = \arg\max_g \sum_j W_{ij}\delta(g_j g) \quad (2)$$

where W_{ij} is the sum of weights on the edges between nodes i and j of the adjacency matrix of the network, δ is equal to *1* when its arguments are the same, and *0* otherwise.
In other words, the nodes sequentially adopt the labels shared by most of their neighbors taking into consideration the weights of the edges. The propagation ends when the labels no longer change.

It is important to note that in our case study nodes are represented by clusters of customers created by DBSCAN. Neighborhoods of clusters are defined individually by the distance between the centers of clusters. The upper limit of neighboring is usually predefined by the manager or analysts, so the number of neighboring clusters is variable.

The business goal of the study is to obtain a higher quality of definition of customer communities from the marketing viewpoint. Therefore, in the approach the Recency Frequency Monetary value method has been integrated with graph clustering to give clusters of higher quality compared to the traditional mono-algorithm clustering. Various data sources, different quality measures, and business orientation provide more up-to-date and richer information for decision makers and marketing analysts.

The paper is structured as follows: in the next section, the basic characteristics of customers profiles and behavior are provided. The information is saved in the database and available using Upsaily platform. The third section describes a method of clustering of customers of the internet store and the measures to evaluate quality of the results. The fourth section details the label propagation algorithm and a method of discovery of customer communities. The results of the case study on real life database are presented and discussed in the last section. A general conclusion summarizes the outcomes of the proposed approach.

II. Analysis of Customer Behavior and Properties

The first studies of the subject of customer behavior were conducted more than 60 years ago [17], [18]. They focused on customer identification in offline stores, analysis of customer characteristics, and studies on buying-behavior patterns. It is quite common to find customer-behavior research based on questionnaires filled by researcher and a customer who would have accepted to take part in such a study [19]-[21]. Such research is time- and resource-consuming; however, but what is more important is the fact that customers behave differently when they are aware of participation in research.

Currently, analysis of customer behavior in e-commerce is much more convenient and more options can be applied. It is possible as today's e-commerce databases collect data about every single action the customer undertakes (visit, transaction, search, and many more) [17], [19]. Such systems concentrate on a delivery of the best fitting proposal for a customer in a perspective of the selected customer segment, desired product, and conditions under which the product is offered. Those issues were examined by the authors in [12] using customer clustering based on the RFM method, considering customer recency, frequency of purchases, and monetary value of orders. RFM method has been shown to be very useful in determining the proper point in time to provide customer with an offer.

This paper concentrates on another set of characteristics describing customer behavior. The proposed segmentation was inspired by direct interviews with e-commerce managers who independently observed two principals in terms of profit generation, but also contrary segments of customers. One of the segments brings together fashion-driven customers (they focus on new and fashionable items). Second one is "bargain hunters" – discount-driven customers who are ready to purchase products present on a market for a longer period of time. This segmentation is called "fashion vs. discount".

Being aware of such an observation, the authors extracted data from transactional database of the structure presented in Fig. 1. Due to a large number of tables and attributes in the source database, only tables and fields used in the experiment are presented. The data used in the experiment come from a fashion store (clothes and related items).

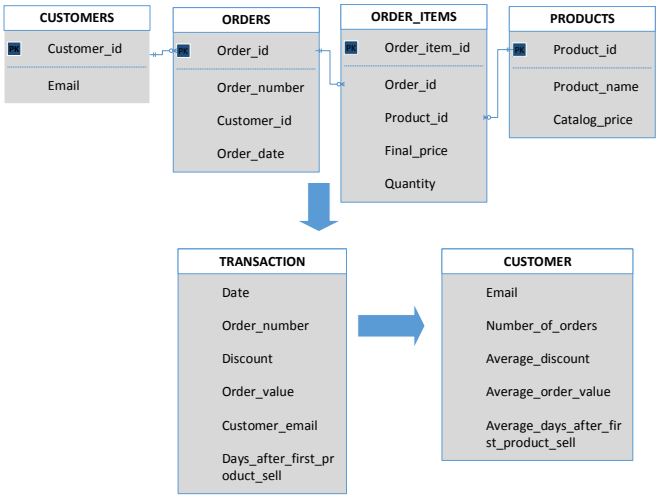

Fig. 1. Structure of source database and result tables

Having such a source database, the following measures characterizing transactions are computed (TRANSACTION table in Fig. 1):

- Value – as a quantity of items multiplied by price.
- Discount – as a percentage the difference between the highest transactional price of a specific item and its price in the current transaction.
- Days after first item sell – as the number of days from the first transaction of a given item and the current transaction.

In order to build customer profile (CUSTOMER table in Fig. 1): the above measures have been aggregated to define:

- Loyalty – expressed in the number of orders. Such an attribute differentiates the one-off buyer customer from the loyal customers.

- Average discount – high percentage discounts are typical for bargain hunters.
- Average number of days after first product sell – determines whether the customer is interested in new (fashionable) items or accepts purchasing items launched in previous seasons.
- Average order value – determines the amount of money the customer is able to spend for a single purchase.

