Proceedings of the Federated Conference on Computer Science and Information Systems pp. 801–810
DOI: 10.15439/2018F217
ISSN 2300-5963 ACSIS, Vol. 15

Collective clustering of marketing data— recommendation system Upsaily

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

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

***Abstract*—The article discusses the importance of the recommendation systems based on data mining mechanisms targeting the e-commerce industry. The article focuses on the use of clustering algorithms to conduct customer segmentation. Results of the operation of many clustering algorithms in segmentation inspired by the RFM method are presented, and the method of collective clustering using the positive effects of each algorithm is separately presented.**

I. Introduction

THE first seminars and conferences of the 90s on advisory systems [1],[2],[3] were a significant stimulus for the rapid interest in the methods and techniques of automation of recommendations not only in practice, but also by research. In recent years, under the influence of IT development, social networks, and artificial intelligence methods, the concept of the recommendation system and the scope of its main functionalities has significantly expanded. Today, the recommendation system constitutes a complex interactive system that allows one to determine the rank of a product or preferences that the customer should assign to a given product or group of products [4]. In the literature, this system is considered in three main perspectives. From the managerial perspective, the recommendation system is a decision support system that uses large, heterogeneous data and mechanisms generating recommendations related to the sales strategy and promotion of the products offered. From the client's perspective, it is an advisory system facilitating selection of products in accordance with one's interests, needs, and preferences. From an IT perspective, the recommendation system is an interactive computing platform containing a number of data analysis and exploration models, integrated with transactional systems of the online store and the environment. This platform must guarantee not only access to various information resources, but also scalability of applications operating on a large number of information collections.

The specific economic benefits of a personalized recommendation achieved by e-commerce tycoons (Amazon, Alibaba, eBay, Booking, etc.) have proven the increasing effectiveness of recommendations systems. It has resulted not only in increased sales and marketing effectiveness, but also in significant analytical and decision support for marketing managers. Modern recommendation systems are not limited to giving the recommendation *"You bought this product, but others who bought it, bought / watched X, Y, Z products"*. Many of them have based their recommendations on the customer profile, product characteristics, behavioral, and psychological analysis of customers.

Currently, the systems are distinguished by four categories of advisory mechanisms: recommendation by collaborative filtering of information, content-based recommendation, knowledge-based recommendation, and hybrid recommendation [5],[6],[7],[8],[9],[10]. Recommendation by collaborative filtering is the most common method based on recommending products highly rated by clients with similar profile and preferences [11],[12],[13]. The key issues here are: designation of the similarity between clients and choosing the customer segmentation method. These issues will be discussed in more detail later in the article. The content-based recommendation is founded on the analysis and data mining of products purchased by the customer [1],[14]. In contrast to the previous method, the key issue here is to analyze a customer's purchase history and determine the similarity of the products. The third group of methods builds recommendations based on analysis of product features with reference to its usefulness for the client [15]. In order to take advantage and reduce the negative features of the aforementioned methods, hybrid recommendation systems are increasingly being designed [16],[17].

For several years, we have also been observing a growing interest in recommendation systems by owners and managers of online internet shops in Poland. In 2010, every third online shop used a recommendation based on a simple analysis of CBR and Business Intelligence systems [18]. In recent years in Poland, artificial intelligence, personalized recommendation, and digital marketing have dominated the orientation of developers of e-commerce systems which until recently had focused on the efficiency of shopping services [19]. Currently, almost all big online stores use recommendation systems. However, these systems are to a large extent based on a simple business analytics, limited computational intelligence and reduced possibility of dynamic customer profiling.

The aim of the article is to present methods of analysis and profiling of clients available in the Upsaily[1] recommendation system targeting online internet stores. It is a hybrid system combining recommendation techniques through collaborative filtering and through contextual analysis. In the development of recommendations, in addition to transactional data, the system also uses geo-location and social network data. The data is a source of information for many clustering algorithms in the system. These algorithms can work autonomously or collectively, cooperating with each other in order to achieve semantically rich segmentation that is interesting in business interpretations. This second approach is the subject of the article. Although the source data set is the same, the innovativeness of the solution manifests itself in the selection of algorithms; each of them was selected from a different computing class and applies different similarity criteria. Among the algorithms, in addition to the commonly used k-means that uses Euclidean distance measure, we chose for the Gaussian Mixture Model based on probability distributions the DBSCAN algorithm taking into account the density of observation and the RMF involving the manager engagement. The unification of clustering results in our application is specific to the e-commerce applications – not all the clusters are used, but only one or several clusters. The cluster selection criteria include both statistical metrics as well as external, mainly economic, criteria.

The structure of the article is as follows. The next section describes the main functionalities of the Upsaily recommendation system and sketches its functional architecture. The third section defines the problem of individual and collective clustering together with descriptions of the applied algorithms. The concepts of similarity and criteria for unification of clustering results have also been outlined. The last section of the article describes the experiments carried out and further shows the advantages of collective clustering on real marketing data.

II. Functional Architecture and Functionalities of Upsaily System

The Upsaily system, based on the B2C model, is oriented towards current customers of the online internet shops. In the system database, not only all customer transactions are stored, but also basic data about their demographic and behavioral profile. The system is able to record customer reactions to offers directed at them through various contact channels. Functionally, the system can be classified as a Customer Intelligence solution, i.e. the one whose primary interest is current customers, and the aim is to increase customer satisfaction that translate into increasing turnover through the Based on literature [21],[22],[23],[24] and drawing conclusions from the research carried out as part of the RTOM project [25], the schema of advanced data analysis in marketing has been proposed (fig. 2). The schema is helpful in organizing marketing activities. Depending on the specific purpose, a group of clients to be covered by the campaign should be selected. In general, for the defined clients, the subject of the campaign is selected, e.g. product groups that customers making follow-up purchases, increasing the value of individual orders by cross-selling or more valuable products (up-selling). The immediate goal of the system is not to help in acquiring new customers. 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 clients and send them a personalized marketing message [20].

