# Survey on Technique and User Profiling in Unsupervised Machine Learning Method

**[1]Andri M Kristijansson and [2]Tyr Aegisson**
Industrial Engineering, University of Bifröst, Bifröst, 311, Iceland.
Tyraeg675@yahoo.com

**ArticleInfo**
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmc202202002
Received 25 April 2021; Revised form 15 July 2021; Accepted 12 October 2021; Available online 05 January 2022.


**Abstract** – In order to generate precise behavioural patterns or user segmentation, organisations often struggle with pulling information from data and choosing suitable Machine Learning (ML) techniques. Furthermore, many marketing teams are unfamiliar with data-driven classification methods. The goal of this research is to provide a framework that outlines the Unsupervised Machine Learning (UML) methods for User-Profiling (UP) based on essential data attributes. A thorough literature study was undertaken on the most popular UML techniques and their dataset attributes needs. For UP, a structure is developed that outlines several UML techniques. In terms of data size and dimensions, it offers two-stage clustering algorithms for category, quantitative, and mixed types of datasets. The clusters are determined in the first step using a multilevel or model-based classification method. Cluster refining is done in the second step using a non-hierarchical clustering technique. Academics and professionals may use the framework to figure out which UML techniques are best for creating strong profiles or data-driven user segmentation.

**Keywords** – Machine Learning (ML), User Profiling (UP), Unsupervised Machine Learning (UML), Internet of Things.

## I. INTRODUCTION

The Internet of Things (IoT), Neurology, Machine Intelligence, and Data Gathering have fuelled the thirst for data for rational decision and personalisation. The availability of vast volumes of datasets for the aims of dividing the client base, delivering personalised service, as well as collecting valuable knowledge offered by diverse data sources is a significant competitive edge for modern organisations. In data mining technique, deep learning is used to actionable insights from unstructured information. Machine Learning (ML) [1] is "the study of computer techniques to systematize the process of information accumulation from instances." It is classified into two types: supervised and Unsupervised Machine Learning (UML). There is no variable of goals in UML, and the input datasets are just supplied. Surveillance computer vision algorithms, on the other hand, are provided a particular purpose (- for example target variables). This research concentrates on the application of UML for segment customers and behavior analysis based on data. The practice of acquiring information particular to each person, either directly or implicitly, is alluded to as UP. A user profile often comprises data such as spatial, psycho - graphic, or behavioural data.

Unsupervised learning is a ML type whereby the learning data is presented to systems without pre-allotted ratings or labels [2]. The methods of unsupervised learning should, resultantly, first discover any physical prevailing structures in the training datasets. Clustering is a prominent example, in which the learning algorithm clusters its learning. Principal components analyses, in which the algorithms identify methods to condense the training data by finding which characteristics are most beneficial for distinguishing between various training instances and rejecting the rest, and categorizing examples into groups with comparable attributes. This in contrast to supervised learning that makes use of pre-allocated group classes in the learning algorithms (typically by humans, or from outputs of unsupervised classification algorithms). Other intermediary degrees of supervision include learning algorithms, in which simply numerical scores rather than extensive tags are supplied for each training instance, and semi-supervised learning, in which only a fraction of the learning information has been labelled.

Unsupervised learning has distinct advantages, integrating low workloads for auditing and preparing the learning data contrasted to supervised learning approaches that necessitates fundamental amounts of expert human labor to assign and confirm the initial tag and to liberate, utilize and recognize initially undiscovered trends, which could have been unnoticed according to [3]. This typically requires the application of unsupervised methods. Using a larger quantity of training information and achieving reasonable performance more gradually. During the experiential approach, this enhanced computing and storage criteria, as well as a prospective high sensitivity to objects or discrepancies in the learning algorithm that are highly irrelevant or identified as erroneous by an individual, but are given undue advantage by the unsupervised learning method.

Corporations are frequently unable to get actionable insights from data, resulting in a significant waste of chances, finances, and marketing initiatives. Furthermore, many marketing teams are unfamiliar with data-driven grouping approaches. Consider the number of groups and the classification algorithm to use are critical concerns in methodological considerations for data-driven segmentation. Furthermore, whereas statistical data techniques [4] are well-understood and

generally accessible, categorized and mixed data techniques are less common and clear. For example, unlike statistical information, categorical data lacks default ordered relationships on feature values, making creating distance measures and grouping methods more difficult. The creation, efficacy (i.e., precision), and efficiency of different UML algorithms were the subject of previous study. Furthermore, distinct clustering approaches could either manage massive data but just manage numeric and categorical qualities, or they can handle both kinds of information but are ineffective with massive data. As a result, choosing the right technique is a tough and time-consuming operation.

The study topic, parameters used to define objects, type of data, dataset quantity, information complexity, proximity measurements, and outliers are all important things to think about. None of the research, on the other hand, included a description of UML techniques as well as the dataset's numerous needs and features. With regard to significant data qualities, this study focusses on proposing an approach and model of UML approaches based on the two-clustering technique. The paradigm is designed to assist academics and practitioners in picking the best methods and, as a consequence, achieving accurate segmentation findings. The following is the study's question: What approach should be used to outline Unsupervised Machine Learning (UML) techniques for User Profiling (UP)? In that case, this survey has been organized as follows: Section II focusses on a methodology for the research. Section III presents a critical survey of a technique and paradigm for User Profiling (UP using Unsupervised Machine Learning (UML) method. Section IV presents a brief discussion of the survey whereas Section V concludes the survey.