Sample data being the input to the experiment is presented in Fig. 2. Whole data set included 264127 rows (customers).

email	average days after first product sell	average discount	average order value	number of orders
00000b3a11d76fdd26b0490c4eb0bdaf@unity.pl	204,88	44	183,00	8,00
00005bad570278ac77399df05c96ebf1@unity.pl	203,25	8	244,00	1,00
0000739e973086436944624a10acb44a@unity.pl	224,24	24	294,00	9,00
000084e2f4ccebeb412ccc3bb8270e20@unity.pl	245,63	57	158,00	2,00
0000aaab4a6e7bec76ba668d507c0b9d@unity.pl	114,00	33	103,00	1,00

Fig. 2 Sample input data

The Upsaily platform, directed to retail companies working in both B2C and B2B models, is geared towards current customers of the online internet shops. The experiment presented in this paper is based on database originating from B2C store. In the database, not only all customer transactions are stored (which is presented in Fig. 1), but also the basic data about their demographic and behavioral profile. Functionally, the solution can be classified as a Customer Intelligence system, i.e. one whose primary interest is current customers. The aim is not to help in acquiring new customers, but to increase customer satisfaction that translates into increasing turnover. It can be achieved by customers making follow-up purchases, increasing the value of individual orders (cross-selling) or more valuable products (up-selling). The Customer Intelligence approach is related to conducting analytical activities leading to creation of a clear image of the customer so that one can find the most valuable customers and send them a personalized marketing message. The system is equipped with several analyses including customer segmentation. The main screen of the platform where a manager can search for desired analysis is presented in Fig. 3.

The multi-level approach to discover customer communities will be described in the following steps:

1. Customers clustering using HDBSCAN algorithm.
2. Dimensions reduction using Uniform Manifold Approximation and Projection (UMAP) method in order to base next steps on two dimensions.
3. Centroid calculation for each cluster according to UMAP result.
4. Graph generation with k-NN algorithm.
5. Communities detection according to LPA algorithm.
6. Marketing interpretation.

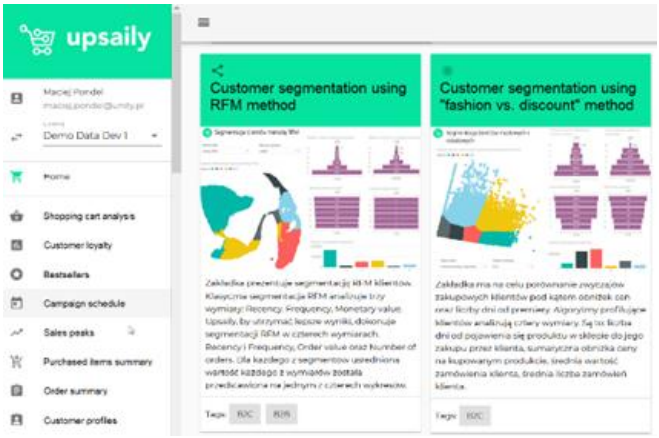

Fig. 3 Main screen of Upsaily platform

Details on the particular steps will be given in the next sections.

III. Clustering of Customers of the Internet Store

Upsaily platform is equipped with two main business customer segmentations based on RFM and the "fashion vs. discount" method explained earlier. Upsaily uses four algorithms, namely:

- k-means based on the Euclidean distance between observations.
- Bisecting k-means acting on a basis similar to k-means, however, starting with all the observations in one cluster and then dividing the cluster into two sub-clusters, using the k-means algorithm.
- Gaussian Mixture Model (GMM), which is a probabilistic model based on the assumption that a particular feature has a finite number of normal distributions.
- HDBSCAN which is an extension of DBSCAN algorithm presented in [22]. The original DBSCAN identifies clusters by measuring density as the number of observations in a designated area. If the density is greater than the density of observations belonging to other clusters, then the defined area is identified as a cluster.

Experiments with each algorithm indicating their strengths and weaknesses as well as collaborative approaches have already been presented in [12].

In the current experiment, we would like to identify small but very precise segments of the most profitable customers. A profitable customer is one whose order values are high and at the same time they don't seek discounts. Authors have done corresponding clustering using k-means clustering algorithm in order to evaluate proposed method by comparison with the typical approach. Source data included 264127 rows describing customers (presented in Fig. 2). 140 segments were generated. Fig. 4 presents visualization of 7 selected clusters of the most profitable customers. Customers assigned to clusters (indicated by color) are presented in left hand side. Distribution of order values is presented in right hand side.

Methods of clustering assessment were presented in [12], [23]. For the interesting clusters, the measure of scatter within the cluster using the Davies-Bouldin index is computed. In general, the lower the value of the measure, the more consistent a cluster is. In this experiment measure of scatter was between 38.68 (best value) and 168.01 (worst value). Average value in seven selected clusters is 80.47.