The results of research conducted as part of the RTOM project on Polish online stores operating in various industries that showed that in each of them over 75% of all customers are one-off customers, meaning they never returned to the store after making a purchase form the basis of such orientation of the system. Analysis of the average value of the order for a one-off customer shows that it is lower than for customers who make subsequent purchase. Interestingly, it can be noticed that the general trend of an increase in the average value of the order with the increase in customer loyalty expressed in the number of purchases made by them. This observation is presented in Figure 1. The average value of orders have been hidden due to the company's confidentiality. From this observation, it was concluded that it is worth sacrificing the resources of the online store to build customer loyalty, for the simple fact that a loyal customer is ultimately more valuable than a one-off customer.

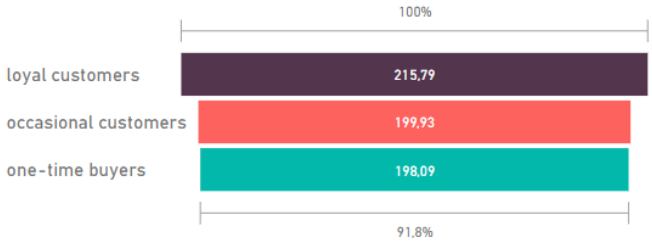

Fig. 1 Graph of the dependence of the average order value on the total number of orders placed by customers.

It should also be pointed out that acquisition of a new customer is always related to the extra cost to be incurred to reach the customer with the marketing message in a selected medium. Usually by acquiring a client then sending them a general message. Without knowing the customer's previous transactions, we are unable to propose an effective offer tailored to client's preferences, therefore in many cases the presentation of a marketing message will not cause projected customer reaction. In case of communication with current clients whose contact details are available and for whom all necessary marketing consents are established - at least at the assumptions level, it can be stated that reaching the customer should cost significantly less and the effectiveness of messages should be definitely higher.

[1]Upsaily system was developed by the Unity S.A., Wrocław, in the framework of the Real-Time Omnichannel Marketing (RTOM) project, RPO WD 2014-2020.

they will potentially be interested in. The final stage is defining the conditions under which customers will be offered participation in the campaign. As the schema shows, cluster algorithms have a wide application in this approach, and this will be shown later in the article.

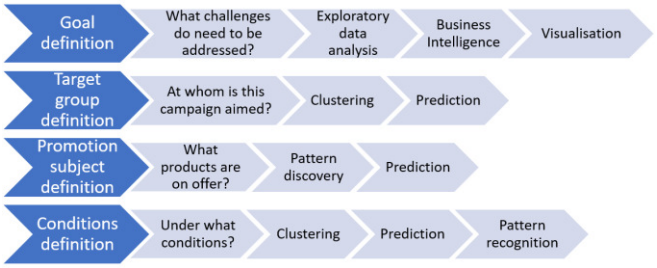

Fig. 2 Stages of building a marketing message with the proposal of using methods and tools for data analysis.

The functional architecture of the recommendation system Upsaily is presented in Figure 3. The Upsaily system collects data from many sources, but the basis of its analysis is transactional data. Data from other sources such as marketing automation systems, social media, systems analyzing activity on the store's website enrich the customer profile and, thus, expand the set of input data for analytical modules that, thanks to them, are able to provide better analyzes and better predictive models. The research platform on which the experiments are carried out has a significant place in the architecture of the system.

These experiments are evaluated in terms of business suitability and when their effects are positive, then they are transformed into regular modules operating in a production manner.

The system information outputs are integrated with:

- Marketing panel or application presenting the results of conducted analyzes, visualizing identified trends, found patterns, and segmentation effects. The recipient of this application are primarily managers and marketing analysts who, in using it, expand their own knowledge on the clients and their behaviors,
- A real-time recommender, an application whose aim is to offer an online store an offer that is as congruent as possible with its needs.
- Module "*campaign for today*", which is based on discovered trends and customer behavior patterns, at the moment of launching it is able to automatically indicate groups of customers, and the product that they may be interested in at that moment.

The results of the Upsaily system will be detailed in the next sections of the article.

III. COLLECTIVE CLUSTERING ASSESSMENT METHODS

There are many algorithms that can be used in collective clustering approach [26], [22], [23]. In the project the composition idea was based on maximum variability and differentiation of clustering paradigms. Therefore the following algorithms were chosen:

- k-means based on the Euclidean distance between observations,
- Bisecting k-means acting on a similar basis to k-means, however, starting with all the observations in one cluster and then dividing the cluster into 2 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,
- DBSCAN identifying clusters by measuring density as the number of observations in the 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.

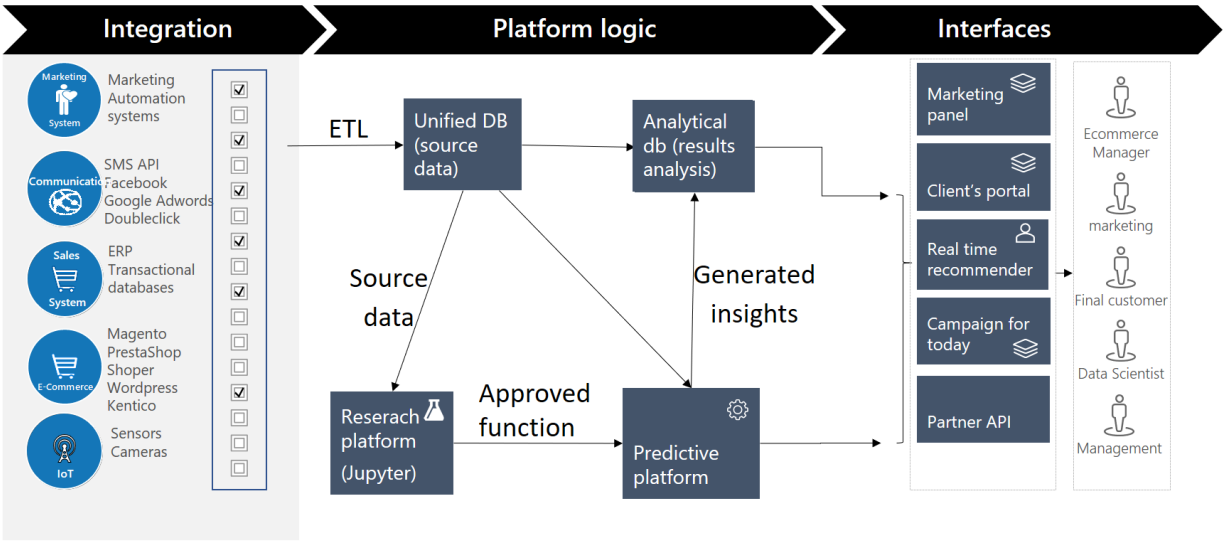

Fig. 3 Functional architecture of the Upsaily system.