## II. METHODOLOGY

The techniques provided in [5, 6, 7, 8, 9, 10] will be used to perform a comprehensive research study. The researcher may construct a method and framework by analyzing the basic ideas of UML methods. Google Scholar, Web of Science, and Scopus are among the science browsers that have been explored. Articles are sorted by relevance, with the headline, abstract, and date of publication being evaluated first. The papers are next evaluated based on the number of references and lastly by reading the whole content. Section III presents a literature survey of the paper.

## III. SURVEY

*Machine Learning (ML)*

In [11], computing Machinery and Intelligence marked the foundation of Artificial Intelligence (AI) in educational writings. Machine Learning (ML) has emerged as the preferred AI technique for producing effective methods. They suggest that the rapid decline in the cost of processing power, as well as the access of increasing volumes of data, are the two reasons driving the advancements in ML. In order to get insights from unstructured information, ML might be useful in data mining methods. ML is defined as "the research of computer techniques to systematize the approach to information acquisition from instances," as per experts. One key characteristic is that ML is capable of forming its own prediction model based on data and responses, rather than being designed to follow certain classification methods to produce outcomes. Unsupervised and supervised learning are the two primary kinds of ML methods, which are discussed in the following sub-sections.

*Unsupervised Machine Learning (UML)*

In [12], no goal parameter is given, and just required information is supplied in UML. UML is a term used to describe grouping and its variants. Grouping is a multidimensional approach that groups items in such a way that each item is comparable to the other items in the cluster while being distinct from objects in other groups. Studying customer behavior by finding homogenous groups of consumers, recognizing new product possibilities by clustering items or trademarks, connection discovery, or information minimization are just a few instances. Marketing strategy is one of the most important strategic concerns in sales, and clustering may be thought of as market segmentation. The accuracy of the (data-driven) target markets that are created determines the effectiveness of focused marketing campaigns. As a result, one of the advantages of clustering is that it allows an organization to adapt its offers to the demands of a specific consumer group, giving it a strategic edge in the industry.

In [13], factors used to define objects, type of data, data quantity, information complexity, estimation methods, outlier identification, and understandability are all important concerns and needs for clustering algorithms. Model-based, Grid-based, Density-based, Partitioning-based, and Hierarchical-based are the five most common core clustering techniques. The intensity, connectedness, and border of objects are used to distinguish them in density-based approaches. The intensity of objects is investigated in this study in order to discover the functionalities of databases that impact a specific item. The range of data items is divided into grids in grid-based approaches. Model-based approaches maximize the fit between both the data and a statistical model depending on the premise that the information contains a combination of fundamental likelihood distributions. Automatic determining the number of groups and accounting for outliers is possible using model-based approaches. In [14], researchers have built Neural Networks like Self-Organising Charts as illustrations.

Usability testing has been a key aspect of recommendation systems since their inception, according to research. However, the International Journal of Advance Foundation and Research in Computer (IJAFRC) has just published an article on UP. Search Personalization, Adapted Website, Customer Relationship Management System Adaptive Web Store are just a few examples. In this part, we'll look at a few examples of such implementations. A lot of the development has already been done on user profiles for research paper suggestions. In [15], scientists created the program, which divides user profiles into three subprojects: profile mining, integration, and interest identification. In [16], authors utilized a profile

display phase to depict a profile created by a system that employed an ontological method in a similar way. Another implementation that might profit from UP is an e-Tourism-based webpage.

Depending on the specified location, this technology was capable of giving tailored data. Because the tourism industry is so reliant on demographic data like location, the program was capable of making suggestions for nearby tourist attractions depending on a new specified location. Currently, power management is a critical issue. Many large businesses are grappling with how to handle energy more efficiently and effectively. In [17], the authors intelligent energy management technology has shown to be effective for this purpose. For smart energy management, they employed user profiles and micro accounting. Getting a job remains one among the most time-consuming tasks that everyone must do at some point in their lives. As a result, one of the fantastic ideas that this author has implemented is establishing a system that would automatically propose employment to users centred on their experience and qualification. Case-Based Profiling for Electronic Recruitment (CASPER) is the name of the system. The algorithm considers the user's profile information and makes employment recommendations to each person.

After some data about the customers has been gathered, it is required to divide the users into distinct groups in order to provide response to the program. This may be accomplished by categorizing users depending on their behaviour. This approach is also known as filtration, and it has been the subject of a lot of study. Content-based filtering and cooperative filtering are examples of filtering approaches.

Contextual filtering [18] also refers to content-based filtration. It chooses those things whose contents fit the contents integrated in distinct items depending on the contrast of the material of the products with the material of a user account. Each item's content is expressed by a collection of descriptors or keywords, which are often the words that are connected with it. The user identity is described in the same words and is created by assessing the content of things that the user has seen. This strategy is based on users' express scores or preference for a certain item, and it attempts to discover users who have provided comparable rankings for the item; however, in practice, users seldom offer explicit scores to the program. As a result, a system is required that will implicitly detect a user's rating or preference for a certain item.

