

Using the Latest Methods of Cluster Analysis to Identify Similar Profiles in Leading Social Networks

Bohdan Zhurakovskiy ¹, Ihor Averichev ² and Ivan Shakhmatov ²

¹ National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Peremogy Avenue, 37, Kyiv, 03056, Ukraine

² State University of Information and Communication Technologies of Kyiv, Solomyanska St., 7, Kyiv, 01310, Ukraine

Abstract

In the modern world, social networks are not only a showcase of individuality but also reflect the unique aspects of a user's personality. In light of this, networks accumulate colossal volumes of data daily, posing the challenge of deep analysis of this data to study communication dynamics and form interaction strategies. To address this issue, the potential of cluster analysis as a statistical analysis tool is considered, allowing the identification of homogenous groups based on data similarity or differences. For such analysis, clustering methods such as k-means, c-means, and others are examined, as well as their advantages and limitations in the context of social network analysis. Moreover, this information can be used to understand collective consciousness, analyze trends, and business activity. Delving deeper into the research, it's evident that, on the whole, it highlights the significance of cluster analysis in understanding social media users. This information expands the possibilities for understanding social interaction structures, studying current trends, shaping interaction strategies, and ultimately opens new horizons for scientists, analysts, and businesses in the social media sphere. Additionally, thanks to the proposed practical implementation within this research, a specific algorithm for detecting similar profiles in networks has been suggested. Its model is demonstrated using UML diagrams and Python code, and a series of tests confirm the high efficiency of the methodology. Thus, the use of cluster analysis becomes a crucial tool for analysts wanting to understand user behavior on social networks and identify significant trends. This study provides valuable insights into how data from social networks can be used to gain understanding beneficial for both scientists and businesses.

Keywords ¹

Cluster Analysis, Social Networks, Clustering, k-means Algorithm, Mini batch k-means, Big Data, Data Normalization.

1. Introduction

In the era of the global information age, every like, comment, and tweet reflects the personality of a user, transforming social networks into a showcase of individuality. Feedback on Facebook, business contacts on LinkedIn, reactions to current events on Twitter - each aspect of communication reveals nuances of the modern internet user. Social networks are a bottomless source of information, and here lies the challenge for researchers and businesses - how to extract valuable information from this massive data set? Cluster analysis reveals patterns and relationships, helping to understand user behavior, interests, and needs, much like the work of an archaeologist. But why are giants like Twitter or LinkedIn the focus? They reflect the collective consciousness of modern society. By analyzing activity on these platforms, one can get insights about current trends and business activity. These studies, in turn, provide valuable information for marketing, PR, and communication strategies.

We will focus on how cluster analysis reveals the understanding of social media users. We will consider different clustering methods, their characteristics, advantages, and disadvantages. This provides a better understanding of how these methods can be applied to analyze large volumes of data and detect deep trends in user behavior. The aim of this study is to identify effective ways to use cluster analysis in social networks. Through the study of user interactions, their interests, interactions, and community formation, we can


Information Technology and Implementation (IT&I-2023), November 20-21, 2023, Kyiv, Ukraine

EMAIL: bogdan68@ukr.net (B. Zhurakovskiy); iaverichev19@gmail.com (I. Averichev), ivan.shakhmatov@gmail.com (I. Shakhmatov)

ORCID: 0000-0003-3990-5205 (B. Zhurakovskiy); 0009-0008-9766-0115 (I. Averichev); 0009-0004-9628-0365 (I. Shakhmatov)

© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

uncover new opportunities for business, scientific research, and sociocultural analysis. Cluster analysis, where data is grouped based on certain characteristics, allows for the identification of patterns that might be hidden from a superficial analysis. Using examples of approaches, we will demonstrate how these techniques can be applied to achieve specific results. We will also look into specific clustering methods, tools, and approaches that researchers use to analyze social networks. We will also offer recommendations and best practices for those who want to apply these methods in their projects.

The practical implementation of this study focuses on creating a method to search for similar accounts on social networks. Standard clustering algorithms, such as k-means, might not always be optimal for the vast data of social networks. For this reason, it was decided to adapt and optimize the algorithm using a modification of mini batch k-means, which is more efficient for processing large data sets. This approach involves several stages: data acquisition and normalization, building a vector matrix, determining the optimal number of clusters, and applying mini batch k-means for clustering. In subsequent studies, the proposed product should allow the user to obtain a list of accounts that are similar based on certain criteria.

2. Literature review and problem statement

As the number of Internet users grows rapidly, the amount of data required for processing also increases. The use of conventional statistical methods to analyze such large amounts of data becomes inefficient, which makes it necessary to develop methods for analyzing groups of users by certain common features, which makes it possible to provide users with the kind of advertising that will interest them. Cluster analysis is used as a method for analyzing large amounts of data to create homogeneous groups based on certain characteristics. A significant contribution to the development of clustering data processing was made by such scientists as M. Jamboux in hierarchical cluster analysis and correspondence, I. Yenyukov in methods of clustering objects with categorization features, L. Meshalkin in classification and dimensionality reduction, and S. A. Ayvazyan in the development of classification of multidimensional observations.

3. Formulation of the problem

The active development of social networks in the global information space encourages modern technologies to solve the urgent task of developing and implementing a method of big data analysis to search for similar accounts in social networks based on cluster analysis.

4. The main section

The growing number of Internet users (Fig. 1) compels us to study the dynamics of interactions in the digital space. This is especially relevant in the context of social networks, where the dynamics offer perspectives for analysis in various fields, including marketing, sociology, psychology, and others. When analyzing big data, tools such as cluster analysis uncover patterns of behavior, showing which aspects of social networks are the most relevant.

