Specificity of information clustering in application to the problem of messages classification in social media

D V Moiseev, L I Lukina and O S Mikhailova
Federal State Autonomous Educational Institution of Higher Education the Sevastopol State University

E-mail: dmitriymoiseev@mail.ru

Abstract. The paper presents research and development materials obtained by a team of authors that co-working at the Department of information technologies and computer systems of Sevastopol state University within the framework of an internal grant "Development of agent-based modeling and big data methods for social media analysis in post-conflict societies (grant №28/06-31)". To perform the task of clustering messages in social media, it is necessary to take into account that not only the power of the clusters themselves, but also the power of many clusters can be variable relative to time intervals $t_1, t_{i+1}, \ldots, t_K$. Next, we will discuss the problem statement and some aspects of decision-making on classifying messages in social media in the framework of dynamic clustering.

1. Introduction
When analyzing big data in the social sciences and humanities in real time, it is advisable to apply the approach of clustering a set of messages.

Given that clusters are groups of homogeneous objects, the task of cluster analysis is to divide their set into $m$ ($m$ – integer) clusters based on the characteristics of messages in social media, so that each message belongs to only one partition group. However, social media posts that belong to the same cluster should be homogeneous (similar), and social media posts that belong to different clusters should be heterogeneous.

Various object clustering algorithms are known, for example:

- Hierarchical and flat. Hierarchical algorithms (also called taxonomy algorithms) build a system of nested partitions instead a single partition of a sample into disjoint clusters. Thus, the output is a cluster tree with the entire sample as its root and the smallest clusters as its leaves. Flat algorithms build a single partition of objects into clusters.
- Clear and fuzzy. Clear (or disjoint) algorithms assign a cluster number to each selection object, i.e. each object belongs to only one cluster. Fuzzy (or overlapping) algorithms assign each object a set of real values that show the degree of the object's relationship to clusters. In other words, each object belongs to each cluster with some probability [1-3].
- Static and dynamic. Static algorithms are implemented on a static set of objects. dynamic algorithms imply the possibility to work within a multidimensional space (deformable classes...
and dynamic objects).

2. Problem statement

Both (1) and (2) can be applied to the problem of clustering messages in social media, but due to the specifics of the problem, it is advisable to apply (3).

Figure 1 shows the scheme of the main types of clustering algorithms proposed by the authors, including for messages in social media.

![Figure 1. Main types of clustering algorithms, including social media posts.](image)

Shown in figure 1 the diagram reflects the classification of clustering algorithms for complex objects used in various tasks, including social media posts. Thus, the allocation of dynamic clustering methods to a separate type of algorithms is due to the requirements of a variety of applied problems that take into account the variability of processes both filling clusters and the dynamics of a power of the set of clusters.

Let’s describe the problem in a generalized form. To do this, we denote the objects that make up the studied sets by the following set: \( I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\} \), where \( i_j \) — objects included in the analyzed sets; \( n \) — total number of objects. The clustering task is to build a set of: \( C = \{c_1, c_2, \ldots, c_k, \ldots, c_g\} \), where \( c_k \) — cluster containing similar objects from a set \( I \):

\[
c_k = \{i_j, i_p \mid i_j \in I; d(i_j, i_p) < \sigma\},
\]

where \( \sigma \) — a value that defines the proximity measure for including objects in the same cluster; \( d(i_j, i_p) \) — a measure of proximity between objects, called distance; nonnegative value \( d(i_j, i_p) \) is called the distance between the elements \( i_j \) and \( i_p \), if the following conditions are met:

1. \( d(i_j, i_p) \geq 0 \), for all \( i_j \) and \( i_p \),
2. \( d(i_j, i_p) = 0 \), only when \( i_j = i_p \),
3. \( d(i_j, i_p) = d(i_p, i_j) \),
4. \( d(i_j, i_p) \leq d(i_j, i_r) + d(i_r, i_p) \) [2].

The problem of constructing a classification and regression function in the simplest form can be
formally described as the problem of selecting a function with a minimum degree of error:

$$
\min_{f \in F} R(f) = \min_{f \in F} \frac{1}{m} \sum_{i=1}^{m} s(y_i, f(x_i)),
$$

(1)

where $F$ – set of all possible functions; $s(y_i, f(x_i))$ – loss function, where $f(x_i)$ – the value of the dependent variable found using the function $f$ for a vector $x_i \in T$, $y_i$ – its exact known value. In this case, the loss function takes non-negative values. For clusters with a number greater than two, each type of clustering error generally introduces its own type of loss so that a matrix of size $\xi \times \xi$ (where $\xi$ – is the number of clusters) is obtained.