Clustering people with similar interests' groups [19] is a common information retrieval strategy. This is predicated on the assumption that people who have consented to something in the past are likely to cooperate with it again in the coming years. As a result of this strategy, people with similar interests are organized into groups of peers, allowing the aforementioned notion of suggesting articles to the same group of customers to be realized. The success of this strategy is largely determined on how effectively the users' profiles are clustered. Alternative ways to inform the system have also been presented. Geographic filtering systems [20], for instance, employ data like education, gender, age, and location to determine the sorts of people who enjoy particular products. [21] utilizes the similarity fuzzy classification for profiles of users, which analyzes web logs to evaluate similarities between distinct users. Another technique employed a user centric profile based on a perceptive preferences questionnaire to examine features of perceptual preference based on knowledge processing, knowledge training and knowledge development. The scope of this work is restricted to a review of hierarchy and partitioning-based approaches. Furthermore, [22] examined the literature review for standards of different clustering approaches and discovered that most segmentation application (i.e. approximately 72%) applied either the hierarchical or the non-hierarchical classification techniques.

*Hierarchical and Non-Hierarchical Clustering*

The goal of hierarchical classification techniques is to locate a framework in the data (hierarchical) centred on the nearness medium that is showcased in a tree-like model known as dendrogram. The clustering algorithm could either be divergent (top-down) or agglomerative (bottom-up). The agglomerative clustering approaches with one item for every cluster and therefore integrates it with more than a single cluster that could be comparable in a stepwise manner. A divisive form works in the reverse manner, starting with the data as a single cluster and then recursively assigning items to the most relevant clusters. However, hierarchical approaches have the disadvantage of being unable to handle huge datasets or high dimensionality. One benefit of hierarchical approaches is that the classifications do not need specifications earlier. Ward's technique, Centroid's technique, Complete Linkage, and Single Linkage techniques are among the five agglomerative methodologies. Non-hierarchical classification techniques split data items into multiple divisions, each representing a cluster. Because they are substantially less costly, non-hierarchical approaches are often employed to handle enormous datasets. Clustering that isn't hierarchical may be Soft or Hard. Hard classification, also known as an exclusive cluster isolation, is used in the most basic approaches. Each item must be assigned to one of the groups. Techniques like fuzzy clustering, which are used in soft approaches, reduce this need.

*UML Algorithms*

A critical stage in classification study is assessing the technique and similarity metric for calculating item proximity. For constant information, similar metrics are pretty well known and commonly accessible, but for categorical information, it is not that simple. Categorical information, unlike continuous data, lacks default order connections on feature values, making building distance measures and grouping methods more difficult. Large, complicated, or high-dimensional datasets are a common feature in data extraction operations. Millions of items with hundreds of properties may be found in a single dataset. As a result, ML algorithms must be versatile and able to deal with a variety of characteristics. Because these sets of information are most typically found in actual statistics, classification methods that can handle huge data of numerical or categorical parameters are relevant. Most classification techniques, on the other hand, can either manage big data sets but

only have numeric characteristics, or they can handle both forms of information but are ineffective when dealing with large databases.

In [23], authors devised the k-means technique for non-hierarchical classification, which is particularly suitable for data extraction jobs since it can effectively handle enormous datasets. The mean of all points reflecting the arithmetic average is the center of the k averages method. It uses the (squared) Distance measure metric to repeatedly search for cluster centers and modify object memberships to minimize the Within Cluster Sum of The Squares (WCSS). The fact that k-means works best with statistical information is a disadvantage. The non-hierarchical k-modes approach was developed by the authors to cluster big categorized datasets. The main distinctions are that the k-modes employ a modest corresponding measure of distance (i.e. the hamming proximities) instead of the Euclidean distances, that cluster means are replaced by modes, and that cluster modes are updated using a frequency-based technique. The total discrepancies of relevant feature groups of the two objects are used to calculate the k-modes dissimilarity metric. As a result, the better the resemblance between things, the fewer incompatibilities there are. Additionally, as contrasted to k-means, kmodes is quicker since it corresponds in less iterations. In [24], authors developed a similar approach known as k-medoids, which considers medoids rather than centroids or phases. It's centered on the asteroid's most centrally situated item, therefore it's less prone to anomalies. As a result, k-medoids are good for categorizing information and dealing with outliers (i.e. noise), but they struggle with huge datasets.

The preceding non-hierarchical approaches are best for dealing with numerical or categorical properties. Real-world databases, on the other hand, often include a combination of data kinds. The k-prototype approach, which can group mixed-type objects and handle big databases and high dimensions, was developed by combining the k-means and k-modes methods. For numeric characteristics, the technique uses a squared Euclidean distance measurement, and for categorical features, it uses a simple corresponding dissimilarity measure. To prevent favoring a sort of characteristic in which the researcher's understanding of the data is a major influence, a particular weight is applied. There are many methods in the literary works for hierarchical clustering. CURE, a hierarchical technique for clustering huge datasets, was developed and implemented in [25]. To capture the structure and size of the cluster, the algorithm uses the dispersed points as representations. At each phase, the clusters are formed by combining the nearest pair of representing points. It can handle not just huge databases and also high dimensions and is more noise resistant since reducing dispersed points towards the mean decreases susceptibility to outliers.