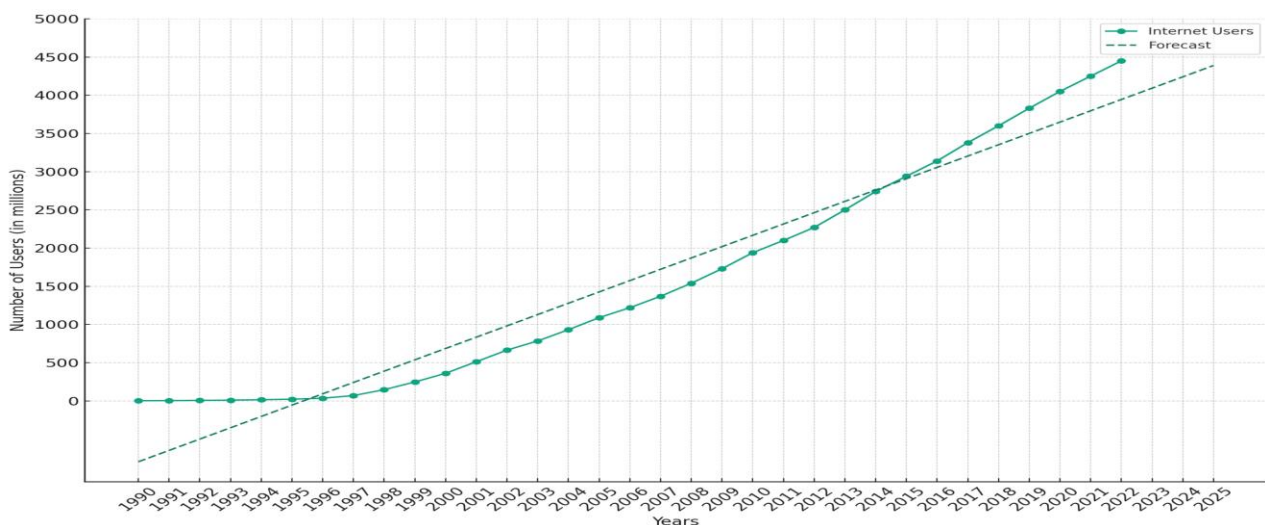


Figure 1: Dynamics of the increase in Internet users

Cluster analysis is becoming increasingly relevant in the context of social network analysis. A study on global scientific collaboration emphasizes the importance of social network analysis and data mining in the context of studying interactions in scientific publications [1]. This is corroborated by a study that considers the exchange of construction safety information in the Twitter environment, where social network analysis can detect and analyze data exchange in online communities [2]. From a technical perspective, clustering is important for identifying similar user accounts. Cluster analysis often employs machine learning methods based on PyTorch and Scikit-Learn [3]. Additionally, the importance of optimizing clustering algorithms for processing large data sets is highlighted in a study focused on fast clustering [4].

Some modern techniques, especially focused on social network analysis, propose using probability distributions of random variables (Gaussian distribution) for clustering Twitter users. These approaches allow processing large-scale data in real-time, adapting the analysis to the needs of dynamic social network data [5]. As for similarity assessment methods, there are algorithms that consider not only the primary but also additional user attributes. They help to increase the accuracy of recommendations by analyzing more detailed information about users [6]. Another important aspect to consider during clustering is the filtering out of inorganic profiles. Determining the authenticity of a Twitter account is critical for ensuring the quality of analytical data and increasing the reliability of results [7]. In the field of data analysis, there are various methods for grouping information. Additional methods, such as hierarchical clustering and the K-means algorithm, are noted for their effectiveness, especially when it comes to grouping users based on demographic behavior. These methods help reveal patterns and connections between different user groups, reflecting their similarities or differences in interests and behavior [8]. Regarding innovative approaches to clustering, some researchers propose viewing it not only as a tool for grouping data but also as a way to optimize decision-making processes in groups. It's interesting to note that clustering methods may play a key role in shaping group thinking and promoting more objective and reasoned decision-making [9, 10].

Cluster analysis, which is one of the primary techniques of machine learning and data analysis, is increasingly drawing the attention of scientists and analysts due to its potential. Its core value lies in the ability to effectively identify subgroups within complex data sets, highlighting patterns and relationships that may remain unnoticed with a superficial analysis. One of the key recommendations that researchers often emphasize in the context of cluster analysis is the need for the correct selection of sample size. To achieve quality results, it's ideal to have between 20 to 30 observations for each expected subgroup. Adhering to this recommendation ensures high reliability and accuracy of the results, which forms the basis for further analytical conclusions and decisions based on them. Multidimensional scaling plays a crucial role in the data analysis realm, offering researchers a powerful tool for examining intricate data sets. This method enhances the definition and separation of clusters by transforming multidimensional data into simplified, yet informative visualizations. The particular value of multidimensional scaling becomes evident when researchers encounter multi-dimensional normal distributions. In such cases, traditional methods might not always provide sufficient clarity. However, with "fuzzy" clustering or mixture modeling, one can achieve a more detailed and precise analysis that captures the essence of the data [11].

Social networks are becoming an endless source of information and user interaction, and it is here that cluster analysis demonstrates its immense potential. Researchers gain the opportunity to study global trends in scientific collaborations, as well as to develop more effective recommendation systems using adaptive machine learning methods. By considering additional parameters such as demographic information and user behavior patterns, even greater accuracy and relevance of clustering results can be achieved.

However, like any other research method, there are certain challenges. For instance, the Fuzzy C-Means (FCM) algorithm can sometimes introduce noise into the data, especially during lengthy computations, and its initial bias can affect the results. To address these issues, researchers propose a combined approach [12], which merges the capabilities of meta-learning and competitive learning. Such an integrated approach can be extremely effective, especially in challenging conditions where data sets are large or where the data structure constantly changes. With the growth in data volumes and increasing complexity of tasks, contemporary clustering methods are constantly evolving and improving. One of such innovative approaches is the hybrid method KM-SML. This method, which was thoroughly described in scientific paper [13], enhances the capabilities of the traditional K-means algorithm by integrating principles of supervised learning. The primary advantage of such integration lies in the ability to conduct dynamic validation and correction of data during the clustering process. This reduces the risk of potential errors and significantly improves the quality of grouping, making the results more accurate and aligned with the actual data structures.

