UDK 004.9

Y. Stekh, V. Artsibasov
Lviv Polytechnic National University
CAD Departament

ADAPTIVE CLUSTERING ALGORITHM FOR RECOMMENDER SYSTEMS

In this article adaptive clustering algorithm for recommender systems is developed.
Key words: Recommender system, clustering algorithm, group of users, user profiles.

Розроблено адаптивний алгоритм кластеризації для рекомендаційних систем.
Ключові слова: рекомендаційна система, алгоритм кластеризації, групи користувачів, користувацькі профілі.

Introduction

Recommendation systems have many advantages that make this technology attractive for users. This technology allows users to spend a little time to find the information you need on the Internet, choose the most appropriate products and services [1]. These systems are compared to data collected from different users and create a list of items recommended by the user. They are an alternative search algorithm, because they help users to find items and information that they could not find on their own. A crucial role in recommendation systems are classification algorithms – k-neighborhood, which is used in a social network user, as well as collaborative filtering algorithm. Important for improving the performance of recommendation systems is the development, exploration and use of effective methods and algorithms for clustering and classification. These methods and algorithms are used to find groups (clusters) of users with common interests and preferences [2], [3]. The aim of this article is to develop and study a method k-means, which does not require an initial set of cluster centers and the provision and allows you to search for groups (clusters) of user profiles of a spherical shape with an automatic choice of the radius of the sphere.

Method and algorithm

The most common methods of non-hierarchical clustering algorithm is k-means [4], [5], [6]. This algorithm minimizes the quality score, defined as the sum of squared distances of all points within the cluster region, the center of the cluster. The algorithm consists of the following steps.

Step 1. Selects the initial K cluster centers $\mathbf{Z}_1$ (1), $\mathbf{Z}_2$ (2),..., $\mathbf{Z}_k$ (1).

This choice is arbitrary and, usually used as the initial centers of the first k sample results from a given set of information models of objects.

Step 2. At the *l*-step of iteration given set of information model objects are distributed objects on K clusters according to the following rule:

$$\mathbf{X} \in K_j \text{ if } \left\|\mathbf{X}-\mathbf{Z}_j(l)\right\| < \left\|\mathbf{X}-\mathbf{Z}_i(l)\right\|,\ i, j = \overline{1,n} \tag{1}$$

where K_j-set of images that are part of a cluster with center $\mathbf{Z}_j$ (l). In the case of a tie in an arbitrary way decisions are made.

Step 3. Based on the results of step 2 are determined by the new cluster centers $\mathbf{Z}_j$ (l +1), j = 1,2,..., k, based on the condition that the sum of the squares of the distances between all information model objects that belong to $\mathbf{S}_j$ (l), and new center cluster should be minimized. New centers of clusters $\mathbf{Z}_j$(l +1) are chosen in such a way as to minimize the quality measure

$$F_j = \sum_{\mathbf{X} \in K_j} \left\|\mathbf{X}-\mathbf{Z}_j(l+1)\right\|^2 \tag{2}$$

Center for $\mathbf{Z_j}$ (l +1) minimizes quality measure and is sample mean, defined on the set K_j. New cluster centers are defined as follows:

$$\mathbf{Z}_j(l+1)=\frac{1}{n_j}\sum_{\mathbf{X}\in K_j}\mathbf{X},\ j=\overline{1,K} \tag{3}$$

where n_j – number of sample information model objects included in the set *Kj*.

Step 4. If $\mathbf{Z}_j(l+1) = \mathbf{Z}_j(l)$, $j=\overline{1,K}$, then STOP, otherwise go to step 2.

The quality of the algorithm depends on:

- the number of pre-selected cluster centers;
- the choice of initial cluster centers;
- the sequence of viewing images;
- the geometric features of the data.

Practical application of the algorithm requires experiments with a choice of different values of K and the initial placement of cluster centers. Select the number of clusters and the initial position is a difficult task. If no assumptions about this number, then set increasing sequence the number of clusters (2, 3, 4,...). and then compare the results. The initial location of cluster centers chosen at random. The optimality of the results obtained estimate of the distance between the centers of clusters and the number of information models in each cluster. The distance between the centers of clusters should be maximized, and the amount of information models in each cluster – the minimum. Advantages of the algorithm: ease of use, speed of convergence, clarity and transparency of the algorithm.

