https://doi.org/10.15407/scine17.03.078

SYDOROVA, M. (https://orcid.org/0000-0001-7795-0459),
BAYBUZ, O. (https://orcid.org/0000-0001-7489-6952),
VERBA, O. (https://orcid.org/0000-0003-1030-4377),
and **PIDHORNYI, P.** (https://orcid.org/0000-0002-6005-9739)
Oles Honchar Dnipro National University,
72, Gagarin Ave., Dnipro, 49010, Ukraine,
+380 56 744 7683, mzeom@ukr.net

INFORMATION TECHNOLOGY FOR TRAJECTORY DATA MINING

Introduction. *Advanced technologies allow almost continuous tracking and recording the movement of objects in space and time. Detecting interesting patterns in these data, popular routes, habits, and anomalies in object motion and understanding mobility behaviors are actual tasks in different application areas such as marketing, urban planning, transportation, biology, ecology, etc.*

Problem Statement. *In order to obtain useful information from trajectories of moving objects, it is important to develop and to improve mathematical methods of spatiotemporal analysis and to implement them in high-quality modern software.*

Purpose. *The purpose of this research is the development of information technology for trajectory data mining.*

Materials and Methods. *Information technology contains the three main algorithms: revealing key points and sequences of interest with the use of density-based trajectories clustering of studied objects; detecting patterns of an object movement based on association rules and hierarchical cluster analysis of its motion trajectories in the time interval of observations, similarity measure of the motion trajectories has been proposed to be calculated on the basis of the DTW method with the use of the modified Haversine formula; new algorithm for revealing permanent routes and detecting groups of similar objects has been developed on the basis of clustering ensembles of all studied trajectories in time. The clustering parameters are selected with multi-criteria quality evaluation.*

Results. *The modern software that implements the proposed algorithms and provides a convenient interaction with users and a variety of visualization tools has been created. The developed algorithms and software have been tested in detail on the artificial trajectories of moving objects and applied to analysis of real open databases.*

Conclusions. *The experiments have confirmed the efficiency of the proposed information technology that may have a practicable application to trajectory data mining in various fields.*

Keywords: information technology, pattern mining, trajectory of motion, points and sequences of interest, cluster analysis, and similarity measure.

In today's world, where everything is in constant motion, mobility is a key concept. Prevalence of GPS-enabled devices and wireless communication technology leads to the accumulation of huge amounts of information about the movement of objects in space and

Citation: Sydorova, M., Baybuz, O., Verba, O., and Pidhornyi, P. Information Technology for Trajectory Data Mining. *Sci. innov.* 2021. V. 17, no. 3. P. 78—86. https://doi.org/10.15407/scine17.03.078

time. Having analyzed these data, we may better understand the features of the motion and behavior of objects, as well as other interesting patterns.

One example of practical application is the analytics or marketing industry, where it is important to correctly analyze the audience and categorize it fairly precisely by interest category in order to be able to more accurately engage a new target audience or draw conclusions based on the interests of an existing audience. Consider, for example, the trajectory of human movement. We can determine which areas may be important for him: work, gyms, shopping malls, etc. On the basis of several trajectories, it is even possible to track some habits and patterns of human behavior: go to the cafe every morning, an evening jog around the park every Friday, and so on.

Most mobile apps installed on mobile devices occasionally request access to a user's location. The location and movement of the users is used in operating systems and mobile applications to provide various quality promotional recommendations for nearby places of interest or, for instance, timely notification of traffic jams.

Analyzing the trajectories obtained by tracking the movement of animals, it is interesting to determine which geographic areas are important for an animal, in which it spends a certain part of its time or movement patterns such as spatio-temporal expression of behaviours, e.g. in flocking sheep or birds assembling for the seasonal migration, migration patterns of traveling for better access to food, water, and shelter.

In transportation field — tracking vehicle movement, traffic planning, detecting hotspots, for taxi pick-up point recommendation; in ecology — tracking down pollution incidents etc.

There are also some successful examples in various AR applications: such as Ingress or PokemonGO, where user location analysis is the main mechanism for managing the gameplay.

Raw trajectory data in the form of geographical coordinates and timestamps are meaningless to humans. All these data are of great value only if it is properly processed. Therefore, the task of trajectory data mining is of interest among researchers and very popular in practical applications.