The K-means algorithm has long been used in various scientific and technical fields due to its simplicity and efficiency. However, with the development of technology and increasing data volumes, there arises a need for its modifications and improvements. One such innovation is the integration of a noise algorithm,

which assists in accurately pinpointing critical points in urban environments, making the algorithm even more adaptive [14].

Furthermore, the process of automating the determination of the optimal number of clusters, as well as the initial initialization of cluster centers, can solve some of the traditional problems of K-means. This, in turn, significantly enhances its ability to adapt to specific data features in social networks, allowing for more accurate and informative results. In all spheres of data research, one common problem is the absence of complete information, or missing values in datasets. This can complicate the analysis process and lead to unreliable results. However, with the development of the latest algorithms, such an issue becomes less critical.

Algorithms specialized in clustering data with missing values are used to address this particular challenging situation. Their approaches are based on density-oriented methods, preserving the natural structure of the data. Furthermore, the application of Bayes' theory can ensure simultaneous imputation and clustering, improving the quality of results [15]. Among such algorithms, the methods of CI-clustering and LI-clustering stand out for their ability to efficiently place incomplete points in high-density areas, ensuring more accurate and consistent grouping. Modern data analysis, especially in social networks, often requires the use of advanced data preprocessing and machine learning methods. When it comes to finding similar profiles or understanding user behavior patterns, integrating techniques such as vectorization, dimensionality reduction, and the application of machine learning algorithms can be key to success. As demonstrated in the material [16], such a comprehensive approach can greatly enhance the quality and speed of profile analysis. Additionally, image analysis on social networks like Instagram can provide a unique insight into user behavior. In particular, studying photos posted by tourist organizations and visitors not only allows an understanding of the popularity of certain places but also identifies patterns in user interests and activities [17]. Such a deep analysis can be used to create effective marketing strategies, as well as for cluster analysis aimed at identifying users with similar interests or spatial activity within social networks.

Time series analysis plays a pivotal role in studying data dynamics. Especially in the context of studying user behavior, time series can reveal changing patterns and trends that other methods might miss. One of the leading techniques in this direction is the deep time series clustering method (DTSC), which offers an innovative approach to data grouping based on movement behavior [18].

Instead of traditional methods, DTSC employs modified DCAE architectures specifically adapted for working with time series. This ensures high accuracy and operational capability during the analysis of behavioral patterns. For researchers and marketers focusing on social networks, such a tool can be extremely beneficial for studying user behavior, their preferences, and interaction with content.

In the realm of digital image processing, the automatic detection of categories from unlabeled images has become a pivotal task. Various methods are employed for this purpose, including the "bag of visual words", which aids in extracting the principal visual features of an image. Moreover, the research community is increasingly embracing methods grounded in self-supervised learning and transfer learning [19]. Among these, transfer learning stands out, proving to be exceptionally beneficial for understanding and clustering large image sets based on their visual content. Attempting to analyze the dynamics of events on social media, the Embed2Detect method focuses on a combined approach incorporating word embedding and hierarchical clustering [20]. Instead of merely concentrating on structural data characteristics, this method broadens the analysis capabilities to the semantic understanding of textual information. Such in-depth analysis can be key to identifying user profiles on social networks with similar interests, views, or interactions.

An examination of contemporary research indicates the active application of cluster analysis across various research domains, notably within the context of social networks. Machine learning methods, such as deep learning and transfer learning, demonstrate particular efficacy in detecting intricate data structures and relationships among users. Algorithms designed for clustering specific data types, like time series or unlabeled images, expand the horizons of analysis, accounting for the real challenges a researcher might encounter. The importance of data preprocessing and integration of diverse information sources underscores the necessity for a comprehensive approach to clustering in modern studies.

In summary, a review of publications confirms the potential and adaptability of cluster analysis as a vital tool for studying interactions and structures in large datasets. Concurrently, it emphasizes the need for further research and development to better accommodate the nuances of different application areas and data sources.

Cluster analysis in the context of machine learning falls under the category of "unsupervised learning", where models aim to identify structure in the data without prior annotations or labels. This contrasts with "supervised learning", where models are trained based on known input data and corresponding outputs. Visual representation of data can be beneficial for interpreting clustering outcomes and discerning natural groups within the data (Fig. 2).

Clustering Algorithm:

- Load data.
- Set initial cluster centers according to the selected clustering method.
- Classify each element to the appropriate cluster.
- Update cluster centers considering the position of the assigned elements.
- Repeat steps 3-4 if there are changes in cluster composition.
- Finalize clustering stage.