Usually the results of clustering algorithms are evaluated according to internal and external criteria. The internal criteria relate to the hierarchy of clusters, taking into account the similarity of observations within clusters and the similarity

between clusters. The Davies-Bouldin[2] and Dunn[3] metrics are usually applied for assessment measures. In addition to the mentioned measures, other functions of assessment are used, such as the silhouette index, measures of cluster cohesion, cluster separation measure, and intra-class scattering matrix [26],[27].

According to external criteria, the results of clustering are evaluated using external data, not considered in the clustering process. Such data are observations which membership in the cluster is assigned earlier by experts. Then the assessment of clustering results from comparing of the content of clusters marked by experts with clusters created by the algorithm. Among the measures used, one should mention the clusters homogeneity index[4] , Jaccard index[5], Rand index[6] . In addition to the specified measures of the assessment, other indicators are also used, such as Kappa, F-score, Fowkes-Mallows index, etc. [26],[27].

In the case of using many clustering algorithms, the obtained results usually differ from each other not only by the number and hierarchy of clusters, but also by the allocation of observations to clusters. In the article, we treat the set of algorithms as a collective of experts whose task is to make the grouping of the set of observations from the business point of view as best as possible. Discrepancies in grouping that appear in the results of the algorithms must be minimized. The solution to this problem is determined by the unification process.

In order to assess the results of clustering, it is often helpful to assign a category to collected observations. In the case of very large data sets, it is not possible to assign all observations by experts. Therefore, it has been proposed to enable the assignment of observation to the clusters through decision rules that define clusters selected by the expert, in the form:

$$X_i \in C_j \mid if\ [(w_{11} \cap w_{12} \cap \ldots \cap w_{1k}) \cup (w_{21} \cap w_{22} \cap \ldots \cap w_{2m})\ \ldots]$$

Where Xi is a given observation, C_j is a cluster in the conditional expression. Attributes used in conditional clauses indicate their importance and usefulness in the characteristics of clusters.

The decision rules are determined by the algorithm of inductive decision tree algorithm C4.5 [28]. These rules make in possible, on the one hand, to interpret the obtained clusters and, on the other hand, to symbolically determine the observations belonging to individual clusters. This solution enable finding of similar semantic clusters generated by different algorithms. The symbolic interpretation of clusters is complemented graphically, which facilitates a quick identification of similar clusters. It should be noted that these works generally require significant involvement of marketing analysts.

In general, in the recommendation systems, the manager is only interested in a few clusters describing similar clients, similar products, or similar transactions. Therefore, for the analyst the first task involves identifying clusters that are still subject to unification. Although the task can be performed algorithmically, our experience has shown that much better results are obtained through selection of clusters by the analyst. If the visual selection is difficult, finding for a cluster C_i, a counterpart among clusters $C_j \in C_k$ obtained from another algorithm, then the formula of similarity between clusters $S(C_i,C_j)$ can be applied:

$$S(C_i,C_j) = max\ (|C_i \cap C_j| / |C_i|).$$

In cases where the cluster's observations C_i are distributed into several clusters from C_k, the assignment should take into account the distribution of S values and the weights of related cluster similarities.

After selecting the clusters obtained from different algorithms, one can start unifying the results. There are many methods of unification [29]. The most commonly used methods are the following:

- Consensus methods [30],[31],[32],[33], which are used more in the first phase of unification to create initial clusterization than to unify the results
- Multi-criterial grouping methods [30],[31] are mainly used to harmonize the criteria of different algorithms,
- Clustering methods supported by domain knowledge [35],[36],.

The last group of unification methods was used in the Upsaily system. The domain knowledge of marketing has been used to direct the unification process of selected clusters. In the system, the earlier created decision rules were used to govern the process of unification, in particular, the conditional expressions of which are treated as grouping constraints. The idea of the proposed method consists in determining semantic relationships-constraints indicating observations that must be included in the cluster (called must-link), and those that

[2] The Davies-Bouldin index is computed according to the formula: $DB = 0.5n\Sigma\ max\ ((si + sj) / d(ci,cj)$ where n is the number of clusters, the cluster centroids, si and sj mean d distances between the elements of a given cluster and the centroid The algorithm that generates the smallest value of the DB indicator is considered the best according to the criterion of internal evaluation.

[3] The Dunn index is calculated according to the formula: $D = min\ (d(i,j) / max\ d'(k)$ where $d(i,j)$ means the distance between clusters i i j and $d'(k)$ the measure of distances within the cluster k. The Dunn index focuses on cluster density and distances between cluster. Preferred algorithms according to the Dunn index are those that achieve high index values.

[4] Cluster homogeneity index is computed according to the formula: $CH = 1/N\ \Sigma max\ |\ m \cup d\ |$ where M is the number of clusters created by the algorithm, D is the number of expert classes.

[5] The Jaccard index measures the similarity between two sets of observations according to the following formula: $WJ = TP / (TP + FP + FN)$, where TP means True Positive error, FP False Positive, FN False Negative. In the case of two identical sets of $WJ = 1$.

[6] The Rand measure is calculated according to the formula: $WR = (TP + TN) / (TP + FP + TN + FN)$. The Rand index, as well as the previous ones, is based on a comparison with the benchmark given by an expert. It informs about the similarity of the assessment of correct decisions between the results of the clustering algorithm and the benchmark.

should not be included in it (called cannot-link). In order to improve the quality of clusters, fuzzy logic is proposed in some works [37],[38],[39] or characteristics of clusters such as values of inter-cluster distances, density [40], [41].

Let us now follow the entire unification process step by step aiming to achieve consensus on the content of the final clusters without a significant loss of quality of the partitions. Let us assume that they were pre-designated as similar two clusters C_j i C_l, each generated by a different algorithm. As indicated, the interpretation of each cluster is given in the form of decision rules, namely:

C_j | *if* $[(w'_{11} \cap w'_{12} \cap ... \cap w'_{1k}) \cup (w'_{21} \cap w'_{22} \cap ... \cap w'_{2m})$ ...]