Disadvantages of this algorithm is the unsolved problem of the choice of the initial number and positions of the centers of clusters. We proposed a modified algorithm for the threshold which is based on the traditional threshold algorithm [5, 7]. The difference is that first defined the center of the cluster as the most remote from all other objects of the search space. Next is found the center of mass of the cluster, and it moved the center of the cluster. Next to this cluster are assigned all objects whose distance is less than or equal to the threshold T. The process continues until the center is not stabilized. Points that are included in the cluster are excluded from further consideration and the final form the first cluster. After that, among the many remaining objects chosen a new center of the cluster which is most distant from all other centers and the above procedure is performed. The algorithm is executed as long as all objects are not clustered. Feature of the algorithm is that the first cluster center is chosen as the most remote of all the objects search space (step 4). The remaining centers of clusters are chosen as the most remote from the centers have already been found (step 11). The flowchart of adaptive clustering algorithm is shown in Fig. 1.

The algorithm consists of the following steps.

Step 1. Determine the minimum and maximum distance between data models of objects

$$\begin{aligned} l_{\min} &= \min\left\{ l_{ij} \,\middle|\, i,j=\overline{1,n} \right\}, \\ l_{\max} &= \max\left\{ l_{ij} \,\middle|\, i,j=\overline{1,n} \right\} \end{aligned}. \tag{4}$$

Step 2. Determine the maximum value of the threshold

$$T_{\max}=\frac{l_{\min}+l_{\max}}{2}. \tag{5}$$

Step 3. Determine the step and the initial value of threshold

$$\Delta T = b\cdot T_{\max}, T=\Delta T, \tag{6}$$

where $b\in(0,1)$.

Step 4. Select the first point of the cluster center

$$m_i=\sum_{j=1}^{n}\left(\frac{l_{ij}}{\sum_{l=1}^{n} l_{jl}}\right),\ m_p=\min\{m_i \,|\, i=\overline{1,n}\}, \tag{7}$$

$$\mathbf{Z}_1=\mathbf{X}_p. \tag{8}$$

Step 5. Set the initial value of the index cluster

$$q = 1,\ B_q = D. \tag{9}$$

Step 6. Determine the set of points that belong to q – cluster that follows

$$\mathbf{X}_i \in K_q, \text{ if } l_{iq} < T, \tag{10}$$

where $l_{iq} = \left\| \mathbf{X}_i - \mathbf{Z}_q \right\|$, $i = \overline{1,n}$.