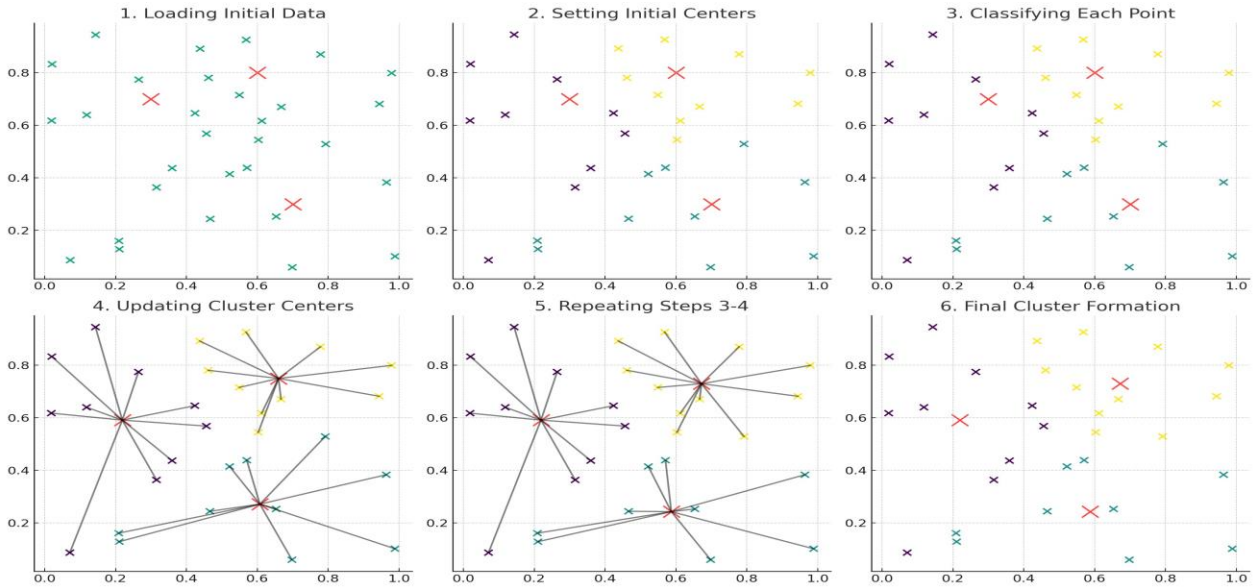


Figure 2: Clustering stages: from determining the initial points to completing the process

The fundamental principle of cluster analysis is to ascertain the degree of similarity between objects. This similarity is often illustrated as a spatial distance between objects, where each coordinate represents a specific characteristic of the object (Fig. 3).

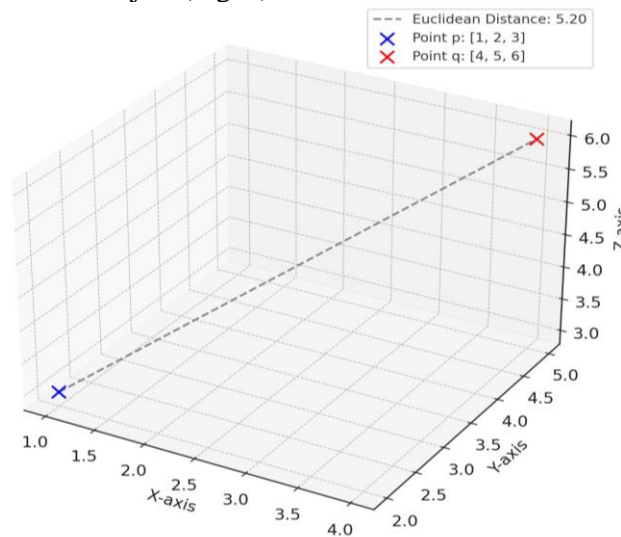


Figure 3: Establishing similarity between two objects using the Euclidean formula

One of the most commonly used metrics for measuring this distance is the Euclidean distance, which represents the "standard" way of determining the distance between two points in space. Its formula is:

$$d(q, p) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}, \quad (1)$$

where q_i and p_i are points in an n -dimensional space.

Contrary to the Euclidean distance, the Mahalanobis distance takes into account correlations between variables, making it ideally suited for datasets with an elliptical distribution. Its formula is:

$$d_M(x_i, x_j) = (x_i - x_j)F^{-1}(x_i - x_j)^T, \quad (2)$$

where F - is the covariance matrix.

By using these distance metrics, we can assess the similarity between objects (Fig. 4), facilitating the clustering process.

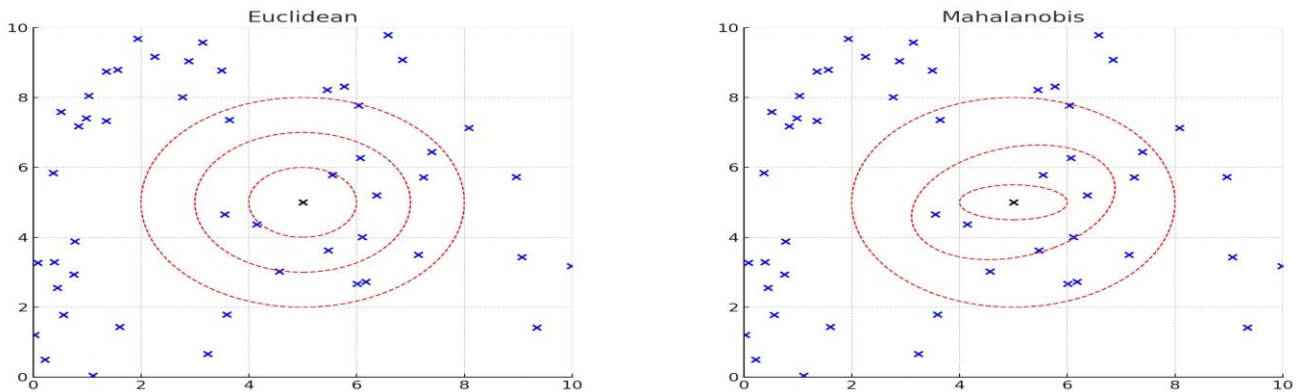


Figure 4: Illustration of point placement on a plane using Mahalanobis and Euclidean methodologies

Effective search for similar profiles in social networks requires the application of highly efficient data analysis methods. One such method is hierarchical clustering. The hierarchical clustering method offers an intriguing approach to data aggregation. It operates based on a structured tree of clusters, where each tree level showcases a unique partitioning of the dataset. This structure allows the researcher to observe different aggregation levels, eliminating the need to predict the exact number of clusters.

The primary approach in hierarchical clustering is agglomerative clustering. According to this method, each object is initially viewed as an isolated cluster. However, with each step, the two most similar clusters merge, until they form one large cluster encompassing all objects. On the other hand, divisive clustering proposes the opposite approach. It starts with one large cluster containing all objects. This cluster is then progressively divided into smaller subclusters until each object becomes an individual cluster.