Trajectory data mining is the process of interesting, useful knowledge and patterns discovery in large data sets of objects` motion history. The mathematical apparatus of spatial-temporal analysis methods is used to study the trajectories of motion [1—3]. In [4] the authors divided the methods into two categories: primary (clustering, classification) and secondary (pattern mining, outlier detection, prediction) and showed relationships between them.

A systematic review of different methods and examples of practical application of trajectory analysis was performed in [4—6]. In [6] links to different datasets can also be found.

An important task is finding points of interest or points of stops [7—8]. Generalizing the various methods, the fact can be noticed, that in most cases density algorithms (DBSCAN and its modifications) are used for solving this problem. In cases where there is no access to the entire amount of data, for example, in real-time analysis, the algorithms based on analysis of the spatial and temporal differences between individual segments of trajectories are used in order to highlight the centroids as points of interest. Due to their low computational complexity, they are widely used in mobile devices. Sometimes a probabilistic approach is used based on models such as a mixture of Gaussian distributions, Bayesian and Markov models.

In many pattern mining tasks, it is necessary to determine the similarity of trajectories. Most approaches to determining it are distance measures that can be divided into two categories: spatial ones that focus only on spatial changes and ignore the temporal attribute, and spatio-temporal ones that use data of both spatial and temporal changes. Spatio-temporal distance measures, for the most part, have the same camputation principles as when dealing with time series. Widely used approaches to distance measure are, for example, DTW, LCSS, TWED and so on. Among the spatial measures of distance, three types are most com-

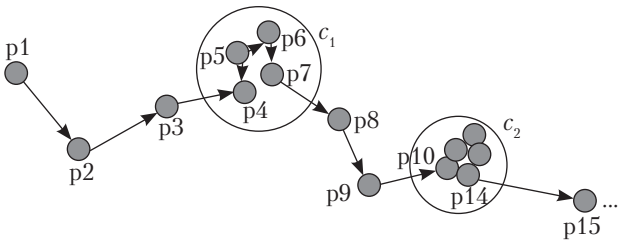

Fig. 1. Determination of points clusters on the trajectory

mon: proximity to the direction of motion, proximity of geometric shapes and spatial proximity.

Visualization and visual analytics is an important and effective tool for the study of trajectories that has been addressed, for example, in [9—10].

Although there are a lot of theoretical works in this area, there are few software solutions and most of them are for highly specialized tasks.

The object of study is trajectories of objects' motion in space and time.

The subject of study is methods of revealing useful information and pattern mining in trajectory databases.

The purpose of the work is to study existing approaches for solving the problem, to develop the information technology for trajectory data mining and to apply the created software to the analysis of real datasets from different subject areas.

Materials and Methods

Let a set of objects of observation be given as $O = \{O_k; k = \overline{1, N_{obj}}\}$, each object is characterized by a set of trajectories $O_k = \{T_i^k; i = \overline{1, N^k}\}$, $T_i^k = \{p_{ij}^k; j = \overline{1, N^k}\}$ — i-th trajectory of the k-th object of observation, $p_{ij}^k = \langle t_{ij}^k, x_{ij}^k, y_{ij}^k \rangle$, where x_{ij}^k, y_{ij}^k — latitude and longitude of the j-th point in the i-th trajectory of the k-th object of observation, t_{ij}^k — the moment of time corresponding to the j-th point in the i-th trajectory of the k-th object.

To analyze this data, it is necessary to develop algorithms and software for:

- searching for key points and sequences of interest;
- detecting objects movement patterns in the period of observation;
- revealing permanent routes, identifying groups of similar objects based on an analysis of all studied trajectories over time.

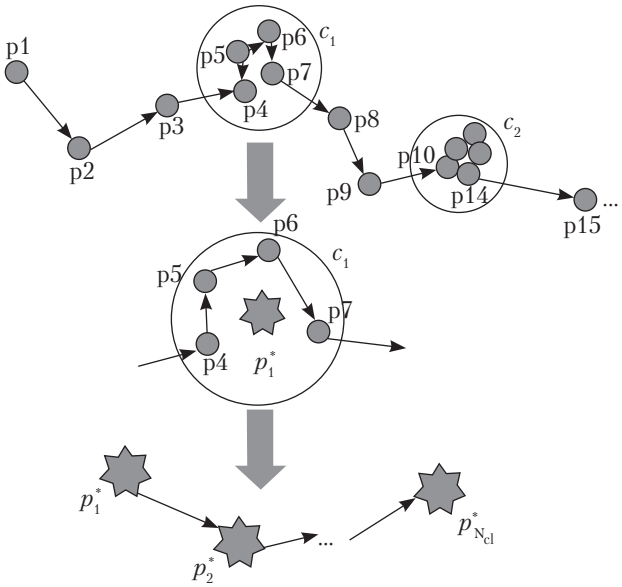

Fig. 2. Transition from the initial trajectory to sequence of key points (interests)

The first step in analyzing trajectories is to identify key points or so-called points of interest. It can be interesting to researchers as an independent task (identifying visited by an object places that are, places of particular interest to him) and may be used as an ancillary task to reduce data dimensions and to eliminate noise.