It can only be used with statistical information, though. Chameleon, a hierarchical method based on adaptive modeling, was developed and utilized in [26]. In choosing the most comparable pair of classifications, it takes into account interconnection and proximity, which is a critical quality. Whenever the nearness and interconnection (proximity) of clustering is greater than the interconnectedness and closeness of items inside the cluster, two groups are merged. The Chameleon system's dynamic model of classifications is suitable to all sorts of information, huge databases, and high dimensions, according to the authors, as much as a similar matrix can be given. The ROCK method, which works with both categorical and numerical data, was developed in [27]. The ROCK method merges neighboring data points using a links-based metric rather than a distance-based one. While the ROCK method can handle huge datasets, it struggles with high dimensional and noise. The authors used 10 distinct similar indices to test the effectiveness of 10 different hierarchy clustering algorithms. Single, complete, average, centroid’s techniques, flexible-beta, density linkage, two-phase density linkages and the Ward’s approach were all investigated in the research.

Definitive linkage, Ward's technique, and flexibility beta were shown to be the best methods in terms of efficiency. Nevertheless, as contrasted to the Chameleon, ROCK, and CURE techniques, the later cluster formation approaches are computationally costly and sluggish when dealing with huge datasets and high dimensions. The model-based technique is frequently used in scholarly literature for classification, in addition to hierarchical and non-hierarchical approaches. Neural Networks have been a more common use for clustering algorithms in the literature, according to [28]. The Self-Organising Maps (SOMs) method, developed in [29], is the most widely utilized form of neural network, as per academics. Grouping, classifying, and forecasting models are all possible using SOMs. The purpose of SOMs is to transform a high-dimensional incoming signal into a finite map that is easier to understand. It's also utilized for dimension reduction and information visualization. SOMs group output nodes are grouped, with nodes closer together having more in common than nodes farther away. SOMs are less susceptible to activation, and the number of nodes is not necessary to be specified in advance.

*Two-Stage Clustering and Data Size*

Classifying algorithms are heavily influenced by the number of structures determined a priori. One of the greatest unresolved challenges in clustering evaluation is the difficulty of picking the number of clusters. Many literatures proposed a method depending on interior index comparisons as one of the initial attempts. [30], on the other hand, offered a two-stage clustering approach, recommending that clusters be identified first based on the Ward’s approach or the average interconnections (i.e. hierarchical classification), then cluster refining using non-hierarchical classification. They came to the conclusion that a two-stage method outperforms a hierarchy or non-hierarchical method alone. [31] used a two-stage strategy to clustering, using hierarchical and non-hierarchical clustering, and came to the same conclusion about getting better outcomes. [32] suggested using self-organizing charts (i.e. model-oriented) to evaluate the clusters using the k-means algorithm. The authors noted that their two-stage technique worked well for finding the starting segments and had fewer misclassifications than traditional methods.

As a consequence, using hierarchical clustering to determine the quantity of clusters before performing a non-hierarchical approach is a good way to get reliable clustering findings. When dealing with huge datasets or computational

complexity, hierarchical clustering algorithms are computationally costly and sluggish. As a result, a study is done to provide guidance on what data sizes are considered too big or too little. Non-hierarchical approaches often outperform hierarchical methods on big data sets, however hierarchical techniques' performance degraded as the number of data points rose. [33] looked at scholarly literature for classifying analysis approach for data-based segmenting markets and discovering that the bottom dataset had just 10 items, the biggest had 20.000, and the mean was 700. The average number of observations in hierarchical clustering approaches was 530, whereas non-hierarchical methods had 927. The databases' number of parameters varied from 10 to 66, with an average of 17 parameters. Thus, ten parameters may be considered low dimensionality, whereas more than ten parameters could be considered high dimensionality. On different information sizes, [34] has used hierarchical clustering algorithms. [35], for example, compared the effectiveness of both the hierarchical and non-hierarchical clustering technique on sets of data that ranges from 4,000 to 36,000 rows, with distinct levels and numbers of clusters.

Hierarchy classification worked best on a limited data with minimal dimensions, according to the findings. As a result, a data set with less than 4000 events might be regarded as small enough for clustering algorithms and processing time. With the exception of the Chameleon, ROCK, and CURE methods, databases with more than 4000 records may be deemed huge and less suited for hierarchical classification approaches. Because there are no information size restrictions, the only suggestion is to check whether the dimension is appropriate for the quantity of incidents to be categorized. A technique to evaluate the bottom sized data is to integrate at least 2k incidents (where k is the variables numbers), and preferably 2k * 5.

In summary, a clustering algorithm is useful when there are several classification solutions to consider or when the size of data is modest. Hierarchy classification may be used to identify the number of classifiers, and then a non-hierarchical technique groups all data using the nodes or previous capital points to generate precise clusters affiliations.