To visualize the clustering process, dendrograms are often used. This is a graphical representation that displays the sequence of merges and splits of clusters during hierarchical clustering. Alongside hierarchical methods, there are other clustering approaches. For instance, k-means clustering is a popular method focused on partitioning objects into a pre-specified number of clusters. This method is renowned for its efficiency and simplicity, although it has certain limitations. Another method, c-means clustering, is an extension of k-means incorporating fuzzy logic concepts. It allows objects to belong to multiple clusters simultaneously, having varying degrees of membership in each. To determine the optimal number of clusters in k-means, researchers often apply the elbow method. This method is based on analyzing the dependence of within-cluster spread on the number of clusters. In the context of social network analysis, hierarchical clustering may be less suitable for processing large datasets. Whereas methods like k-means or c-means can provide more precise and faster data partitioning. Let's compare the two main clustering methods (Fig. 5): k-means and c-means. Here, one can observe the key features of each method, their similarities, and differences.

Density-based clustering opens new horizons in the realm of data analysis, offering the capability to recognize and isolate regions with varying object concentrations. The core of this method is rooted in the principle of determining groups of objects based on their mutual proximity rather than a pre-defined criterion.

The distinctive feature of this method is its ability to detect areas of high object concentration, distinguishing them from those where the concentration is low. This is achieved using algorithms such as DBSCAN, which can identify "noise" – objects that don't belong to any of the primary clusters.

Additionally, density-based clustering doesn't necessitate the pre-specification of the number of clusters. Owing to this, it demonstrates flexibility when working with data of diverse shapes and structures. However, the method has its limitations, especially when dealing with data where there is a significant disparity in densities between clusters. The density-based method often proves effective for exploring complex datasets or detecting anomalies that other methods might overlook. However, for standard datasets where a certain cluster structure is anticipated, methods like k-means might be more appropriate. When investigating social networks, developers frequently turn to these platforms' APIs for data collection. Platforms such as Twitter, Facebook, and LinkedIn provide their APIs, allowing users to fetch user data in formats like JSON or CSV. Nevertheless, each API comes with its characteristics, restrictions, and authorization requirements. Analyzing such data necessitates additional processing, especially given their volume and intricacy. After retrieving the data via the API, it's crucial to normalize it before applying clustering methods. Ultimately,

appropriately chosen techniques will allow for efficient analysis of user profiles and interactions on social networks. This, in turn, can unveil new opportunities for business and research.

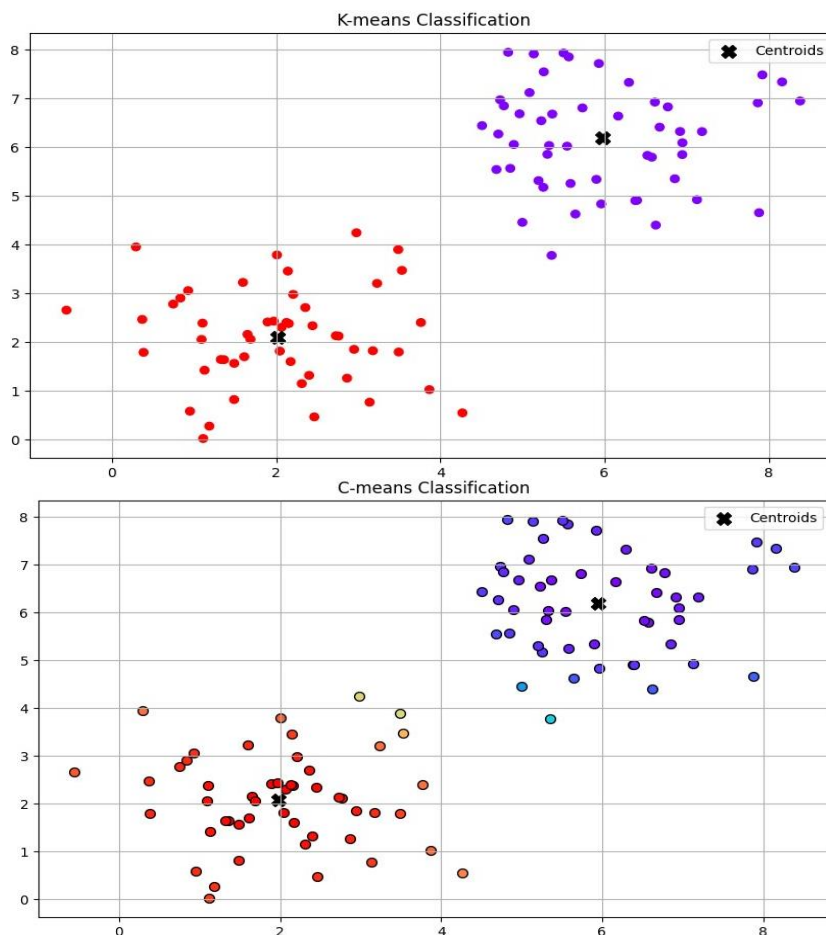


Figure 5: Comparison between k-means and c-means

Normalization of textual data in social networks is an integral part of information processing. Thanks to this, users and analysts can easily and effectively draw conclusions based on data analysis, filtering out irrelevant information and noise. Ensuring normalization requires developers to utilize several key methods. For instance, the segmentation process enables breaking down large text blocks into smaller units like words or sentences for further analysis. Other methods, such as TF-IDF, can pinpoint keywords that are most relevant to a particular context. In contrast, NLTK allows for an in-depth textual analysis using various linguistic tools.

Additionally, when designing software for analyzing social networks, several key aspects must be considered. This includes a modular architecture that ensures system flexibility, as well as accommodating asynchronous requests, allowing for smooth system operation regardless of the volume of data being processed. When developing the user interface, it's crucial to focus on usability and intuitiveness. This will enable users to effortlessly make queries, view results, and interact with the system. Using tools like Figma can simplify the design process, allowing the team to easily create and test interface mockups.

In a broad sense, an effective system for analyzing data from social networks combines in-depth textual data analysis, a flexible modular software architecture, and an intuitive user interface. Such an approach ensures high productivity and meets user needs.