To determine the points of interest, we first apply the DBSCAN dense clustering algorithm, which will allow us to find clusters of points on the trajectory (Fig. 1):

1. Set the parameters of the algorithm ε (allowable neighborhood of a point during cluster formation) and *minP* (the minimum number of points for cluster formation). All points of the studied trajectory are considered as a set of non-clustered points *S*.

2. Of the set *S* choose an arbitrary point *r* and calculate distances d_{ri}, $i \in S$. As geo-position coordinates are analyzed, we propose using a modified Haversine formula as a measure of distance:

$$d_{ij} = \arctan\left(\frac{\ldots \sqrt{[\cos x_j \sin\Delta y]^2 + [\cos x_i \sin x_j - \sin x_i \cos x_j \cos\Delta y]^2}}{\sin x_i \sin x_j + \cos x_i \cos x_j \cos\Delta y}\right)$$

3. Identify *countP* — the number of points contained in ε-neighborhood of point *r*, that is, such that $d_{ri} < \varepsilon$,

- if *countP* < *minP*, then denote the point *r* by noise and proceed to item 2,
- otherwise, from the point *r* and its ε-neighborhood form a cluster and proceed to item 4.

4. For each point belonging to the cluster find points belonging to its ε-neighborhood and add them to the cluster.

5. All points of the formed cluster exclude from the set *S*.

6. If there are non-noise objects in the set *S*, proceed to item 2, otherwise finish the clustering.

For each cluster c_l, $l = \overline{1, N_{cl}}$ find a point of interest $p_l^* = \langle t_l^*, x_l^*, y_l^* \rangle$ as the averaging of the values of all points belonging to it:

$$t_l^* = \frac{1}{|c_l|}\sum_{p \in c_l} t, \quad x_l^* = \frac{1}{|c_l|}\sum_{p \in c_l} x, \quad y_l^* = \frac{1}{|c_l|}\sum_{p \in c_l} y.$$

Based on the implemented algorithm for accessing Google Places for more information about each point received, we move from the point of interest to the semantic location.

Having identified all the points of interest of the trajectory, we obtain a sequence of key points (points of interests) $P^* = \{p_l^*, l = \overline{1, N_{cl}}\}$, $p_l^* = \langle t_l^*, x_l^*, y_l^* \rangle$ (Fig. 2).

The following algorithm is proposed to detect patterns of object movement within the observation time interval.

1. Based on all the trajectories of the object being moved, identify the sequence of key points (interests) as described above.

2. Determine a hierarchical cluster structure of sequences of interest. By finding groups of similar trajectories, it is possible to draw conclusions about the permanent routes, preferences and habits of the studied object.

3. Using the apriori algorithm, we find associative rules that allow detecting the regularities between related events in the form of "if the object visited point A, then with probability *p*, it will also visit point B". As related events we consider the points of interest in one trajectory. This approach is described in detail in the previous work of the authors [11].