C_l | *if* $[(w''_{11} \cap w''_{12} \cap ... \cap w''_{1k}) \cup (w''_{21} \cap w''_{22} \cap ... \cap w''_{2m})$ ...]

The final cluster can be created by merging of conditions containing variables (attributes) indicated by the analyst based on domain knowledge. This operation can be called a subsumption according to which the more detailed condition are covered with a less detailed one. However, the resulting cluster may contain too many observations that are too far away from the class sought (as shown in Fig.4). In narrowing the cluster's space, the observations given earlier by the expert might help, defined as a must-link or cannot-link marked in Fig. 4 in green and red respectively.

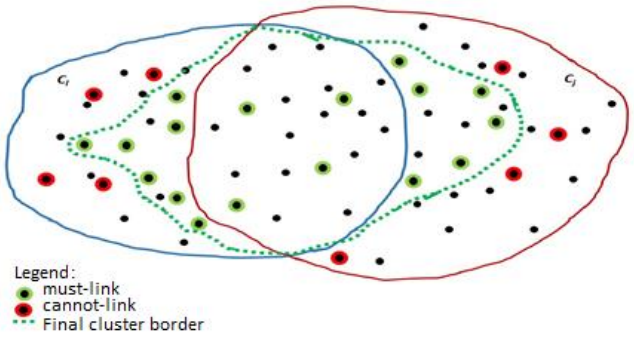

Fig. 4 Example of space of merged clusters.

The boundary of the final cluster (green dotted line) is determined between the sum of observations belonging to two clusters minus the surroundings of ε observations belonging to the *can not-link* relationship and the intersection of the observation plus the surroundings ε observations belonging to the *must-link* relationship. It can therefore be noted that the unified cluster includes observations lying in the space $|C_i \cup C_j|$ - ε *Xi/cannot-link* and $|C_i \cap C_j|$ + ε *Xi/must-link*. The radius of the surroundings ε can be determined based on ½ distance between the closest observations belonging to the *can not-link* and *must-link* relationships.

After the first unification of clusters, the process should be repeated for all similar clusters obtained from all algorithms. It should be noted that the order in which the clusters are selected influences the calculation time. We suggest choosing the most numerous clusters of interest in the first place. The next chapter will show examples of unification of the results of collective clustering.

Due to the thematic orientation of the conference and the restricted volume of the article in the next chapter, we will concentrate only on the business assessment of the results of clustering (domain knowledge). The RFM analysis will be used which is a traditional approach to analyze the customer behavior in the retail industry. Its acronym comes from the words "recency" (period from the last purchase), "frequency", and "monetary value". In this type of analysis, customers are divided into groups, based on information on time which has elapsed from last purchases, how often they make purchases, and how much money they spent (see [42]).

The following observations explain why RFM is interesting for retail companies:

- Customers who have recently made purchases are more likely to make a new purchase soon
- Customers who frequently make purchases are more likely to do more shopping
- Customers who spend a lot of money are more likely to spend more money

Each of these observations corresponds to one of the dimensions of RFM.

In the next section, the usefulness of this approach for assessing clustering algorithms is shown on the real marketing data.

IV. The Results of Experimental Research

In order to show the usefulness of the collective clustering method in specific business conditions, this chapter presents an experiment aimed at finding customer segments with similar behavior on the market. The clustering method should support a process of customer assignment to particular segment, assessment of proposed segments and interpretation of characteristics of these segments. The segmentation example was inspired by the RFM method. The customer is described by the following characteristics: frequency of their purchase (frequency dimension), the number of days which has passed since the last order (recency dimension) and the average order value (monetary value). We extended the customer description by information about the number of orders. Such dimension is essential in the case of an online store in order to determine the loyalty customer. The customers were divided into 6 segments. For each segment, we calculated its value (the sum of all customers' orders from a given segment). The number 6 was chosen arbitrarily. Marketing employees were able to prepare 6 different marketing communication policies addressed to individual customers. With more segments, it would be very difficult for the marketing analyst to interpret segments and subsequently develop a tailored communication policy for selected customers. A larger number of segments will be justified only if the automatic recommendation mechanism uses this segmentation.

The experiment was carried out using three clustering algorithms: bisecting k-means, Gaussian Mixture Model and

DBSCAN[7]. After each experiment, an expert evaluated the results of the segmentation. The analysis covered 56 237 customers who made at least 2 purchases in the online store.

When assessing segmentation, it is very helpful to visualize the data. Having 4 dimensions and ability to present it on surface (with only two dimensions). We used two methods for projecting the multidimensional space into a smaller number of dimensions. In order to prepare the visualization in the experiment, the four dimensions were reduced to two (X and Y), while the color means the segment number to which the given customer was assigned. One of those methods is The Principal Component Analysis[8] (PCA). PCA is a popular technique for reducing multidimensional space [43].

An example of RFM segmentation using the k-means algorithm and visualization using the PCA method is presented in Figure 5. One dot represents one real customer on the visualization (on left hand side of picture). After hovering over the selected dot, one can read the values describing the selected customer. This solution will help the marketing analyst to understand the prepared segments.

In the right part of the report there are funnel charts, presenting the average value of the given dimension attribute in individual segments; for example, average customer from segment 1 purchases with frequency of 14.22 days.

The column chart located in the bottom right corner of the report shows the sum of customers' orders in a given segments. It can therefore be observed that the highest revenue was generated by customers from segment no. 6, while the smallest in segment no. 5.

Another method of reducing dimensions that is useful for visualization is the Uniform Manifold Approximation and Projection (UMAP) [44]. It is a novel manifold learning technique for dimension reduction. UMAP seeks to provide results similar to t-SNE, which is the current state-of-the-art for dimension reduction for visualization, with superior run time performance. A theoretical framework of UMAP is based on Riemannian (a non-Euclidian) geometry and algebraic topology. In overview, UMAP uses local manifold approximations and patches together their local fuzzy simplicial set representations.

It is based on the approximation of the local manifold (local manifold approximations) and fuzzy simplicial sets. In contrast to a simple method such as PCA, where the projection is mainly based on two dimensions, the UMAP method takes all dimensions into account equally. An example of a visualization made using the UMAP method is presented in Figure 6.