5. Discussion of experimental results

The development of a method for identifying similar accounts on social media platforms is an intriguing and timely challenge. During the research, it was determined that many existing clustering algorithms don't always perfectly suit the analysis of large data sets, especially when dealing with social media data. This led to the idea of creating a specific method for this purpose.

Initially, the k-means algorithm was chosen as the foundation due to its universality and suitability for user profile analysis. However, considering that data volumes can be vast, a modification—mini batch k-

means—was proposed. This algorithm differs from the standard version in that it makes a randomized selection of a data subset at each iteration step, allowing for significantly faster performance.

When determining the cluster centers, the primary objective is to minimize the distance between objects and their corresponding centers. Let's examine the data set

$$T = \{x_1, x_2, \dots, x_p\}, x_i \in R^{m \times n} \quad (3),$$

where each element x_i represents an object in the form of an n-dimensional vector, and where m - is the total number of objects in the set T, The determination of the cluster centroids C (with a predetermined number of clusters) aiming to minimize the distance to the data T is carried out based on the following formula:

$$\min \sum_{x \in T} ||f(C, x) - x||^2, \quad (4)$$

where $f(C, x)$ indicates the cluster centroid closest to the object.

A crucial step in this method is the preprocessing of data. Clustering algorithms, especially k-means, are often sensitive to noise. Therefore, prior to clustering, it's essential to normalize the data, which will depend on the specific characteristics of the data in question. The methodology for searching for similar accounts begins with gathering user data from the API, normalizing this data, constructing a vector matrix, determining the optimal number of clusters, and then applying the mini batch k-means for clustering. As a result, the user receives a list of similar accounts. This method aims to provide an efficient search for similar accounts on social media platforms, taking into account the peculiarities of the data and clustering algorithms. A block diagram (Fig.6) of the methodology will help visualize each step of this process.

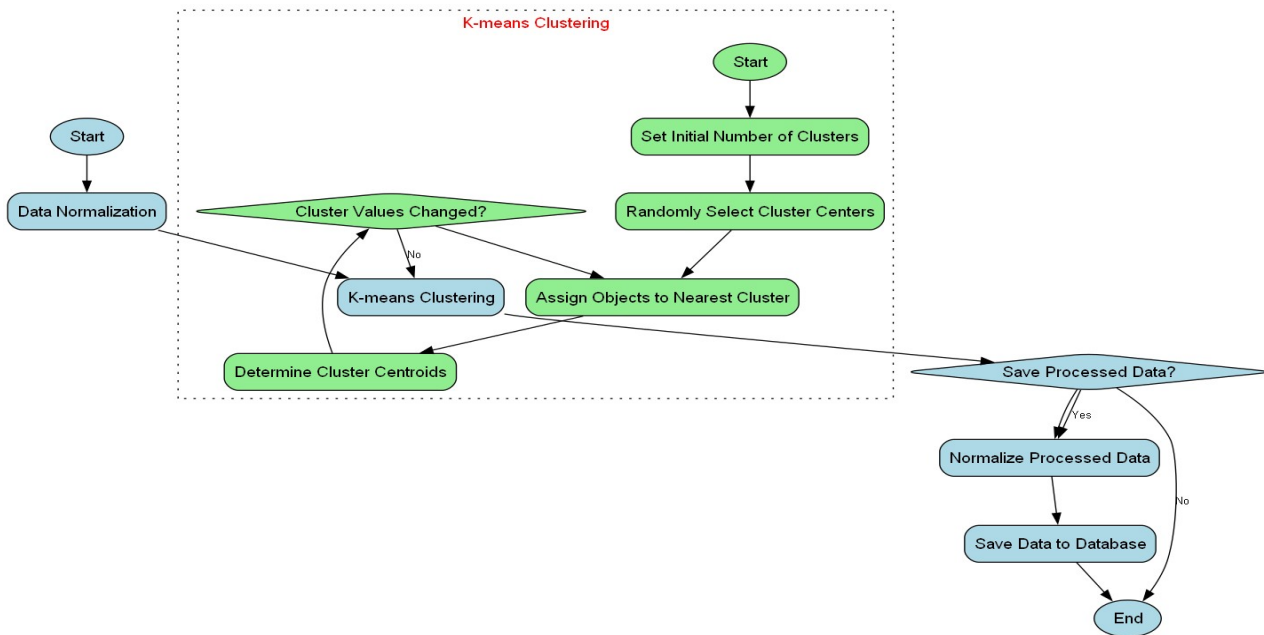


Figure 6: Block Diagram of the Methodology for Identifying Similar Profiles in Social Networks through Cluster Analysis

The algorithm's solution is a tool for analyzing user accounts on social networks, designed to detect similarities among them based on specified characteristics. Main functions of the software include:

- **Data Collection and Preparation:** The software ensures the generation of a test dataset that mimics real user profiles. In real-world conditions, such data can be acquired through social network APIs.
- **Data Normalization:** To simplify calculations and enhance the accuracy of analysis, user profile characteristics are normalized to the range [0, 1].
- **Cluster Analysis:** The software employs the k-means algorithm for grouping users based on their characteristics, enabling the detection of similar groups.
- **Cluster Optimization:** Using the elbow method, the software recommends the optimal number of clusters for precise analysis.

Visualization (Fig. 7) illustrates how the software looks during its operation. A straightforward and comprehensible interface enhances user convenience.