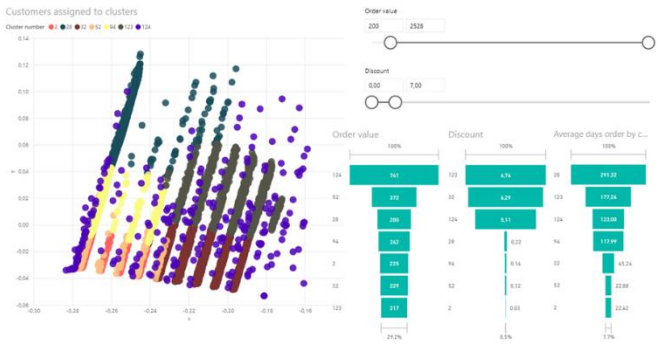

Fig. 4 Most profitable customers in k-means segmentation

In order to perform current experiment, HDBSCAN method was selected because of its marketing usage in effective discovery of clear patterns in given set of observations. It is worth mentioning that other algorithms are focused on assigning observations to a specific number of clusters defined by user upfront. HDBSCAN generates the number of clusters based on the number of patterns found in the data. It can also leave some observations without assigning them to any cluster if no pattern is found.

To understand the idea of HDBSCAN, the basic DBSCAN has to be explained.

The algorithm can be abstracted also into the following steps [24]:

1. Find the points in the ε-neighbourhood of every point and identify the core points with more than `minPts` neighbours.
2. Find the connected components (subgraph) of core points on the neighbor graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε-neighbor, otherwise assign it to noise (outliers).

The DBSCAN algorithm can be parameterized by ε (eps) defining the minimum distance between two points and `minPts` denoting the minimum number of points to form a dense region.

DBSCAN algorithm in pseudo code is given [25]

```
DBSCAN(SetOfPoints, Eps, MinPts)
  //SetOfPoints is UNCLASSIFIED
  ClusterId := nextId(NOISE);
  FOR i FROM 1 TO SetOfPoints.size DO
   Point := SetOfPoints.get(i);
   IF Point.CiId = UNCLASSIFIED THEN
    IF
ExpandCluster(SetOfPoints,Point,ClusterI
d,Eps,MinPts) THEN
     ClusterId := nextId(ClusterId);
    END IF;
   END IF;
  END FOR;
END; // DBSCAN
```

`ExpandCluster` function checks all points in neighbourhood of a given point. If number of those points is higher than `minPts` parameter, they are assigned to cluster otherwise to noise.

HDBSCAN converts DBSCAN into a hierarchical clustering algorithm, and then uses a technique to extract a flat clustering based on the stability of clusters. The following steps present the idea of HDBSCAN [22]:

1. Transform the space according to the density/ sparsity.
2. Build the minimum spanning tree of the distance weighted graph.
3. Construct a cluster hierarchy of connected components.
4. Condense the cluster hierarchy based on minimum cluster size.
5. Extract the stable clusters from the condensed tree.

The result of step 1, customers assigned to clusters using HDBSCAN algorithm, is presented in Fig. 5.

email	Cluster number
00005bad570278ac77399df05c96ebf1@unity.pl	562
0000739e973086436944624a10acb44a@unity.pl	-1
000084e2f4ccebeb412ccc3bb8270e20@unity.pl	1428
0000aaab4a6e7bec76ba668d507c0b9d@unity.pl	2001
0000c55228e230f7328cb8d30cfa6a56@unity.pl	1126
0000ec3ee95d687941ffdb40a2978e10@unity.pl	1454
000106e34137ad831fcadb9f1a136c30@unity.pl	384
00014afc8bd6d898b44829fe59bf4a3a@unity.pl	1570
000230f75c317ad9c09cf9fcaa885f9d@unity.pl	445
00027aefbaecbdf9e12a325befd2e9bf@unity.pl	1999
0002c0d61e903bbe749507093563f03b@unity.pl	1207
0003067477d1c77200568dadcf8728cc@unity.pl	1447
00039e775949311810649e8ec21359b2@unity.pl	445
0003b08584f2cb8e3bce38a7750a6590@unity.pl	-1
0003d34a40ef266469c39697157f89ba@unity.pl	707

Fig. 5 Customers assigned to HDBSCAN generated clusters

The result of step 2 and 3 which concerns a dimension reduction using UMAP method [26] and centroid calculation is presented in Fig. 6. Centroids are described by *x* and *y* coordinates.

Cluster number	x	y	Number of customers
-1	7,39	0,13	89830
0	13,06	4,62	126
1	21,54	-6,74	79
2	21,25	-6,34	41
3	15,38	-8,85	423
4	30,61	-6,73	130
5	24,94	-13,68	95
6	28,60	-11,19	99
7	28,60	-11,19	36
8	16,47	-6,44	43
9	27,12	-8,20	111
10	15,79	7,53	34
11	16,59	-11,57	86
12	25,85	4,09	124

Fig. 6 Cluster summary with cluster centroid calculated

Source data included 264127 rows describing customers, out of which 174297 were assigned to clusters by HDBSCAN algorithm. The remaining customers (89830) were assigned to cluster –1, which means that insufficient density was found in the area they were located (no pattern was detected). 2046 small but very consistent clusters were discovered. Such quantity of clusters is too high to effectively address managers' needs, hence the reason why the aim of the next step will be to aggregate small clusters into customer communities.