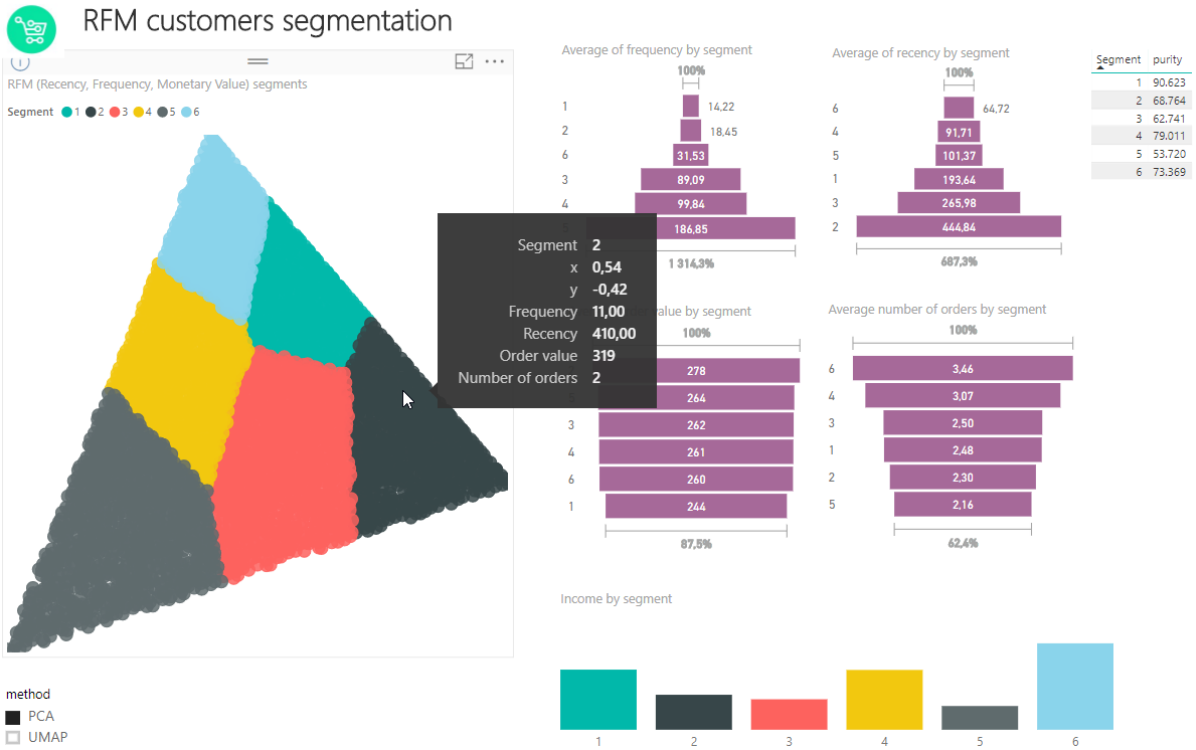

Fig. 5 Segmentation using the k-means algorithm and PCA visualization.

[7] The HDBSCAN algorithm was used, which is an extension of the DBSCAN algorithm. A library available on the GitHub platform was used for this purpose: https://hdbscan.readthedocs.io/en/latest/index.html

[8] The purpose of the PCA method, in brief, is to find a linear subspace (in our case 2-dimensional) in which the variance after projection remains the largest. The PCA method, however, is not to easily reject the dimensions with the lowest variance. It builds a new coordinates system in which the remaining values are the most diverse.

Visualization using two methods as well as presentation of the values of individual dimensions in clusters allow the analyst to better understand the individual customer segments and make an expert assessment of clustering.

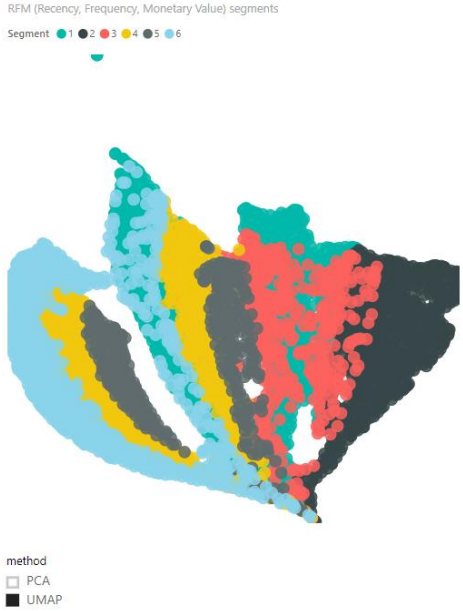

Fig. 6 Visualization of segmentation using the UMAP method.

Clustering using the k-Means algorithm based on the Euclidean distance between observations has many drawbacks. These include the fact that, when assigning customers to segments, the most varied dimensional values have the greatest impact (in our case, recency and frequency). The other dimensions impact less, and this can be observed in the low differentiation in the dimensions of average orders' values. In addition, it should be noticed that the boundaries between individual segments are not sharp. For example, segment 6, with the lowest average value, recency dimensions, includes both customers with a value of 0 and customers with a value of 147, these customers from the perspective of the RFM method, made their purchases relatively long time ago. The main advantage of this algorithm is the fact that the segments are relatively well balanced (their size is relatively similar). It makes those segments worth creating a dedicated marketing policy.

The next algorithm of clustering used in the experiment was the bisecting k-means. In the case of this algorithm, greater diversity was observed in individual dimensions than in the case of k-means. The clusters were again relatively balanced, however, the problem of the slight diversification of the 1 dimension remained, and in some segments there were clients located far away from the average value on a given scale.

Subsequent clustering was performed using the Gaussian Mixture Model algorithm. That method resulted with significant differences in the value of individual dimensions, due to which we can observe interesting cases of outliers (e.g., segment 1 includes customers with a very large number of orders and very high value of orders). Unfortunately, the size of such segments is relatively small (in this case 34 customers), which makes the legitimacy of building a special communication policy targeted to the customers from such a segment questionable. The same experiment was repeated for the DBSCAN algorithm. In case of this algorithm, the number of clusters was defined. Algorithm takes as parameter only the minimum size of the cluster. The disadvantage of this approach is the fact that a large part of the observations were not assigned to any cluster, and also that the majority of clusters are very small. The advantage is that the average values of the dimensions in the indicated segments are very diverse. The use of this algorithm to build communication policies is therefore debatable, but its advantage is the fact that clusters of relatively few but very similar observations are found, which can be used in the automatic recommendation mechanism.