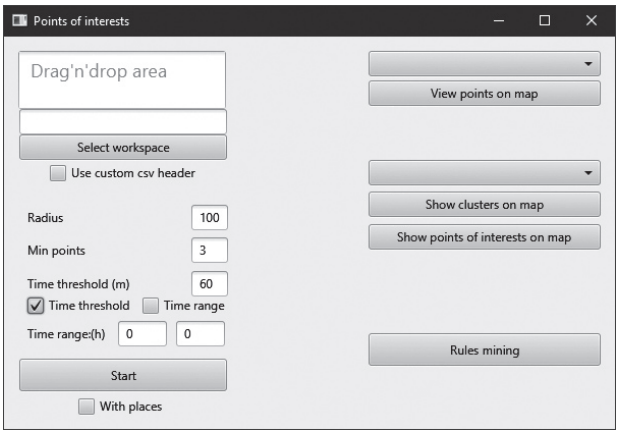

Fig. 3. The main form

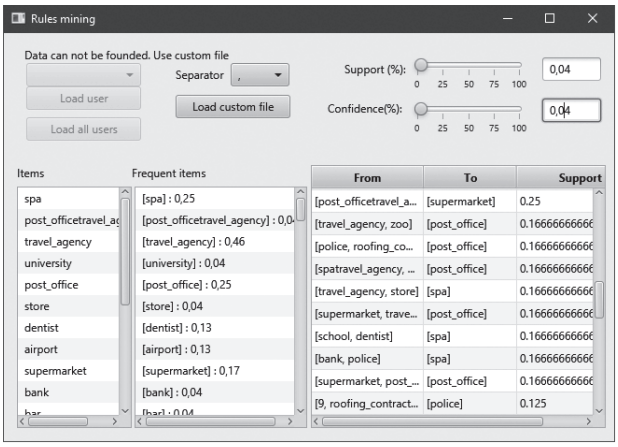

Fig. 4. Form of searching for associative rules

The algorithm for agglomerative hierarchical clustering of trajectories:

1. Consider each trajectory of interest $P_i^*, i = \overline{1, N^k}$ to be a separate cluster g_i, $|g_i| = 1$. Calculate the distance matrix $D = \{d_{ij}\}$, $i, j = \overline{1, N^k}$, where d_{ij} — distance between trajectories P_i^* and P_j^*, which in this work is calculated by an algorithm DTW, but it is suggested to use the above modified Haversine formula instead of the basic Euclidean metric.

2. In the distance matrix *D* find the minimum element d_{ij} and clusters g_i and g_j unite $g_{i+j} = g_i \cup g_j$, $|g_{i+j}| = |g_i| + |g_j|$.

3. From the matrix *D* remove the distances from g_i and g_j to other clusters and add distances corresponding to the new cluster g_{i+j}.

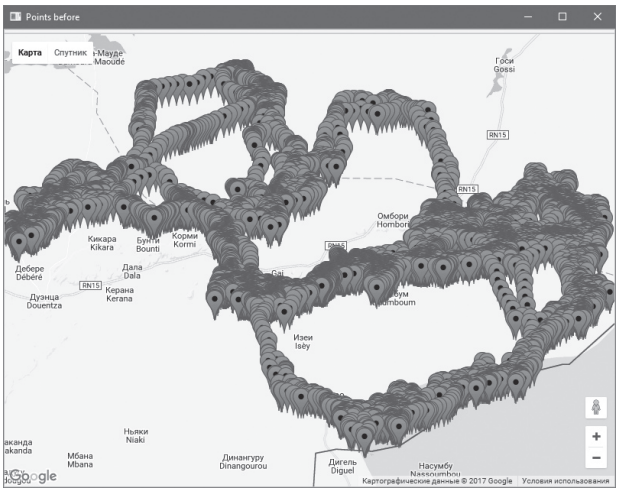

Fig. 5. Visualization of trajectories

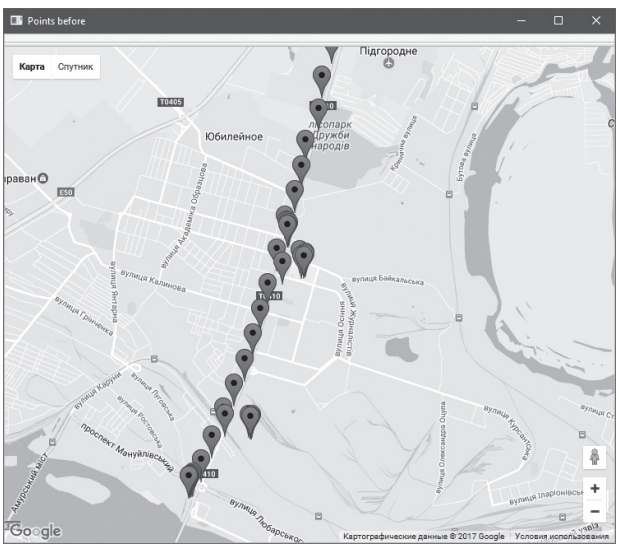

Fig. 6. Visualization of the selected trajectory

To calculate the distance between clusters, there is a general Lance-Williams formula: $d(g_{i+j}, g_h) = \alpha_1 d(g_i, g_h) + \alpha_1 d(g_j, g_h) + \beta d(g_j, g_j) + \gamma d(g_i, g_h) - d(g_i, g_h)$. By setting different parameter values α_1, β_2, β, γ, we will get different types of agglomerative hierarchical methods: single-link, complete-link, average-link and more.