IV. Discovery of Customer Communities - Label Propagation

In [27], communities are defined as "*groups of vertices within which connections are dense, but between which connections are sparser*". According to [28], such communities can be considered as fairly independent spaces of a graph, sharing common properties and/or playing similar roles within it.

In our study, communities are groups of customer clusters whose elements share common properties and allow managers to apply the same measures to them or to identify strong similarities between groups in the same community.

Label Propagation Algorithm has been proposed by Raghavan et al. [14] for detecting communities in networks represented by graphs. The algorithm, due to its linear time complexity of O(m) for each iteration, simplicity, and ease of implementation, is commonly used to identify communities in large-scale real-world networks, such as social media.

An advantage of the algorithm is that it does not require prior information about number of communities or their cardinalities to run; neither does it require any parameterization. The number of iterations to convergence is barely dependent on the graph size, but it grows very slowly.

In [29] the LPA has been compared with other clustering algorithms: Louvain algorithm [30], Smart Local Moving (SLM) [31] and Infomap algorithm [32]. Results of that experiment favors LPA to be used with large scale data as it outperforms other algorithms for well-defined clusters.

These characteristics of the LPA method was the main reason for choosing it for detecting communities of customers and proposing a new method combining multiple methods: HDBSCAN creating numerous clusters, UMAP reducing dimensions, k-NN forming graph, and LPA finding communities.

The main idea behind LPA is to propagate labels throughout the graph from a node to its neighbor nodes. As a result, the groups of nodes sharing the same label and whose nodes have more neighbors than nodes in other groups make communities.

The algorithm consists of five steps [33]:

1. Initialize the labels at all nodes in the network. For a given node $x, c_x(0) = x$
2. Set t=1
3. Arrange the nodes in the network in a random order and set it to x
4. For each $x \in X$ chosen in that specific order, let
   $c_x(t) = f\left(c_{x_{i1}}(t),...,c_{x_m}(t)\right),$
   $, c_{x_{i(m+1)}}(t-1),...,c_{x_{ik}}(t-1)$
   where f returns the label occurring with the highest frequency among neighbors. Select a label at random if there are multiple highest frequency labels.
5. If every node has a label that the maximum number of their neighbors has, then stop the algorithm. Else, set $t=t+1$ and go to (3).

Label propagation works as follows: at the beginning all clusters make own communities, by assigning unique labels to every cluster, then the following steps are being executed in a loop. In every iteration all clusters are processed in a random order and the labels are updated to one that occurs with the highest frequency among the direct neighbours. If the label cannot be chosen as there are multiple labels occurring with the same frequency, then one of them should be chosen randomly. If all clusters are processed in this iteration, stop condition is checked: all clusters should be labelled with the one that majority of adjacent clusters have and if the condition is met, the algorithm ends. Otherwise, the iteration is repeated until convergence defined as the stop condition is reached. In this way, labels will propagate across the graph, replacing other labels and eventually some labels will disappear, and others will dominate.

It is important to note that Label Propagation Algorithm operates on graphs, hence the input data must be converted into a graph. In our experiment, it was necessary to performed on "fashion vs. discount" case study a dimension reduction with UMAP, grouping customers with similar properties into clusters and determining centroids of each cluster accordingly.

In order to create a graph, the "k-Nearest Neighbors algorithm" (known as k-NN) was used that is one of the simplest, but perfectly fitting into the experiment context, and as a non-parametric method it is commonly used for classification and regression. For classification, the centroids with Euclidean distance between them are used and transformed into the normalized distances for all nodes while filtering out all that above a given threshold. More details about the results of LPA and k-NN on real marketing data will be given in section 5.

A graph was created, where the nodes represent clusters generated by HDBSCAN and edges are weighted links connecting clusters, determined by applying k-NN algorithm from the previous step and representing normalized distances between clusters. Centroids defined while executing UMAP method were crucial in the creation of a proper graph for LPA method.

Finally, using the data from the previous steps, a large graph was created, consisting of 2046 clusters (vertices) and

15364 links as represented in Fig. 7. The graph is shown only for demonstrational purposes, where lengths of links do not illustrate the real distances between clusters (weights), nonetheless cluster proximities have been preserved. Multiple dense cluster groups can be noticed. These are candidates to form communities and in compliance with the definition of a community, they have many connections within the group and few to clusters outside the group. More detailed information about the graph structure will be given in the next section.

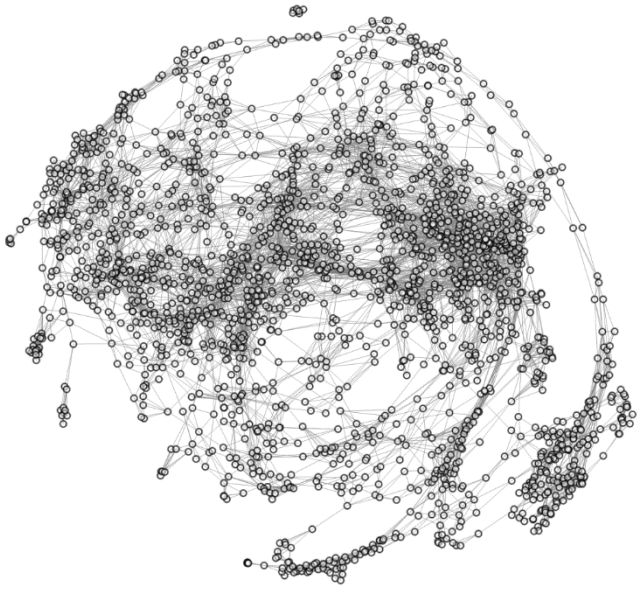

Fig. 7 Visualization of the graph representing communities and connections between them