*Model for User-Profiling based on Unsupervised Machine Learning (UML)*

User-Profiling (UP) strategies that focus on UML and the dataset's specifications and features are visualized using a structure. The system is centered on the research presented in the subsection above. The type of data, quantity of data, and information dimensions all play a role in determining which method to use for UML difficulties. The quality and effectiveness of the classification technique and solution are greatly influenced by these data qualities. For example, k-means and k-modes may be used to analyze huge numerical datasets and large qualitative datasets, respectively. Nevertheless, [36] looked into the scholarly literature for classification evaluation specifications for data-driven market segments and discovered that the slightest information magnitude was only 10 artifacts, the greatest was 20, and the median size was 700. The sets of data had anywhere from 66 to ten parameters, with an average of 17 parameters.

Thus, ten parameters could be considered low dimensionality, while more than ten factors could be considered high dimensionality. [37] also compared the efficiency of the hierarchical and non-hierarchical clustering techniques of datasets with differing dimensions and cluster counts, ranging from 4000 to 36000. Clustering algorithm worked best on a limited data with low dimensionality, according to the findings. As a result, data sizes of less than 4000 may be considered minor enough for hierarchical classification, as well as its simulation time and understandability. With the exception of the Chameleon, ROCK, and CURE methodologies, data points with more than 4000 findings are considered significant and may be less appropriate for hierarchical clustering techniques. The above presumptions give a rough idea of what lower or higher number of dimensions and smaller or larger data sizes are. Nevertheless, they are still hypotheses, and there are no rules in academic literature that govern these classifications. The only suggestion is to consider whether the dimensionality is appropriate for incidents to the classified (2k incidents, especially 2k * 5).

Table 1 summarizes the classification methods in terms of the data attributes. Table 2 shows multiple Users Profiling strategies that focus on UML and data characteristics such as type of data, size, and dimensions. There are approaches for classification, computational, and mixed data in the system. To estimate the number of classifications and recognize initial seeds, the first stage includes a hierarchy or model-based classification process. Then, to provide more reliable cluster subscriptions, a non-hierarchical classification process is used.

**Table 1:** Presentation of the clustering algorithms and dataset attributes

| Algorithms Class | Algorithms | Type of data | Size of data | High-dimensionality Handling | Noise Handling |
|---|---|---|---|---|---|
| **Model-based** | SOMs [29] | Multi-variate datasets | Moderate/Small | True | False |
| **Hierarchical** | Chameleon [26] | Numerical/Categorical | Big | True | False |
| | ROCK [27] | Numerical/Categorical | Big | False | False |
| | CURE [25] | Numerical | Big | False | False |
| | Ward's/Complete linkages | Based on the measure of distance | Moderate/Small | False | False |
| **Non-hierarchical** | k-mode | Categorical | Big | True | False |
| | k-mediod [24] | Categorical | Small | True | True |
| | k-means | Numerical | Big | False | False |
| | k-prototype | Numerical/Categorical | Big | True | False |

**Table 2**: Model outline the UML algorithms for UP centred on a two-stage cluster and attributes of datasets

| Type of data | Size of data | Dimensionality | Stage 1 | Stage 2 |
|---|---|---|---|---|
| **Categorical** | Big | [High]/[Low] | [Chameleon]/[ROCK] | [k-mode]/[k-mode] |
| | Moderate/Small | [High]/[Low] | [Chameleon/Wards]/[Complete Linkages] | [k-mode-k-medoid]/[k-mode/k-medoid] |
| **Numerical** | Big | [High]/[Low] | [CURE]/[CURE] | [k-means]/[k-means] |
| | Moderate/Small | [High]/[Low] | [SOMs]/[SOMs] | [k-means]/[k-means] |
| | Small/Big | [High]/[Low] | [Chameleon]/[ROCK] | [k-prototype]/[k-prototype] |

Note: The dataset sizes of < = 4,000 is visualized as small or moderate. High-dimensionality is >10 variables and low-dimensionality is < =10 variables. There is a lack of guidelines concerning the data attributes in research.

## IV. DISCUSSION

Personalization has become more important in the area of computer engineering, particularly in the context of Recommendation Algorithms. A recommender algorithm must cope with a large number of users, each of whom has their unique set of preferences. The Recommender System must meet the demands of each user by either suggesting user-specific products or updating itself to meet those needs. As a result, user profile aids recommend technologies in understanding user needs and acting accordingly. The technique of collecting information about a person's interest domain is known as user profiles. This information may be utilized by the computer to learn more about the user, and this information can then be used to improve retrieving and ensuring the user's pleasure. UP involves two crucial aspects: effectively understanding people and suggesting things of interest depending on those individuals. In the Recommender System, this study attempts to investigate all features of a Classification Based system. This article looks at user profiles in three different scenarios. The article begins by identifying trends in UP to explain how it originated from recommender systems, then moves on to describing the approaches for profiling people, and finally concludes with several case reports on how users based has been utilized in different sectors.

*Trends*

UP's primary job is to collect data about individuals and their interests. In the subject of recommendation systems, much study has been conducted on profiling, and numerous profiling approaches have been developed throughout time. In general, UP has progressed as a result of the data mining and ML methodology. Its origins may be traced back to the cognitive data analysis paradigm, where many of the phases are similar to those engaged in the customer profiling procedure. UP is referred to in this research as the User Data Discovery (UDD) paradigm.