4. Repeat steps 2—3 until we have the required number of clusters or all the objects are combined into one cluster to construct the dendrogram. The optimal number of clusters and other settings has been determined with the use of the multicriteria quality assessment technology proposed by the authors in [12].

The next step in proposed in this work technology is to identify groups of similar objects based on an analysis of all the trajectories studied over time, as well as finding permanent routes.

Since this problem is more complex than the previous one and the simple application of cluster analysis methods is not possible, an ensemble clustering approach is proposed to solve it [13].

1. Form M subsets of data $U_m = \{P_{mi}^*, i = \overline{1, N_{obj}}\}$, $p = \overline{1, M}$, representing the trajectories of interest of all the objects under study in a given day m (M — the number of days in the period under review).

2. Applying clustering separately to each of the subsets U_m, obtain M partitions $G_m = \{g_1^m, g_2^m, ..., g_{K_m}^m\}$, $m = \overline{1, M}$, $g_i^m = \{P_j^{*m}, j = 1, |g_i^m|\}$ — the i-th cluster in the m-th partition, $i = 1, K_m$, $\sum_{i=1}^{K_m} |g_i^m| = N_{obj}$, $\bigcup_{i=1}^{K_m} g_i^m = U_m$, $g_i^m \cap g_j^m = \varnothing$, $i, j = 1, K_m$.

3. Based on the obtained results, construct adjacency matrices $A_m = \{a_{ij}^m, i, j = \overline{1, N_{obj}}\}$, $m = \overline{1, M}$, where

$$a_{ij}^m = \begin{cases} 1, \textit{ if } (P_i^{*m} \in g_k^m) \wedge (P_j^{*m} \in g_k^m), k \in [1, K_m] \\ 0, \textit{ otherwise} \end{cases},$$

that is $a_{ij}^m = 1$, if the trajectories of objects i and j belong to the same cluster, $a_{ij}^m = 0$ otherwise.

Since the number of clusters K_m in each of m may be different, the degree of similarity of the objects in this case will not be the same. It is clear that the more clusters in a partition, the more similar are the objects that fall into one cluster. So to account for this difference, multiply each of the matrices A_m by the number of clusters in m-th partition. So get that

$$a_{ij}^m = \begin{cases} K_m, \textit{ if } (P_i^{*m} \in g_k^m) \wedge (P_j^{*m} \in g_k^m), k \in [1, K_m] \\ 0, \textit{ otherwise} \end{cases},$$

4. Construct an aggregate matrix $A' = \{a'_{ij}\}$, $i, j = \overline{1, N_{obj}}$, as follows $a'_{ij} = \sum_{m=1}^{M} / K'$, where $K' = \sum_{m=1}^{M} K_m$.

The greater the value a'_{ij}, the more similar the objects i and j are in their trajectories of interest.

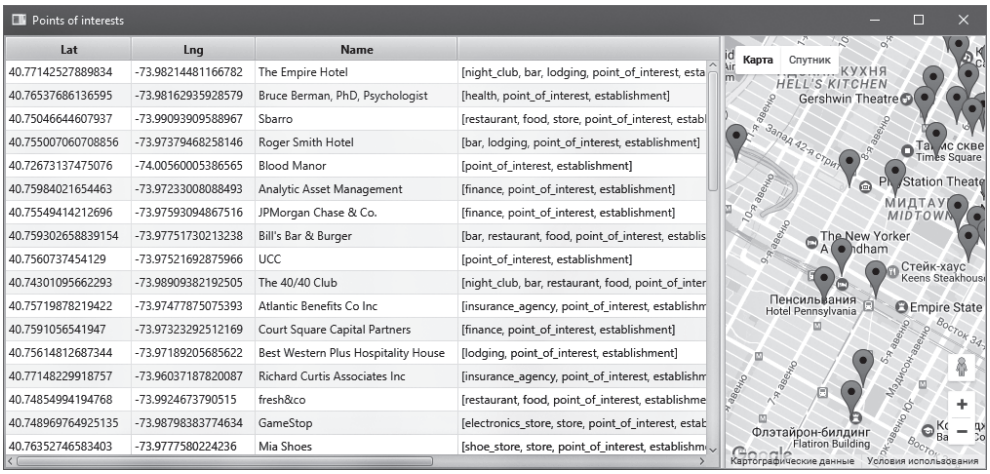

Fig. 7. Visualization of points of interest

Fig. 8. Visualization of the received clusters

Value of a'_{ij} can be considered a measure of closeness, and a matrix A' — a matrix of similarity. Since assigning objects to one cluster on a particular day indicates their closeness across trajectories of interest, and the frequency of their union indicates similarity over time, the measure of similarity thus introduced really reflects the degree of similarity of the two objects along trajectories of interest with regard to time changes.