Densely connected groups reach a common label quickly. When many such dense groups are created throughout the network, they continue to expand outwards until it becomes impossible to do so. Randomization of the order the clusters are processed has consequences: it may not deliver a unique solution, or the final solution may not be found at all (due to fluctuations in label assignment, adjacent clusters can interchange their labels in every iteration, preventing the convergence criteria from being achieved).

In our tests all partitions found were similar to each other, though.

Finally, the graph looks as in Fig. 8. The densest groups of clusters have been marked in color on the graph. They form communities characterized by common attributes.

V. Interpretation of the Experimental Results

In retail businesses, managers would want to know about the customers in order to efficiently tailor offers for selected groups, and to increase efficiency and customer satisfaction, which in turn increases business profitability. On the other hand, there might be a group of customers that may abuse the system, searching for system weaknesses and resulting in a loss or very small benefit for the retailer. The promise of this experiment was to find clusters of most profitable

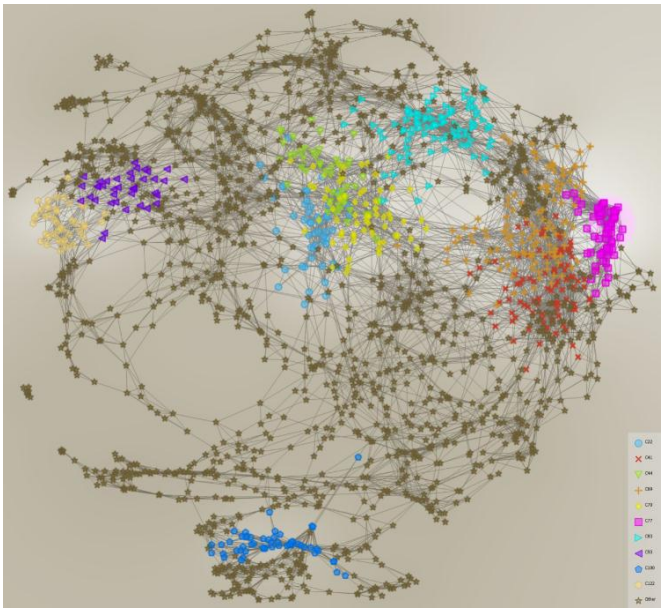

Fig. 8 Graph of clusters with highlighted communities

customers what has been already stated in previous part, and to provide marketing manager or analyst with appropriate knowledge about customer behavior respecting the value of products, discount, and “age” of the product.

In the experiment above Label Propagation Algorithm has been applied on clusters of customers having similar purchase characteristics and identified groups (communities) of similar clusters.

Let us analyze two groups: C77 and C122 among all groups identified by LPA in the experiment, as visualized below (Fig. 9 and Fig. 10).

First group C77, marked in pink color, consists of customers buying goods present in the shop for several months, but always with a price discount. The second group C122, marked in orange, seems to be very similar to C77, but it represents customers buying goods with the highest price discounts. For these two groups the measures applied by managers should be different. For the first group it could be running a marketing campaign in order to increase the average price of the order, while the second group can be used to address seasonal sale campaigns at the late stage

If a manager is interested in clusters of customers with the highest order values (as potentially most beneficial customers), they can filter clusters using the order-value property. Selected clusters of customers who buy goods valued at more than 350 PLN are presented in Fig. 11. There are 133 groups (communities) created by LPA out of 2052 clusters generated by HDBSCAN. On the left-hand side of visualization one dot represents one HDBSCAN cluster and the color of the dot represents LPA cluster (after community detection by LPA).

If a manager is interested in some specific clusters, they can observe the distribution of each feature in clusters. If one takes into consideration clusters C2 and C7 presented in Fig. 11, one can observe that these are customers not looking