There are several possible solutions to this problem: in the two-dimensional case (figure 2), and in the multidimensional case, which will be discussed later. As shown in figure 2 clustering schemes for two and three clusters can be obtained by finding the maximum of the variance ratio between internal $\sigma_v^2$ and external elements $\sigma_w^2$:

$$
A = \frac{\sigma_v^2}{\sigma_w^2} = \frac{n_w \sum_{i=1}^{n_w} (x_{vi} - \mu_v)^2}{n_v \sum_{i=1}^{n_v} (x_{wi} - \mu_w)^2} \rightarrow \max
$$

(2)

The simplest ward’s method for hierarchical clustering is based on this approach.

Figure 2. Social media message clustering scheme for disjoint clusters in the classification feature space: a) for the two-dimensional case, b) for the three-dimensional case and two functions $f(x)$.

For the task of clustering messages in social media, it is necessary to take into account that not only the power of the clusters themselves, but also the power of the set of clusters can be variable relative to time intervals $t_i, t_{i+1}, \ldots, t_K$. At the same time, the same number of clusters for different time periods may correspond to different capacities of these clusters. Next, we will discuss the problem statement and some aspects of decision-making on classifying messages in social media in the framework of dynamic clustering.

The question of classifying messages in social media cannot be resolved a priori, based on the described scheme for analyzing incoming messages and the variable capacity of the set of clusters (to which the classified messages can be assigned). Adaptive configuration of the classification system is expected, which will save time and improve
the efficiency of the system for analyzing messages in social media. Thus, the system for classifying messages in social media is implemented on the basis of a dynamic clustering subsystem and is a hierarchical adaptive scheme with feedback.

Denote \( \xi \) – the number of message clusters in information systems, \( \xi = 1, K \), at time \( t_m \). At time \( t_{m+1} \) the number of clusters can change (incrementally or decrementally). Any change in the number of clusters when analyzing a new social media post represents a new state of the clustering system: \( S_i \).

Thus, the state \( S_i \) is described by a vector that defines the number of clusters and their capacity.

Figure 3 shows a bar chart describing changes in the composition of clusters in the process of analyzing messages in social media as part of identifying the States of the clustering system \( S_1, S_2, S_3 \).

Shown in figure 3 dynamic clustering of social media posts describes the following transitions (state changes): \( S_1(t_{m-1}) \rightarrow S_2(t_m) \rightarrow S_3(t_{m+1}) \), this corresponds to three situations: cluster stability, an increase in the number of elements in the cluster, and a decrease in the number of elements in the cluster.

![Figure 3. Scheme of dynamic clustering of social media messages for the two-dimensional case in the clustering parameter space.](image)

The problem of dynamic clustering of messages in social media in general should be considered as a three-dimensional process:

\[
Q(\xi, \varphi, t_m),
\]

where \( \xi \) – number of classes, \( \varphi \) – the content of the class, \( t_m \) – the point in time at which clustering is implemented.

The scheme for modeling the state dynamics of a social media message classification system can generally be implemented based on the generalized experience of stochastic modeling and taking into account the classification of model schemes.

The process of dynamic clustering of messages in social media can be represented by a Markov model with discrete states, where \( S_0 \) – starting state of the classification system; \( S_n \) – the n-th state that the social media message classification system finds itself in after n-th transition; \( \lambda_{ij} \) – intensity of transition from the state i into j, that is, the intensity of classification of messages in social media, figure
4.

\[
\begin{align*}
\lambda_{01}, \lambda_{12}, \lambda_{2i}, \lambda_{jd}, \lambda_{dp} \\
S_0 & \quad S_1 & \quad S_2 & \quad \ldots & \quad \rightarrow S_n \\
\lambda_{01} & \quad \lambda_{21} & \quad \lambda_{ji} & \quad \lambda_{dj} \\
\end{align*}
\]

**Figure 4.** State graph of the process of changing the composition of clusters when clustering messages in social media based on the process of death and reproduction.

According to the well-known analytical scheme for solving this problem, the probabilities of being in States can be written by a system of differential equations, taking into account the normalization condition \( \sum_{i=1,I} P_i = 1 \):

\[
\begin{align*}
\frac{dP_0}{dt} &= \lambda_{i0}P_1(t) - \lambda_{01}P_0(t), \\
\cdots & = \lambda_{i,i+1}P_{i+1}(t) - \lambda_{i+1,i}P_i(t), \\
\frac{dP_d}{dt} &= \lambda_{jd}P_j(t) - \lambda_{dj}P_d(t), \\
\end{align*}
\]

provided that the initial probability distribution has the form: \( P_0(t) = 1, P_1(t) = 0, \ldots, P_d(t) = 0 \). The final value of the probability of the system being in a state based on (4) in the scheme of death and reproduction is equal to:

\[
P_0 = \left[ 1 + \frac{\lambda_{01}}{\lambda_{01}}, \frac{\lambda_{01}\lambda_{12}}{\lambda_{10}}, \ldots, \frac{\lambda_{01}\lambda_{12}\ldots\lambda_{d-1,d}}{\lambda_{10}\lambda_{21}} \right]^{-1}.
\]