5. The transition from the similarity matrix $A' = \{a'_{ij}\}, i, j = \overline{1, N_{obj}}$ to the distance matrix $D = \{d_{ij}\}, i, j = \overline{1, N_{obj}}$ can be done as follows: $d_{ij} = 1 - a'_{ij}, i, j = \overline{1, N_{obj}}$. That is, the more similar objects i and j in the matrix A' the smaller distance between them in the matrix D.

6. Next, we need to get the final solution of the task, namely the division of the objects of the original set $O = \{O_k, k = \overline{1, N_{obj}}\}$ into clusters. Clustered in one cluster should be those objects that are similar in all trajectories with respect to their temporal changes. Summary partitioning may be gotten by using cluster analysis algorithms that deal with matrix of distances between objects (such as hierarchical, graphical, or fuzzy methods) as input. Hierarchical clustering is used in the work.

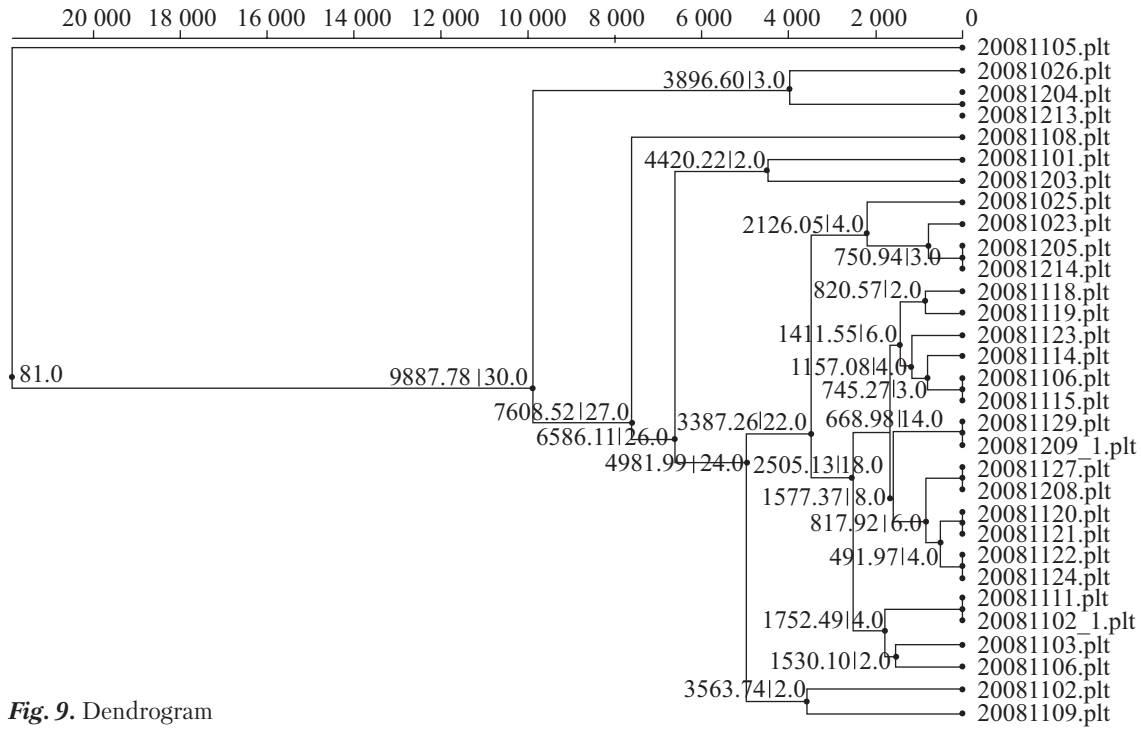

Fig. 9. Dendrogram

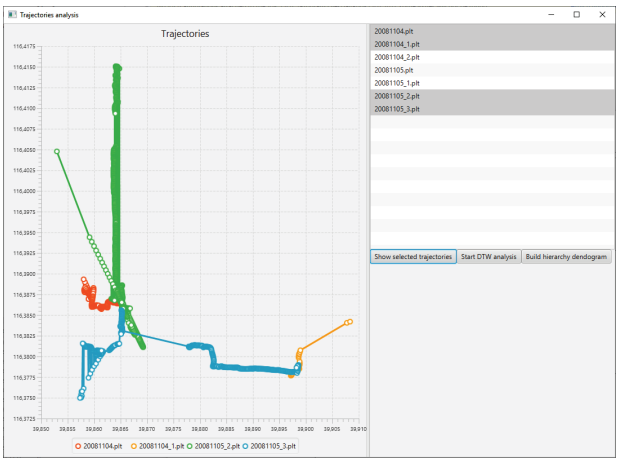

Fig. 10. Trajectory visualization

The result of the work is software that implements the above algorithms for analyzing the trajectories of objects and identifying useful patterns in them. The software is written in Java with the use of JavaFX graphical framework.

8 graphical forms have been created for user interaction with the program and easy navigation between them is provided. The main form allows you to download the data and set the parameters of the algorithms (Fig. 3). The form of searching for associative rules is shown in Fig. 4.

The initial data can be viewed on the map by all (Fig. 5) and only the selected (Fig. 6).

Once the points of interest have been identified, they can be viewed on the map, and with the implementation of the algorithm of contacting Google Places service to obtain detailed information about each point (place name, its type, etc) semantic locations cab be obtained. An example is shown in Fig. 7. Each of the obtained clusters can be viewed in detail both on the map and in tabular data (Fig. 8).

The results of hierarchical clustering are also presented as a dendrogram (Fig. 9). In order to abstract from the cartographic representation of information, investigated trajectories can also be viewed on a simple graph (Fig. 10).

Developed algorithms and software were applied to the analysis of real data of the open database of the project «Geolife» (Microsoft Research

Asia) [14], which contains information about the movement of volunteers collected through mobile phones or special trackers; the open database «African Elephant Range (2012)», which contains the locations of elephants according to international organizations reports IUCN, SSC, African Elephant Specialist Group (AfESG) [15], the open database «Taxi & Limousine Commission Trip Record Data», which contains information about boarding and disembarking passengers for the mentioned taxi service in New York [16]. In addition, to test the proposed algorithms in more detail, a web-based JavaScript application was developed, which allows you to create artificial trajectories of moving objects in a convenient mode of interaction with the map.

Conclusions