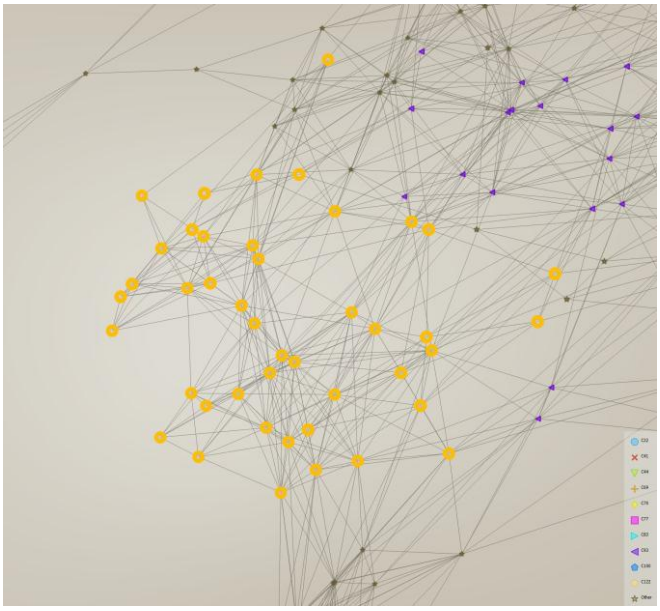

Fig. 9 Clusters of C122 group forming a community

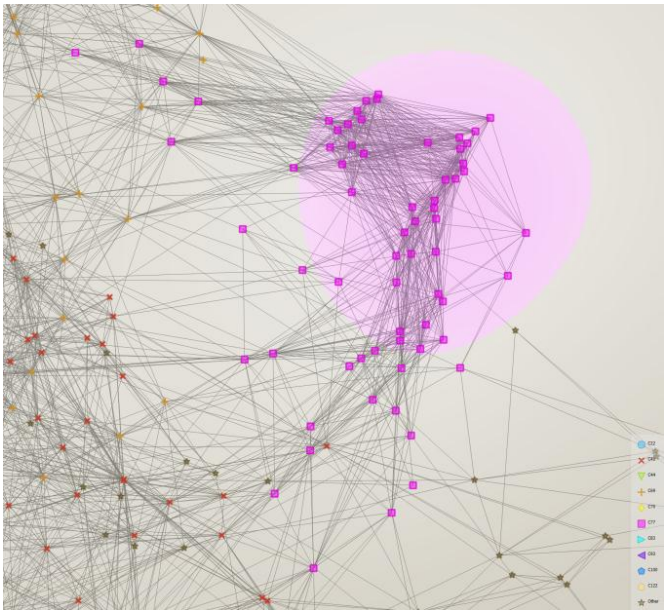

Fig. 10 Clusters of C77 group forming a community

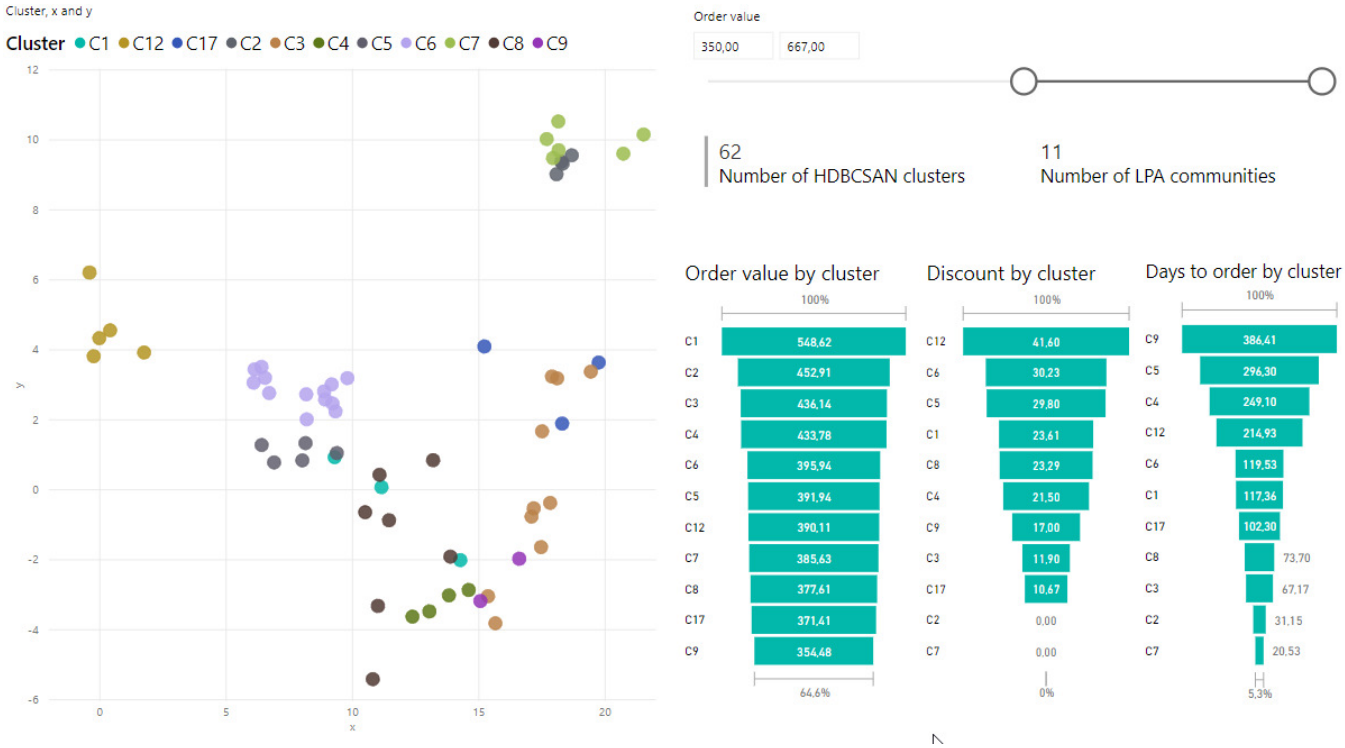

Fig. 11 Graphical representation of analyzed clusters meeting manager's criteria

.

for discounts (they buy with 0% discount), they buy new products (launched accordingly 31 and 20 days before purchase). The difference between those clusters is in the average order value (respectively 452 and 385 PLN). Having such knowledge, the recommendation system makes it possible to tailor the offer in order to meet customer's expectations.