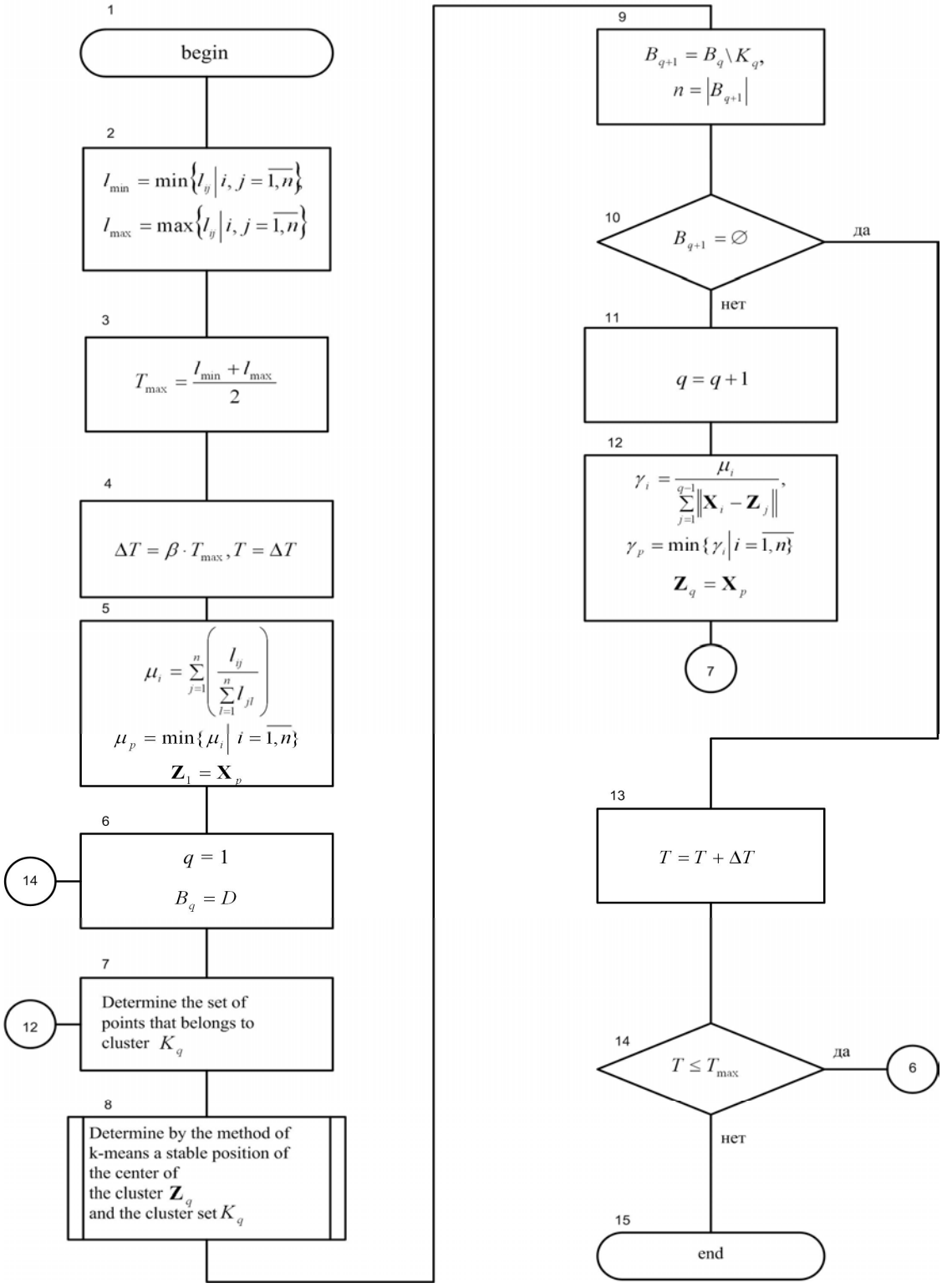

Fig. 1. The flowchart of adaptive clustering algorithm

Step 7. Apply *k*-means method to determine the stable position of q – th center of the cluster $\mathbf{Z}_q$ and the set of points that belong to the q – th the cluster K_q.

Step 8. Delete the set of points K_q in the cluster from further consideration

$$B_{q+1} = B_q \setminus K_q,\ n = \left|B_{q+1}\right|, \tag{11}$$

streamline the indexing of elements of given set B_{q+1} taking into account the index of remote points.

Step 9. If $B_{q+1} = \varnothing$ then go to step 13.

Step 10. $q = q + 1$.

Step 11. Determine the next point cluster center $\mathbf{Z}_q$ with the previous cluster centers

$$g_i = \frac{m_i}{\sum_{j=1}^{q-1} \left\| \mathbf{X}_i - \mathbf{Z}_j \right\|}, \tag{12}$$

$$g_p = \min\{ g_i \,|\, i = \overline{1,n} \},\ \mathbf{Z}_q = \mathbf{X}_p \tag{13}$$

Step 12. Go to step 6.

Step 13. Increase the value of the threshold $T = T + \Delta T$.

Step 14. If $T \le T_{\max}$ then go to step 5, otherwise – STOP.

Conclusion

In this paper we propose an adaptive algorithm for non-hierarchical clustering which is based on k-means algorithm and threshold algorithm. We have developed an algorithm does not require specifying the number of initial centers of clusters, the initial position of the centers of clusters and initial threshold value.

1. Adomavicius G., Tuzhilin A., Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions // IEEE Trans. Knowledge and Data Engineering. – Vol. 17. – P.734–749, Jun, 2005. 2. Christinsen I. A., Schiaffino S. Entertainment recommender systems for group of users // Expert Systems with Applications. – Vol. 38. – P. 14127–14135, 2011. 3. Tang X., Zeng Q. Keyword clustering for user interest profiling refinement with paper recommender system // Journal of Systems and Softwre. – Vol. 2. – P. 87–101, 2011. 4. Saegusa T. An FPGA implementation of real-time K-means clustering for color images // Real Time Image Processing. – Vol. 2. – P. 309–318, 2007. 5. Stekh Y., Lobur M., Faisal M.E. Sardieh, Dombrova M., Artsibasov V. Research and development of methods and algorithms non-hierarchical clustering," in Proc. of the XIth International Conference CADSM, Lviv-Polyana, 2011. – P. 205–207. 6. Lobur M., Stekh Y., Kernytskyy A., Faisal M.E. Sardieh Some trends in knowledge discovery and data mining in Proc. of the IVth International Conference MEMSTECH, Lviv-Polyana, 2008. – P. 205–207. 7. Загоруйко Н.Г. Алгоритмы обнаружения эмпирических закономерностей – Новосибирск: Наука, 1985. – 63с.