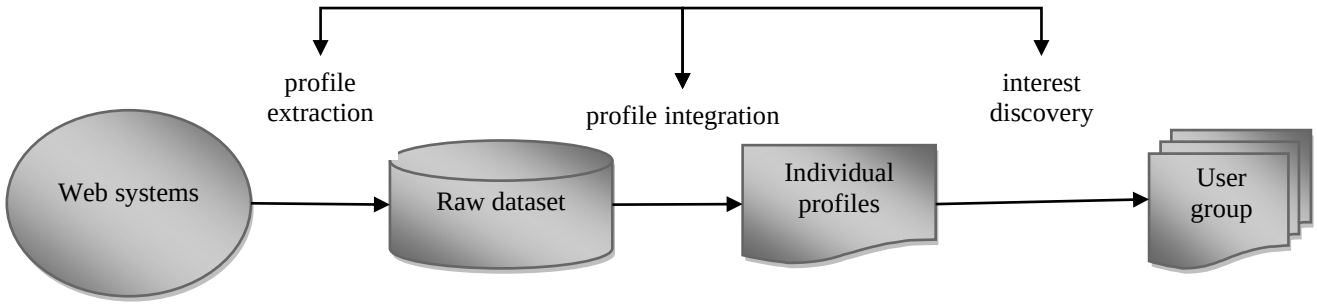


**Fig 1.** Discovery of user data

A simplified Knowledge Discovery in Databases (KDD) paradigm is given in the figure to highlight the similarity between the KDD and UDD processes. In the understanding Data Analysis method, the computer already has a large amount of data, but in the UDD framework, the program has very little data about the customer and must quickly acquire expertise about the user that may be utilized for subsequent actions. The basic goal of both systems is to gain information, but how they go about doing so is the focus here. One requires to extract data from massive datasets, i.e. it should evaluate the prevailing datasets to discover anything interesting, while the other instance involves data with a little amount of data but must nevertheless behave according to the user's expectations. In general, UP began with the simple extraction and collecting of information about the user. Older machines were more focused with receiving data directly from users, which meant that the system would actively ask the users for the information that was required. However, since the user is seldom interested in explicitly providing input, research is currently focusing on intuitively profiling user information based on behaviors taken by the user, a process known as behavioural UP. Many studies have been conducted on this topic, and we can distinguish three primary techniques to UP as specified:

*Explicit UP*

Explicit profiling, according to [38], is the practice of assessing users' static and regular traits. Users' behavior is anticipated using this method by examining the user's accessible data. This information is often obtained by filling out digital forms or survey participation. This is also referred to as static or actual profiling. When we rely only on explicit profiling, we run into issues such as users not wanting to expose their data to anybody because they are worried about their privacy, or the filling out forms' procedure being cumbersome, which the users aim to avoid. Resultantly, reliability of employing this profiling type declines with time.

*Implicit UP*

In [39] also describe an implicit profiling strategy in which, rather than focusing on the present data we have regarding the users, this approach concentrates on whatever we have studies concerning the customers previously, i.e. computers try to learn more about users. As a result, such a technology is also known as Behavioural profiling, adaptable profiling, or lately, user ontology profiling. Such profiling also employs a variety of filtering approaches. Rule-based filtering, text categorization, and content-based filtering strategies are only a few of the filtration approaches discussed in the academic literature.

*Hybrid UP*

In [40], the benefits of both intuitive and intentional UP are combined in this form of UP. In other words, it considers both dynamic and behavioural features of a user. As data is added in real time, this technique makes profiling more effective and ensures the correctness of temporal data.

*Model*

The objective of this contribution is to presents a technique and paradigm for UP and data-driven consumer categorization using UML methods. The capacity to segment a client base, deliver personalised services, and derive useful data from diverse information sources is a crucial competitiveness edge for today's companies. Nevertheless, in order to generate appropriate UP and segments, businesses often struggle to extract information from data and choose suitable ML methods. In addition, many marketing teams lack a basic knowledge of data-driven segmentation strategies. The number of nodes to use and the technique to use were the two most important factors. Furthermore, although quantitative information had a plethora of ways, category and hybrid data had fewer and less obvious options. Prior study centered on the creation, efficacy (i.e., reliability), and effectiveness of different UML methods. None of the research, on the other hand, included a description of UML techniques as well as the dataset's numerous needs and features. The following was the study's question: What approach should be used to outline UML Techniques for UP?

The essential principles of UML, as well as numerous techniques, two-stage grouping, and the features and needs for data attributes, were examined in the literature. A paradigm is provided that outlines multiple UML techniques for UP based on different data attributes. In terms of data size and complexity, it offers a two-stage clustering algorithm for categorized, numeric, and hybrid data. To calculate the number of clusters, the first step involves using a hierarchy or model-based clustering technique. Cluster refining is done in the second step using a non-hierarchical classification method. The methodology adds to a growing body of understanding on UML and data-driven categorization techniques and processes in marketing. Until date, no one has presented a framework that included a two-stage clustering approach for UML techniques, distinct datasets, and database attributes. A two-stage clustering strategy eliminates the disadvantages of employing just hierarchy or non-hierarchical clustering processes, resulting in more stable classification results.