For seven selected clusters meeting the assumed criteria (Fig. 12), the scatter within the cluster was calculated using Davies-Bouldin index in order to compare results with k-means result. In this experiment, the measure of scatter was between 14.44 and 46.27. Average value for selected 7 clusters is 34.13 which in comparison with the value of k-means 80.47 constitutes a significant improvement.

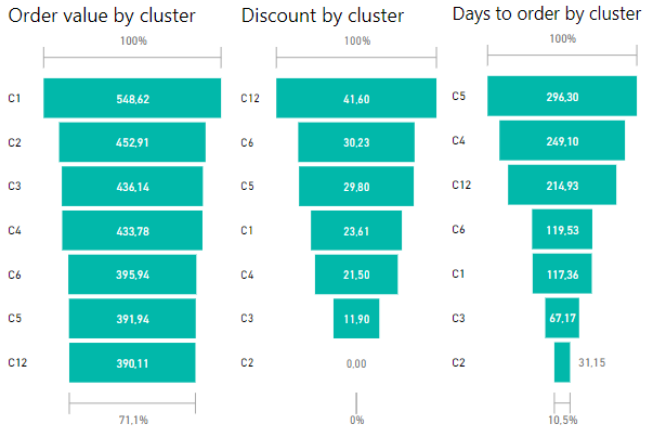

Fig. 12 Interpretation of selected clusters

VI. Conclusions and Future Research

The primary objective of the presented research was to develop a method to discover meaningful customer communities, using data mining techniques and tools. The outcome of this work is a new multi-level hybrid method for community discovery, implemented and validated on the experimental platform Upsaily. The research methodology is composed of six, closely integrated steps. Firstly, relevant information about customers is extracted from large marketing databases and partitioned by the clustering algorithm (HDBSCAN). Secondly, the space dimensions are reduced into two dimensions using the Uniform Manifold Approximation and Projection (UMAP) method. Thirdly, the centroids are computed for each cluster. The graph is generated in step four using the k-NN algorithm. To discover customer communities the Label Propagation Algorithm (LPA) is applied. The final step, most important for decision makers, concerns marketing interpretation of discovered customer communities. These experiments demonstrated that the "customer communities discovery" compared against "segmentation with k-means algorithm", gave much more precise identification of group of customers and allows better understanding of clusters by managers and data analysts.

The multi-level clustering approach described in this paper has shown its advantage over single method clustering. Numerous small clusters were turned into communities sharing common properties. Specifically, running HDBSCAN alone against the data describing customer's purchases resulted in a high number (2046) of dense, but small clusters, making it infeasible to predict customer's needs or address tailored offerings. It is important to mention that using the simplest PCA method for dimensions reduction did not meet expectation. The clusters did not form homogenous communities, which in turn could not provide managers with reliable tools to support decision making processes. When the PCA method was replaced by the UMAP method, the clustering results met expectations and made it possible to calculate meaningful centroids for each cluster. Afterwards, the label propagation method was applied, making it possible to determine customer communities, grouping them based on business needs.

The Upsaily platform used in the experiment, allows for parameterization of the multi-level approach to clustering, described in the paper, by defining the features used for clustering, business-oriented cluster identification, defining data range or specifying the size of expected clusters. The advantage of this customized approach is that it can be widely applied to any type/category of customers and it allows for performing advanced analytics on the business data.

The results obtained so far on real marketing data are very encouraging, in addition they have been positively validated by managers of internet shops. However, many algorithmic and business-oriented issues remain to be extended and tuned. For instance, a desirable extension of the approach will be to refine a method of feature construction describing a customer profile. An interesting future improvement will be on the implementation of collective and cooperative clustering with built-in business-oriented quality measures. One, but not the last, ambitious work will be focused on the dynamics and evolution of customer communities.

References

[1] Barber M. J. (2007). Modularity and community detection in bipartite networks, Physical Review E.,76(6):066102, DOI:10.1103/PhysRevE.76.066102

[2] Codaasco G., Gargano L. (2011). Label propagation algorithm: A semi-synchronous approach, Internat. Journal of Social Network Mining, 1(1):, pp.3-26, DOI:1504/IJSNM.2012.045103

[3] Gregory S. (2010). Finding overlapping communities in networks by label propagation, New J. Pys., 12, 103018, DOI:10.1088/1367-2630/12/10/103018

[4] Han J., Li W., Su Z, Zhao L. and Deng W. (2016). Community detection by label propagation with compression of flow, e-print arXiv:161202463v1, DOI:10.1140/epjb/e2016-70264-6

[5] Liu W., Jiang X., Pellegrini M., Wang X. (2016). Discovering communities in complex networks by edge label propagation, Scientific Reports 6, DOI:10.1038/srep22470