The graph (Fig. 8) depicts the results of using the elbow method to determine the optimal number of clusters. For a deeper understanding of the program's structure and operation, a UML diagram (Fig. 9) is

proposed, which reflects the interactions between different components of the system. This diagram allows for easy identification of the main classes, their attributes, methods, and the relationships between them.

```

account_finder.py
base_clustering.py
base_data_handler.py
clustering.py
data_normalization_handler.py
data_preparation.py
data_preparation_handler.py
data_processing.py
kmeans_clustering_handler.py
main.py
library root
libraries
and Consoles
7 def main():
8     finder = AccountFinder()
9     # Step 1: Getting the initial data
10    finder.retrieve_initial_data()
11    # Step 2: Normalize the data
12    finder.normalize_data()
13    # Step 4: Determining the number of
14    finder.determine_clusters()
15    # Step 5: Performing clustering
16    finder.apply_clustering()
17    if __name__ == "__main__":
18        main()
19
Initial Data:
Age  Number of friends  Number of Posts  The number of groups
49   204                 87              7
51   178                 23              15
20   110                 17              7
58   303                 61              9
51   173                 75              13

Normalized Data:
Age  Number of friends  Number of Posts  The number of groups
0.772727  0.408818  0.878788  0.333333
0.818182  0.356713  0.232323  0.777778
0.113636  0.220441  0.171717  0.333333
0.977273  0.607214  0.616162  0.444444
0.818182  0.346693  0.757576  0.666667

```

Figure 7: Testing process of the algorithm for detecting similar profiles in social networks through cluster analysis

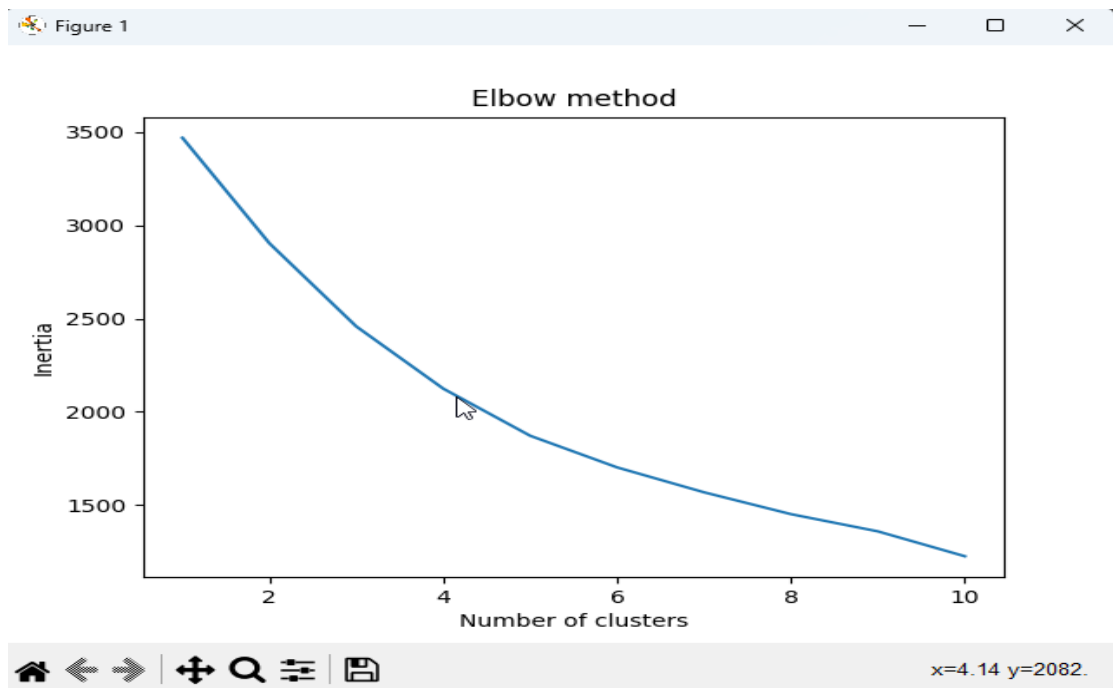


Figure 8: Results of using the elbow method

6. Conclusions

As a result of the conducted research, an analysis of user data from leading social networks, Twitter and LinkedIn, was carried out. Key characteristics of user behavior and their interactions within the networks were identified. An algorithmic methodology for detecting similar profiles in social networks through cluster analysis was proposed. Applying this algorithm will effectively group profiles based on their characteristics, considering not only the content of the publications but also metadata.

A UML diagram was developed to represent the practical implementation of the methodology, reflecting the key components and processes of the system. Specifically, a software model in Python was developed that facilitated the processing of large data sets. A series of tests confirmed the high efficiency of the proposed method, especially after detailed preparation and normalization of data.