## V. CONCLUSION

This research intends to participate in an answer by establishing a method and paradigm of UML techniques with regard to essential data attributes, based on the second classification technique. The paradigm is designed to assist academics and professionals in choosing the best methods and, as a consequence, achieving reliable segmentation accuracy. The following is the questionnaire method: What framework should we use to outline UML techniques for User Profiling (UP)? The program's practical consequences are that it may help academics and practitioners figure out which UML strategies are best for creating comprehensive user profiles and data-driven client segmentation for marketing. Furthermore, Mixed UP: The benefits of both implicit and explicit UP are combined in this sort of UP. In other words, it considers both the user's static features and the user's behavioural data. As data is updated in real time, this technique makes profiles more effective and ensures the correctness of temporal data.

## References

[1]. V. Vanchurin, "Toward a theory of machine learning," Mach. Learn.: Sci. Technol., vol. 2, no. 3, p. 035012, 2021.
[2]. S. Zhao et al., "A review of single-source deep unsupervised visual domain adaptation," IEEE Trans. Neural Netw. Learn. Syst., vol. PP, pp. 1–21, 2020.
[3]. F. Bröker, B. C. Love, and P. Dayan, "When unsupervised training benefits category learning," PsyArXiv, 2021.
[4]. J. Liu, L. Ding, X. Guan, J. Gui, and J. Xu, "Comparative analysis of forecasting for air cargo volume: Statistical techniques vs. machine learning," J. of Data, Inf. and Manag., vol. 2, no. 4, pp. 243–255, 2020.
[5]. A. M. Miksch, T. Morawietz, J. Kästner, A. Urban, and N. Artrith, "Strategies for the construction of machine-learning potentials for accurate and efficient atomic-scale simulations," Mach. Learn.: Sci. Technol., vol. 2, no. 3, p. 031001, 2021.
[6]. N. Käming et al., "Unsupervised machine learning of topological phase transitions from experimental data," Mach. Learn.: Sci. Technol., vol. 2, no. 3, p. 035037, 2021.

[7]. J. Inekwe, E. A. Maharaj, and M. Bhattacharya, "Drivers of carbon dioxide emissions: an empirical investigation using hierarchical and non-hierarchical clustering methods," Environ. Ecol. Stat., vol. 27, no. 1, pp. 1–40, 2020.

[8]. A. Carobene, A. Campagner, C. Uccheddu, G. Banfi, M. Vidali, and F. Cabitza, "The multicenter European Biological Variation Study (EuBIVAS): a new glance provided by the Principal Component Analysis (PCA), a machine learning unsupervised algorithms, based on the basic metabolic panel linked measurands," Clin. Chem. Lab. Med., vol. 0, no. 0, 2021.

[9]. N. K. Papadakis, "Unsupervised stochastic learning for user profiles," in Computational Mathematics and Variational Analysis, Cham: Springer International Publishing, 2020, pp. 279–297.

[10]. L. Zhang and M. Jin, "A two-stage clustering detector for SM-MIMO communications," IEEE Commun. Lett., vol. 25, no. 6, pp. 2019–2023, 2021.

[11]. L. Chen and N. Tokuda, "A unified framework for improving the accuracy of all holistic face identification algorithms: Electoral College for human face identification by computing machinery," Artif. Intell. Rev., vol. 33, no. 1–2, pp. 107–122, 2010.

[12]. J. Tian, Z. Teng, B. Zhang, Y. Wang, and J. Fan, "Imitating targets from all sides: an unsupervised transfer learning method for person re-identification," Int. j. mach. learn. cybern., vol. 12, no. 8, pp. 2281–2295, 2021.

[13]. F. Valdez, O. Castillo, and P. Melin, "Bio-inspired algorithms and its applications for optimization in fuzzy clustering," Algorithms, vol. 14, no. 4, p. 122, 2021.

[14]. S. Bersimis, A. Sgora, and S. Psarakis, "A robust meta-method for interpreting the out-of-control signal of multivariate control charts using artificial neural networks," Qual. Reliab. Eng. Int., no. qre.2955, 2021.

[15]. L. Bzhalava, J. Kaivo-oja, and S. S. Hassan, "Data-based Startup profile analysis in the European smart specialization strategy: A text mining approach," Eur. Integr. Stud., vol. 0, no. 12, 2018.

[16]. Z. Ye, Y. Guo, A. Ju, F. Wei, R. Zhang, and J. Ma, "A risk analysis framework for social engineering attack based on user profiling," J. Organ. End User Comput., vol. 32, no. 3, pp. 37–49, 2020.

[17]. S. Maga, PG scholar, Department of EEE,Dhanalakshmi Srinivasan,College of Technology,Chennai, Tamil Nadu, C. Kavitha, and Assistant Professor,Department of EEE,Dhanalakshmi Srinivasan College of Technology,Chennai, Tamil Nadu, "Power management in micro grid using hybrid energy storage system," Int. J. Bus. Intell., vol. 5, no. 1, pp. 116–120, 2016.

[18]. B. Y. Satria, A. Bejo, and R. Hidayat, "Fingerprint enhancement using iterative contextual filtering for fingerprint matching," in 2021 9th International Conference on Information and Communication Technology (ICoICT), 2021.