The information technology of trajectory data mining has been developed. It allows searching key points and sequences of interest, semantic locations, permanent routes and patterns of behavior as well as identifying patterns of motion of objects in the period of observation. A new approach has been proposed to identify groups of objects by the similarity of their movement routes and the hierarchical structure of all trajectories studied in time based on ensemble clustering. The developed algorithms have been implemented in a modern software complex that may be applied for the intellectual analysis of trajectories in various subject areas.

REFERENCES

1. Atluri, G., Karpatne, A., Kumar, V. (2018). Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Computing Surveys*, 51(4), 83:1—83:41. doi: 10.1145/3161602
2. Andrienko, N., Andrienko, G. (2006). *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer. Berlin. 703 p.
3. Venkateswara Rao, K., Govardhan, A., Chalapati Rao, K. V. (2012). Spatiotemporal data mining: issues, tasks and applications. *International Journal of Computer Science & Engineering Survey (IJCSES)*, 3(1), 39—52. doi: 10.5121/ijcses.2012.3104
4. Mazimpaka, J. D., Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *Journal of spatial information science,* 13(1), 61—99. doi: 10.5311/JOSIS.2016.13.263
5. Tanuja, V., Govindarajulu, P. (2016). A Survey on Trajectory Data Mining. *International Journal of Computer Science and Security (IJCSS)*, 10(5), 195—214.
6. Zheng, Y. (2015). Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3), 29:1—29:41. doi: http://dx.doi.org/10.1145/2743025
7. Suzuki, J., Suhara, Y., Toda, H., Nishida, K. (2019). Personalized Visited-POI Assignment to Individual Raw GPS Trajectories. *ACM Transactions on Spatial Algorithms and Systems*, 5(3), 16:1—16:31. doi.org/10.1145/3317667
8. Huang, J., Liu, Y., Chen, Y., Jia, Ch. (2019). Dynamic Recommendation of POI Sequence Responding to Historical Trajectory. *International Journal of Geo-Information*, 8(10), 433:1—433:15. doi: 10.3390/ijgi8100433
9. Gonçalves, T., Afonso, A. P., Martins, B. (2015). Cartographic visualization of human trajectory data: overview and analysis. *Journal of Location Based Services*, 9(2), 138—166. doi: 10.1080/17489725.2015.1074736
10. Cai, L., Zhou, Y., Liang, Y., He, J. (2018). Research and Application of GPS Trajectory Data Visualization. *Annals of Data Science*, 5(1), 43—57. doi: 10.1007/s40745-017-0132-1
11. Sidorova, M., Pidhornyi, P. (2018, February). Software for spatio-temporal trajectory analysis and pattern mining. *Proceedings of 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*. Slavske, 958—961. doi: 10.1109/TCSET.2018.8336352
12. Sidorova, M. (2012, February).Information technology of evaluation and improvement the quality of cluster analysis. *Proceedings of International Conference on Modern Problem of Radio Engineering, Telecommunications and Computer Science*. Lviv-Slavske, 390.
13. Baibuz, O. G., Sidorova, M. G. (2014). Information technology of the multivariate time series fuzzy clustering on the example of the samara river hydrochemical monitoring. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu,* 2014(5), 11—18 [in Ukrainian].
14. Geolife GPS trajectory dataset — User Guide. URL: https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/ (Last accessed: 24.03.2020).