[6] Rossetti G., Cazabet R. (2017). Community Discovery in Dynamic Networks: A Survey, arXiv:1707.03186, DOI:10.1145/3172867

[7] Wu Z.H. et al. (2012). Balanced multi-label propagation for overlapping community detection in social networks, Journal of Comp. Sc. And technology, 27(3), pp. 468-479, DOI:10.1007/s11390-012-1236-x

[8] Subelj L., Bajec M. (2014). Group detection in complex networks: An algorithm and comparison of the state of the art, Physica A: statistical Mechanics and its Applications, 397, pp. 144-156, DOI:10.1016/j.physa.2013.12.003

[9] Gordon S., Linoff M., Berry J.A. (2011). Data Mining Techniques for Marketing, Sales, and Customer Relationship, Wiley, ISBN:978-0470650936

[10] Aggarwal C.C., Reddy C.K. (2013). Data Clustering: Algorithms and Applications, Chapman & Hall / CRC, ISBN:978-1466558212

[11] Gan G., Ma C., Wu J. (2007). Data Clustering: Theory, Algorithms, and Applications, SIAM Series, DOI:10.1137/1.9780898718348

[12] Pondel M., Korczak J. (2018). Recommendations Based on Collective Intelligence–Case of Customer Segmentation. In Information

Technology for Management: Emerging Research and Applications (pp. 73-92). Springer, Cham, DOI:10.1007/978-3-030-15154-6_5
[13] Witten I.H. et al. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, ISBN:978-0128042915
[14] Raghavan U.N., Albert R., Kumara S. (2007). Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E. 76,036106, DOI:10.1103/PhysRevE.76.036106
[15] Rosvall M., Bergstorm C. T. (2007). An information-theoric framework for resolving community structure in complex networks, Proc. Natl. Acad. Sci., 104, pp. 7327-73-31, DOI: 10.1073/pnas.0611034104
[16] Xie J.R., Szymanski B.K. (2014). LabelRank: a stabilized label propagation algorithm for community detection in networks, Proc IEEE, Network Science Workshop, pp. 386-399, DOI: 10.1109/NSW.2013.6609210
[17] Applebaum W. (1951). Studying customer behavior in retail stores. Journal of marketing, 16(2), 172-178, DOI: 10.2307/1247625
[18] Clover V.T. (1950). Relative importance of impulse-buying in retail stores. Journal of marketing, 15(1), 66-70, DOI: 10.1177/002224295001500110
[19] See-To E., Ngai E. (2019). An empirical study of payment technologies, the psychology of consumption, and spending behavior in a retailing context. Information & Management, 56(3), 329-342, DOI: 10.1016/j.im.2018.07.007
[20] Rustagi A. (2011). A Near Real-Time Personalization for eCommerce Platform. In International Workshop on Business Intelligence for the Real-Time Enterprise (pp. 109-117). Springer, Berlin, Heidelberg,DOI: 10.1007/978-3-642-33500-6_8
[21] Kaptein M., Parvinen P. (2015). Advancing e-commerce personalization: Process framework and case study. International Journal of Electronic Commerce, 19(3), 7-33, DOI:10.1080/10864415.2015.1000216
[22] Campello R.J., Moulavi D., Sander J. (2013, April). Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining (pp. 160-172). Springer, Berlin, Heidelberg, DOI:10.1007/978-3-642-37456-2_14
[23] Pondel M., Korczak J. (2017). A view on the methodology of analysis and exploration of marketing data. In: Federated Conference on co-algorithm to detect community structure in large-scale networks, Phys.Rev. E. 760360106 Computer Science and Information Systems (FedCSIS), IEEE, pp. 1135-1143, DOI:10.15439/2017F442
[24] Schubert E., Sander J., Ester M., Kriegel H.P., Xu X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 19, DOI:10.1145/3068335
[25] Ester M., Kriegel H.P., Sander J., Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231), ISBN:1-57735-004-9
[26] McInnes L., Healy J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint arXiv:1802.03426
[27] Newman M.E.J. (2004). Detecting community structure in networks. Eur. Phys. J. B 38(2), 321-330, DOI:10.1140/epjb/e2004-00124-y
[28] Fortunato S. (2004). Community detection in graphs. Preprint arXiv:0906.0612, DOI:10.1016/j.physrep.2009.11.002
[29] Emmons S., Kobourov S., Gallant M., Börner K. (2016). Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. PLOS ONE 11(7): e0159161. DOI:10.1371/journal.pone.0159161
[30] Blondel V.D., Guillaume J.L., Lambiotte R., Lefebvre E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008(10):P10008. DOI: 10.1088/1742-5468/2008/10/P10008
[31] Waltman L., Eck N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. The European Physical Journal B. 86(11):1–14. DOI:10.1140/epjb/e2013-40829-0
[32] Rosvall M., Bergstrom C.T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences. 105(4):1118–1123. DOI:10.1073/pnas.0706851105
[33] Zhu X., Ghahramani Z. (2002). Learning from labeled and unlabeled data with label propagation (p. 1). Technical Report CMU-CALD-02-107, Carnegie Mellon University