[19]. N. Claro, P. A. Salgado, and T.-P. A. Perdicoulis, "Subtractive mountain clustering algorithm applied to a chatbot to assist elderly people in medication intake," in Advances in Machine Learning, Data Mining and Computing, 2021.

[20]. K. A. Botangen, J. Yu, Q. Z. Sheng, Y. Han, and S. Yongchareon, "Geographic-aware collaborative filtering for web service recommendation," Expert Syst. Appl., vol. 151, no. 113347, p. 113347, 2020.

[21]. R. Belohlavek and M. Krupka, "Grouping fuzzy sets by similarity," Inf. Sci. (Ny), vol. 179, no. 15, pp. 2656–2661, 2009.

[22]. R. Souza de Oliveira and E. Giovani Sperandio Nascimento, "Clustering by similarity of Brazilian legal documents using natural language processing approaches," in Artificial Intelligence, IntechOpen, 2021.

[23]. A. R. Khan, S. Khan, M. Harouni, R. Abbasi, S. Iqbal, and Z. Mehmood, "Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification," Microsc. Res. Tech., vol. 84, no. 7, pp. 1389–1399, 2021.

[24]. D. Pan, Y. Han, Q. Jin, H. Wu, and H. Huang, "Study of typical electric two-wheelers pre-crash scenarios using K-medoids clustering methodology based on video recordings in China," Accid. Anal. Prev., vol. 160, no. 106320, p. 106320, 2021.

[25]. V. N. Phu, V. T. Ngoc Tran, and J. Max, "A CURE algorithm for Vietnamese sentiment classification in a parallel environment," J. Comput. Sci., vol. 15, no. 10, pp. 1355–1377, 2019.

[26]. A. Umamageswari, N. Bharathiraja, and D. S. Irene, "A novel fuzzy C-means based chameleon swarm algorithm for segmentation and progressive neural architecture search for plant disease classification," ICT Express, 2021.

[27]. J. Liu, X.-D. Zhao, and Z.-H. Xu, "Identification of rock discontinuity sets based on a modified affinity propagation algorithm," Int. J. Rock Mech. Min. Sci. (1997), vol. 94, pp. 32–42, 2017.

[28]. J. Chen, M. Xiao, Y. Wan, C. Huang, and F. Xu, "Dynamical bifurcation for a class of large-scale fractional delayed neural networks with complex ring-hub structure and hybrid coupling," IEEE Trans. Neural Netw. Learn. Syst., vol. PP, pp. 1–11, 2021.

[29]. K. Mohammed, A. Ayesh, and E. Boiten, "Complementing privacy and utility trade-off with self-organising maps," Cryptography, vol. 5, no. 3, p. 20, 2021.

[30]. H. Haddadpour and M. Emami Niri, "Uncertainty assessment in reservoir performance prediction using a two-stage clustering approach: Proof of concept and field application," J. Pet. Sci. Eng., vol. 204, no. 108765, p. 108765, 2021.

[31]. J. Park, K. V. Park, S. Yoo, S. O. Choi, and S. W. Han, "Development of the WEEE grouping system in South Korea using the hierarchical and non-hierarchical clustering algorithms," Resour. Conserv. Recycl., vol. 161, no. 104884, p. 104884, 2020.

[32]. A. Barger and D. Feldman, "Deterministic coresets for k-means of big sparse data," Algorithms, vol. 13, no. 4, p. 92, 2020.

[33]. Y. Y. Tan, "Exploring intuitive approaches to protein conformation clustering using regions of high structural variance," bioRxiv, 2021.

[34]. H. Zheng and J. Wu, "Which, when, and how: Hierarchical clustering with human–machine cooperation," Algorithms, vol. 9, no. 4, p. 88, 2016.

[35]. R. V. Kale, B. Veeravalli, and X. Wang, "Design and performance characterization of practically realizable graph-based security aware algorithms for hierarchical and non-hierarchical cloud architectures," in Lecture Notes in Electrical Engineering, Singapore: Springer Singapore, 2018, pp. 392–402.

[36]. S. Dolnicar, "Tracking data-driven market segments," Tour. Anal., vol. 8, no. 2, pp. 227–232, 2003.

[37]. T. H. Nguyen, Hanoi University of Science and Technology, D. T. Dao, and Vingroup Big Data Institute, "Cluster-based routing approach in hierarchical Wireless Sensor Networks toward energy efficiency using Genetic Algorithm," Journal of Science and Technology - Technical Universities, vol. 30.8, no. 147, pp. 14–21, 2020.

[38]. L. L. Costanzo, Y. Deldjoo, M. F. Dacrema, M. Schedl, and P. Cremonesi, "Towards evaluating user profiling methods based on explicit ratings on item features," arXiv [cs.IR], 2019.

[39]. Y. Lin, J. Su, Y. Liu, J. Hou, and F. Wang, "Implicit profiling estimation for semiparametric models with bundled parameters," arXiv [stat.CO], 2021.

[40]. A. Anjali, J. K. Sandhu, and D. Goyal, "User profiling in travel recommender system using hybridization and collaborative method," in 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021.