Thus, the model (4) and its solution (5) allow us to estimate the probability of clustering dynamics in the problem under consideration. However, to analyze class content, it is advisable to consider three types of events: when the power of the class remains unchanged \((q_\phi = \text{const})\), increases or decreases by 1 respectively, (state dynamics, figure 3), in \( t_{m-1} < t_m < t_{m+1} \) – time moment

\[
\begin{align*}
Q_0(t_{m+1}) &= \left( \xi(t_{m+1}), \varphi(t_m), t_{m+1} \right), \\
Q_{+1}(t_{m+1}) &= \left( \xi(t_{m+1}), [\varphi + 1], t_{m+1} \right), \\
Q_{-1}(t_{m+1}) &= \left( \xi(t_{m+1}), [\varphi - 1], t_{m+1} \right),
\end{align*}
\]

where \( Q_0(t_{m+1}), Q_{+1}(t_{m+1}), Q_{-1}(t_{m+1}) \) – estimates of the anticipated capacity of the cluster at \( t_{m+1} \) time moment.

For some small \( \Delta t > 0 \) and the specified values of cluster capacity estimates at the initial time \( Q_+(t_0) = qa, Q_{-1}(t_0) = qb, Q_0(t_0) = qc \) for \( t_{m-1}, t_m \in [0,T] \): \( t_{m-1} < t_m < t_{m+1} \) estimates of the expected cluster capacities can be represented by a system of ordinary linear differential equations under the assumption that \( qa, qb, qc \) are differentiable, and the functions on the right are continuous over the time interval under consideration.
where $\beta_{00}, \beta_{01}, \beta_{+11}, \beta_{-11}, \beta_{-12}$ — parameters that determine the degree of impact of the corresponding functions on the power ratings of social media message clusters, $\eta_0^+, \eta_+^-, \eta_-^-$ — random functions that describe possible uncertainty in estimating the power of social media message clusters.

### 3. Conclusion

Based on (7) and the definition of clustering dynamics (figure 3), along with the general operations of forming new clusters, various dynamic clustering operations are defined: cluster merging, merging and division, and the formation of new clusters, and others, figure 5.

Measure of similarity between clusters formed as a result of transitions of the clustering system to states $S_1(t_{m-1}) \rightarrow S_2(t_m) \rightarrow S_3(t_{m+1})$ can be defined by the expression

$$C(\varphi_i^1, \varphi_j^2) = \frac{2m(\varphi_i \cap \varphi_j)}{(1+u)(m(\varphi_i) + m(\varphi_j) - 2um(\varphi_i \cap \varphi_j))},$$

where $-1 \leq u \leq \infty$ and is determined by the characteristics of the subject area; $m(\varphi_i)$ — denoting the power of a set of clusters $\varphi_i$ [9].

Taking into account (7) and operation schemes, figure 5, the decision maker generates restrictive estimates for specific clusters, for example: $C_{min}^A(\varphi_i^1, \varphi_j^2)$, $C_{max}^A(\varphi_i^1, \varphi_j^2)$, where $A$ — measure of similarity between clusters when forming a cluster $A$. However, for some cluster $B$ these restrictions may be different: $C_{min}^B(\varphi_i^1, \varphi_j^2)$, $C_{max}^B(\varphi_i^1, \varphi_j^2)$.

### Acknowledgments

This work was carried out within the framework of an internal grant "Development of agent-based modeling and big data methods for social media analysis in post-conflict societies (grant №28/06-31).

### References

[1] Kogan J, Nicholas C and Teboulle M 2017 Clustering Large and High Dimensional data Department of Computer Science and Electrical Engineering 12 142-51

[2] Barseghyan A A 2009 Data and Process Analysis 3d edition (Saint Petersburg: BChV-Peterburg)

[3] Yeung D and Ding Y 2003 Host-Based Intrusion Detection Using Dynamic and Static Behavioral Journal of Pattern Recognition 36 229-43

[4] Hastie T, Tibshirani R and Friedman J 2001 The Elements of Statistical Learning (Moscow: Fizmatlit)

[5] Zakharov V M, Nurmeev N N, Salimov F I and Shalagin S V 2000 Classification of stochastic ergodic matrices by cluster and discriminant analysis methods Computer science research 3 91-106
[6] Website of BaseGroupe Labs http://www.basegroup.ru/clusterization/datamining.htm
[7] Wang K, Xu C and Liu B 1999 *Clustering transactions using large items* (Kansas, Missouri: CIKM'99)
[8] Datiev I O 2006 Applied problems of macro-systems operations control 28 368
[9] Andreychikov A V and Andreychikova O N 2000 *Analysis, Synthesis, Planning of Decisions in Economy* (Moscow: F and S)