For the marketing analyst, in order to perform the clustering using all the mentioned algorithms, they should observe the boundaries identified by algorithms on individual dimensions, and then those borders to build their own clusters, which will be referred to as according to their interpretation, e.g.

If average order value> 1000 zł
and number of orders > 10 and recency < 300
and frequency < 60
then segment=„active frequent valuable buyers"

If average order value< 200 zł
and number of orders < 3 and recency > 250
and frequency > 200
then segment=„occasional past cheap buyers"

In the platform, client filtering for clustering assignments can be done "manually" using the provided "sliders" presented in the upper right corner in Figure 7.

In the last phase of the experiment, a collective segmentation was proposed, taking into account the results of the three selected clustering algorithms. Because of similar results of the k-means and bisecting k-means algorithms, the k-means was not finally used in the experiment. We created collective segments basing on the results of 3 algorithms. The label of new clusters is constructed with the 3 numbers of clusters generated by the algorithms: bisecting k-means, GMM, and DBSCAN. For example, cluster 326 means that the customer has been assigned originally to clusters with numbers: 3 - bisecting k-means; 2 - GMM; 6 – DBSCAN.

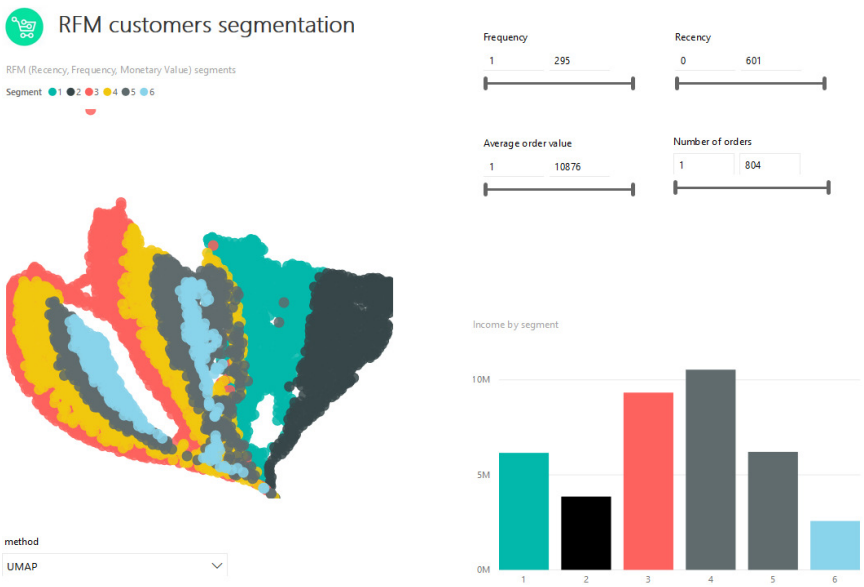

Figure 7 Manual segmentation. Source: Own elaboration in the Upsaily.

As a result, 52 segments were created (on 216 possible combinations), which is presented in Figure 8.

Such a large number of segments, of course, do not allow for an in-depth analysis of each of them and for "manual" preparation of marketing policies. However, these segments can be successfully used in the automated recommendation mechanism.

If the marketing analyst needs to analyze and interpret individual segments, in order to limit the number of clusters, similar segments may be merged. After the analyst decides on the maximum number of clusters or the minimum cluster size, then segments below the thresholds are included in the larger segment meeting the criterion of cardinality. Clusters' merge can be made with the lowest distance between them. The distance of clusters is not determined by the Euclidian measure, as for each of the aforementioned methods, cluster number is just an identifier without any meaning. Such identifiers do not determine similarity of clusters (e.g., cluster 1 doesn't have to be close to cluster 2). Taking the fact into account, distance in this case should be understood as the number of algorithms indicating a different cluster number, e.g., between clusters 525 and 520 the distance is 1 - which means that the clusters differ by the result of 1 method.

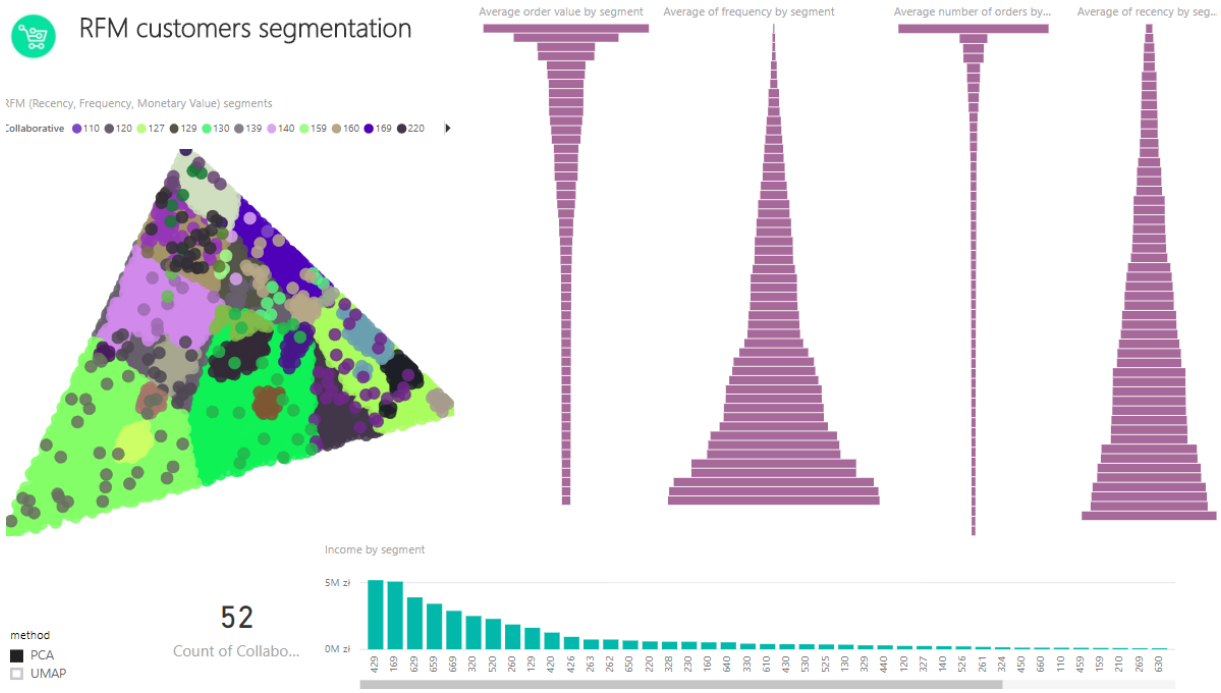

Figure 8 Visualization of 52 segments prepared using a collective approach.