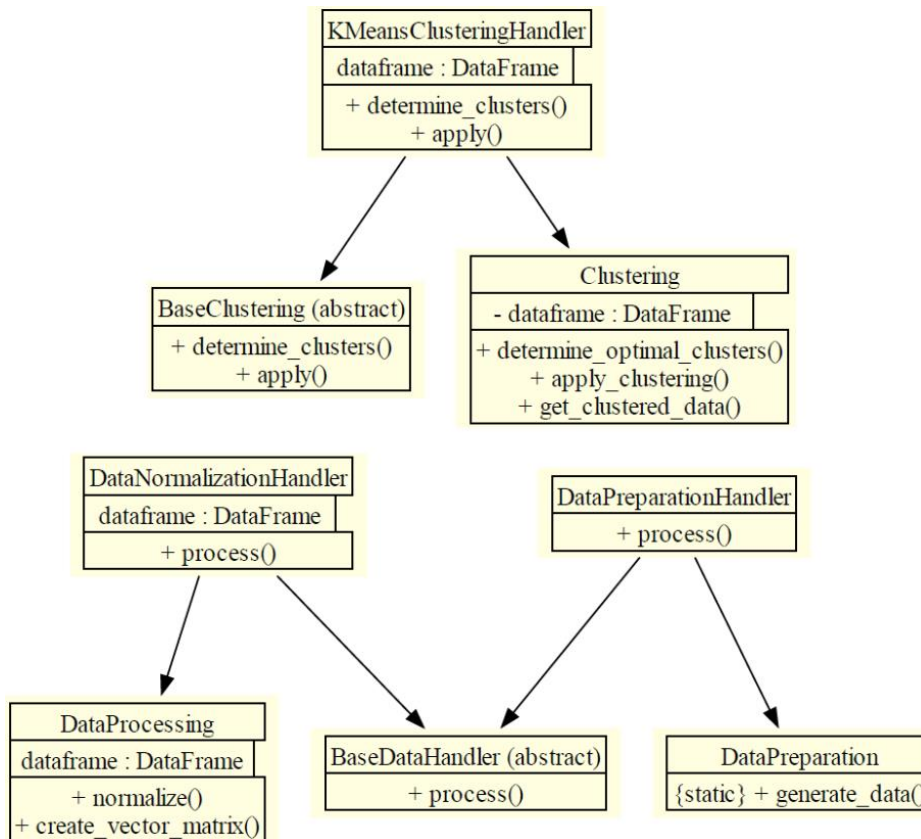


Figure 9: UML diagram of the program model construction for analysis

Thus, thanks to the conducted research, not only was user behavior in social networks analyzed, but also a toolkit for their classification was proposed. This unveils new opportunities for understanding the dynamics of social interactions and can be utilized in a range of practical developments for applications in areas such as targeted advertising, social trend analysis, or strategies for user engagement.

The proposed software is an integrated solution for analyzing accounts on social networks. It serves as a tool for social networks, marketers, and analysts who are interested in studying user behavior or identifying similar user groups. Thanks to its modular architecture and the use of modern machine learning methods, the program can become an essential part of analytical research in social networks.

7. References

- [1] Isfandyari-Moghaddam, A., Saberi, M. K., & Naderbeigi, F. (2021). Global scientific collaboration: A social network analysis and data mining of the co-authorship networks. CLIP. doi: <https://doi.org/10.1177/01655515211040655>
- [2] Yao, Q., Li, R. Y. M., Song, L., & Crabbe, M. J. C. (2021). Construction safety knowledge sharing on Twitter: A social network analysis. Safety Science. doi: <https://doi.org/10.1016/j.ssci.2021.105411>

- [3] Raschka, S., Liu, Y. (Hayden), & Vahid. (2022). Machine Learning with PyTorch and Scikit-Learn. Packt. doi: <https://www.packtpub.com/product/machine-learning-with-pytorch-and-scikit-learn/9781801819312>
- [4] Hicks, S. C., Liu, R., Ni, Y., Purdom, E., & Risso, D. (2021). Fast clustering for single cell data using mini-batch k-means. *plos.org*. doi: <https://doi.org/10.1371/journal.pcbi.1008625>
- [5] Škrjanc, I., Andonovski, G., Iglesias, J. A., Sesmero, M. P., & Sanchis, A. (2022). Evolving Gaussian on-line clustering in social network analysis. *Expert Systems with Applications*. doi: <https://doi.org/10.1016/j.eswa.2022.117881>
- [6] Widiyaningtyas, T., Hidayah, I., & Adji, T. B. (2021). User profile correlation-based similarity (UPCSim) algorithm in movie recommendation system. *Journal of Big Data*. doi: <https://doi.org/10.1186/s40537-021-00425-x>
- [7] Hayawi, K., Mathew, S., Venugopal, N., Masud, M. M., & Ho, P.-H. (2022). DeeProBot: a hybrid deep neural network model for social bot detection based on user profile data. *Social Network Analysis and Mining*. doi: <https://doi.org/10.1007/s13278-022-00869-w>
- [8] Ebrahimi, P., Basirat, M., Yousefi, A., Nekmahmud, M., Gholampour, A., & Fekete-Farkas, M. (2022). Social Networks Marketing and Consumer Purchase Behavior: The Combination of SEM and Unsupervised Machine Learning Approaches. *Big Data Cogn.* doi: <https://doi.org/10.3390/bdcc6020035>
- [9] Fkih, F. (2022). Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison. *Journal of King Saud University - Computer and Information Sciences*. doi: <https://doi.org/10.1016/j.jksuci.2021.09.014>
- [10] Liu, P., Zhang, K., Wang, P., & Wang, F. (2022). A clustering- and maximum consensus-based model for social network large-scale group decision making with linguistic distribution. *Information Sciences*. doi: <https://doi.org/10.1016/j.ins.2022.04.038>
- [11] Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics*. doi: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04675-1>
- [12] Deng, S. (2020). Clustering with Fuzzy C-means and Common Challenges. *Journal of Physics: Conference Series*. doi: <https://doi.org/10.1088/1742-6596/1453/1/012137>
- [13] Borlea, I.-D., Precup, R.-E., & Borlea, A.-B. (2022). Improvement of K-means Cluster Quality by Post Processing Resulted Clusters. *Procedia Computer Science*. doi: <https://doi.org/10.1016/j.procs.2022.01.009>
- [14] Ran, X., Zhou, X., Lei, M., Tepsan, W., & Deng, W. (2021). A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots. *Soft Computing Application to Engineering Design*. doi: <https://doi.org/10.3390/app112311202>
- [15] Xue, Z., & Wang, H. (2021). Effective density-based clustering algorithms for incomplete data. *IEEE*. doi: <https://doi.org/10.26599/BDMA.2021.9020001>
- [16] Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., de Medeiros, D. S. V., & Mattos, D. M. F. (2021). Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. *Decentralization and New Technologies for Social Media*. doi: <https://doi.org/10.3390/info12010038>
- [17] Paül i Agustí, D. (2021). The clustering of city images on Instagram: A comparison between projected and perceived images. *Journal of Destination Marketing & Management*. doi: <https://doi.org/10.1016/j.jdmm.2021.100608>
- [18] Alqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep Time-Series Clustering: A Review. *Computer Science & Engineering*. doi: <https://doi.org/10.3390/electronics10233001>
- [19] Zhang, H., & Peng, Y. (2022). Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research. *Sociological Methods & Research*. doi: <https://doi.org/10.1177/00491241221082603>
- [20] Hettiarachchi, H., Adedoyin-Olowe, M., Bhogal, J., & Gaber, M. M. (2021). Embed2Detect: temporally clustered embedded words for event detection in social media. *SpringerLink*. doi: <https://doi.org/10.1007/s10994-021-05988-7>