15. African Elephant Database. URL: https://www.iucn.org/ssc-groups/mammals/african-elephant-specialist-group/african-elephant-database (Last accessed: 24.03.2020).
16. TLC Trip Record Data. URL: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (Last accessed: 24.03.2020).

Received 10.06.2020
Revised 09.07.2020
Accepted 23.02.2021

М.Г. Сидорова (https://orcid.org/0000-0001-7795-0459),
О.Г. Байбуз (https://orcid.org/0000-0001-7489-6952),
О.В. Верба (https://orcid.org/0000-0003-1030-4377),
П.Є. Підгорний (https://orcid.org/0000-0002-6005-9739)
Дніпровський національний університет імені Олеся Гончара,
просп. Гагаріна, 72, Дніпро, 49010, Україна,
+380 56 744 7683, mzeom@ukr.net

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТРАЄКТОРІЙ РУХУ ОБ'ЄКТІВ

Вступ. Сучасні технічні досягнення дозволяють майже постійно відслідковувати та фіксувати рух об'єктів у просторі та часі. Виявлення цікавих закономірностей у цих даних, популярних маршрутів, звичок та аномалій у переміщенні об'єктів, розуміння мобільної поведінки є актуальними завданнями у різних сферах застосування, таких як маркетинг, містобудування, транспорт, біологія, екологія тощо.

Проблематика. Для отримання корисної інформації з даних траєкторій руху об'єктів важливим є розробка й удосконалення математичних методів просторово-часового аналізу та реалізація їх у вигляді сучасного програмного забезпечення.

Мета. Розробка інформаційної технології інтелектуального аналізу траєкторій руху об'єктів.

Матеріали й методи. Інформаційна технологія містить три основні алгоритми: визначення ключових точок та послідовностей інтересу на основі щільнісної кластеризації траєкторій руху об'єктів дослідження; виявлення закономірностей пересування об'єкта на основі асоціативних правил та ієрархічного кластерного аналізу його траєкторій руху у часовому проміжку спостережень, міру подібності запропоновано обчислювати на основі методу DTW та модифікованої формули гаверсинусів; новий алгоритм пошуку сталих маршрутів та виявлення груп схожих об'єктів за усіма досліджуваними траєкторіями у часі на основі ансамблевої кластеризації. Вибір параметрів кластеризації здійснюється за допомогою багатокритеріальної оцінки якості.

Результати. Створено сучасне програмне забезпечення, що реалізовує запропоновані алгоритми, забезпечує зручну взаємодію з користувачем й різноманітні засоби візуалізації. Розроблені алгоритми та програмне забезпечення детально протестовано на штучних траєкторіях рухомих об'єктів та застосовано до аналізу реальних відкритих баз даних.

Висновки. Експериментально підтверджено ефективність розробленої інформаційної технології, яку може бути впроваджено на практиці для інтелектуального аналізу траєкторій у різних галузях.

Ключові слова: інформаційні технології, виявлення шаблонів, траєкторії руху, точки та послідовності інтересу, кластерний аналіз, міра подібності.