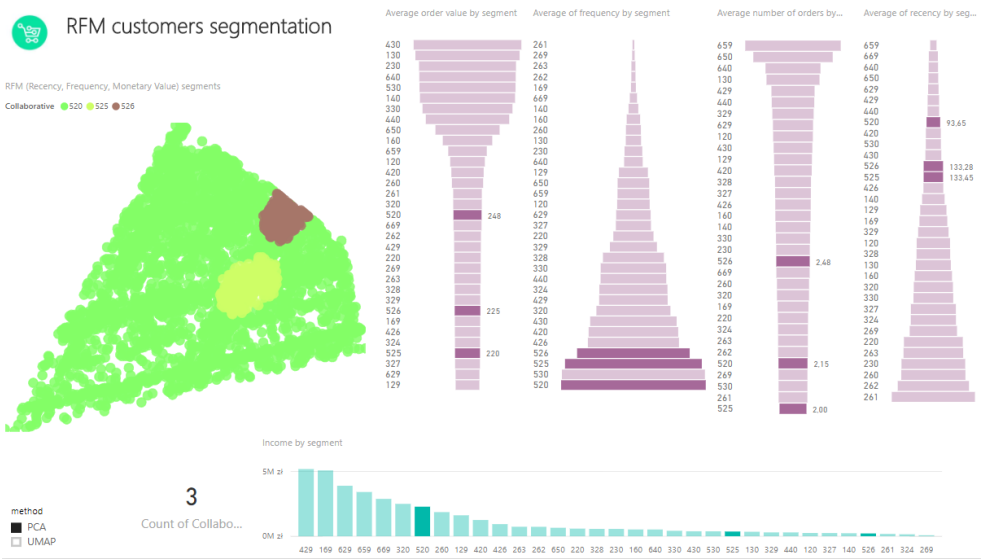

Figure 9 Visualization of the merge of 3 similar clusters prepared using a collective approach.

Between the clusters 320 and 525, the distance is 2. If clusters that should be merged are identified, a number of conflicts is encountered - clusters of the same distance. In this experiment, we will solve the conflict by selecting the highest cardinality cluster to which we attach a cluster that does not meet the criterion of cardinality.

Figure 9 illustrates an example of how to merge clusters 525 and 526 to cluster 520 (as the most numerous).

The k-means, bisecting k-means or GMM algorithms require a pre-determined number of clusters beforehand that we want to receive. The DBSCAN algorithm autonomously selects the number of clusters basing on other parameters, but in its case a large part of the observations are not included in any resultant cluster. We can state that DBSCAN cannot be used in case we would like to define marketing policies covering all clients, but is well suited for identifying smaller groups of observations that are very similar to each other.

Segmentation using a few selected algorithms gives more interesting results from the perspective of the marketing analyst than the segmentation using only one algorithm. First of all, the clusters obtained as a result of collective clustering have better and more useful marketing semantics. In addition, the analyst can decide on their own whether in using the described approach they focus on selecting the optimal number of large clusters, or analyze smaller clusters to identify hidden patterns of customer behavior.

V. Conclusions

The Upsaily system uses clustering as one of the methods for analyzing customer behavior in order to support generation of purchase recommendations. The RFM analysis answers the question when and what value products should be recommended to the customer. Other methods, such as association rules and sequential rules, additionally answer the question of what product / product category to offer to the customer. The Upsaily system also uses classification algorithms to refine the recommendations addressed to the customers.

Segmentation using one algorithm from the marketing analyst's point of view always has disadvantages such as small diversity of segments on particular dimensions or existence of segments with very low cardinality. In order to get rid of these indicated drawbacks and emphasize the advantages of each algorithm, we proposed a collective approach consisting in building a cluster by unification of the segmentation performed by the insights generated by all algorithms. Such segmentation gave us a result of more consistent segments with easier interpretation, however the final number of segments is definitely higher than when using each algorithm individually. Small segments can be useful in situations where we build an automatic mechanism of generating recommendations based on the client's assignment to the segment, where the large number of segments do constitute a problem. Segments consisting of a small number of customers are also useful in the task of identifying atypical clients as outliers.

If we want to provide a marketing analyst with a limited number of segments for the purposes of preparing a tailored marketing policy to each segment separately, then we suggest aggregating segments so that they meet the criterion of cardinality.

In future works, the authors will deal with the subject of collaborative clustering, automatic identification of the optimal number of segments and client clustering based on subsequent dimensions that also take their transactions and purchased products into account.

References

[1] Balabanovic, M., Shoham, Y., Content-based, collaborative recommendation. Com. of ACM 40(3), pp. 66–72, 1997.

[2] Goldberg, D., Nichols, D., Oki, B.M., Terry, D., Using collaborative filtering to weave an information tapestry. Com. of ACM 35(12), pp. 61–70, 1992.

[3] Resnick, P., Varian, H.R., Recommender systems. Com. of the ACM, 40(3), pp. 56–58, 1997.

[4] Konstan, J.A., Adomavicius, G., Toward identification and adoption of best practices in algorithmic recommender systems research. In: Proceedings of the international workshop on Reproducibility and replication in recommender systems evaluation, pp. 23–28, 2013.

[5] Beel, J., Towards effective research-paper recommender systems and user modeling based on mind maps. PhD Thesis. Otto-von-Guericke Universität Magdeburg, 2015.

[6] Jannach, D., Zanker, M., Ge, M., Gröning, M., Recommender systems in computer science and information systems–a landscape of research. In: Proc. of the 13th International conference, EC-Web, pp. 76–87, 2012.

[7] Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds): Recommender Systems Handbook, Springer, pp. 1–35., 2011.

[8] Jannach D., Zanker M., Felfernig A., Friedrich G., Recommender systems – an introduction, Cambridge University Press, 2010.

[9] Lu J., Wu, D., Mao M., Wang W., Zhang W.G.,, Recommender system application developments: a survey, Decision Support Systems, 74, pp. 12-32, 2015.

[10] Said, A., Tikk, D., Shi, Y., Larson, M., Stumpf, K., Cremonesi, P., Recommender systems evaluation: a 3d benchmark. In: ACM RecSys 2012 Workshop on Recommendation Utility Evaluation: Beyond RMSE, pp. 21–23, 2012.

[11] Acilar A. M., Arslan A., A collaborative filtering method based on Artificial Immune Network, Exp Syst Appl, 36 (4), pp. 8324-8332, 2009.

[12] Cornuejols A., Wemmert C., Gançarski P., and Bennani Y. Collaborative Clustering : Why, When, What and How. Information Fusion, 39, pp. 81–95, 2017.

[13] Kashef R., Kamel M.S., Cooperative clustering, Pattern Recognition 43, 6, pp. 2315–2329, 2010.

[14] Konstan J.A., Riedl J., Recommender systems: from algorithms to user experience User Model User-Adapt Interact, 22, pp. 101-123, 2012.

[15] Carmagnola, F., Cena, F., Gena, C., User model interoperability: a survey. User Model. User-Adapt. Interact. 21(3), pp.285–331, 2011.

[16] Burke, R., Hybrid recommender systems: survey and experiments. User Model. User-Adapt. Interact. 12(4), pp.331–370, 2002

[17] Lu J., Wu D., Mao M., Wang W., Zhang G., Recommender system application developments: a survey, Decision Support Systems, 74 , pp. 12-32, 2015.

[18] Kobiela E., Intelligent recommendation systems (pol. Inteligentne systemy rekomendacyjne), Network Magazyn, http://www.networkmagazyn.pl/inteligentne-systemy-rekomendacji, 2011

[19] Gemius 2017, The latest data on Polish e-commerce is now available (pol. Najnowsze dane o polskim e-commerce już dostępne), https://www.gemius.pl/wszystkie-artykuly-aktualnosci/najnowsze-dane-Polish-of-ecommerce-already-dostepne.html.

[20] Nazemoff V., Customer Intelligence. In: The Four Intelligences of the Business Mind. Apress, Berkeley, CA, 2014

[21] Chorianopoulos A., Effective CRM using predictive analytics. John Wiley & Sons, 2016.

[22] Gordon S. Linoff,M., Berry J.A., Data Mining Techniques: for Marketing, Sales, and Customer Relationship, Wiley 2011.

[23] Witten, I. H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[24] Jordan, M. I., MITCHELL, Tom M. Machine learning: Trends, perspectives, and prospects. Science, 349.6245, pp. 255-260, 2015.

[25] Pondel, M., Korczak, J., A view on the methodology of analysis and exploration of marketing data. In Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, pp. 1135-1143, 2017.

[26] Aggarval C. C., Reddy C.K, Data Clustering: Algorithms and Applications, Chapman & Hall / CRC 2013

[27] Gan G., Ma C., Wu J., Data Clustering: Theory, Algorithms, and Applications, SIAM Series, 2007.

[28] Quinlan J., Improved use of continuous attributes in {C4.5}. Journal of Artificial Intelligence Research, 4, pp.77–90, 1996.

[29] Wemmert C., Gancarski P., Korczak J., A collaborative approach to combine multiple learning methods. International Journal on Artificial Intelligence Tools (World Scientific), 9(1), pp.59–78, 2000.

[30] Strehl A., Ghosh J., Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research, 3, pp.583–617, 2002.

[31] Ayad H., Kamel M. S., Cumulative voting consensus method for partitions with variable number of clusters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(1), pp.160–173, 2008.

[32] Nguyen N., Caruana R., Consensus clusterings. In International Conference on Data Mining, IEEE Computer Society, pp. 607–612, 2007.

[33] Pedrycz W., Collaborative and knowledge-based fuzzy clustering. International Journal of Innovative, Computing, Information and Control, 1(3), pp.1–12, 2007.

[34] Faceli K., Ferreira de Carvalho A.C., Pereira de Souto M. G., Multiobjective clustering ensemble with prior knowledge. Volume 4643, Springer, pp. 34– 45, 2007.

[35] Law M. H., Topchy A., Jain A.K., Multiobjective data clustering. In IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pp. 424–430, 2004.

[36] Wagstaff K., Cardie C., Rogers S., Schroedl S., Constrained k-means clustering with background knowledge. In International Conference on Machine Learning, pp. 557–584, 2001.

[37] Belarte, B., Wemmert, C., Forestier, G., Grizonnet, M., Weber, C. Learning fuzzy rules to characterize objects of interest from remote sensing images. In Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE , pp. 2986-2989, 2006.

[38] Guo, H. X., Zhu, K. J., Gao, S. W., & Liu, T., An improved genetic k-means algorithm for optimal clustering. In Conference on Data Mining Workshops, 2006. ICDM Workshops. IEEE, pp. 793-797, 2006.

[39] Grira N., Crucianu M., Boujemaa N., Active semi-supervised fuzzy clustering. Pattern Recognition, 41(5), pp.1851–1861, 2008.

[40] Bilenko, M., Basu, S., & Mooney, R. J., Integrating constraints and metric learning in semi-supervised clustering. In Proceedings of the twenty-first international conference on Machine learning, ACM, p. 11,2004.

[41] Gancarski P., Cornueejols A., Wemmert C.,Bennani Y., Clustering collaboratif : Principes et mise en oeuvre, Proc. BDA'17, Nancy, 2017

[42] Linoff, G. S., Data analysis using SQL and Excel. John Wiley & Sons, 2015.

[43] Ghodsi, A., Dimensionality reduction a short tutorial, Department of Statistics and Actuarial Science, Univ. of Waterloo, 37, pp. 38, 2006.

[44] McLeland I., Healy J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv preprint arXiv:1802.03426, 2018.