Application of the descriptor approach for clustering entities from education sector

Anna Zykina[1,2], **Olga Kaneva**[1,3], **Ivan Sharun**[1,4]
[1]Omsk State Technical University, Mira, ave. 11, Omsk, 644050, Russia

E-mail: [2]avzykina@mail.ru, [3]okaneva@yandex.ru, [4]ivan@sha.run

Abstract. The article proposes the algorithm to solve objects clustering problem for such subject areas as education and labour market. Such objects are competence, discipline, specialty, vacancy, etc. The main problem in clustering algorithm development proved to be the stage of attributes design since the named objects have descriptions in a natural language. Consequently, a descriptive model for the objects was designed at first. The model was based on the fact that all necessary concepts are characterised in the space of descriptors "know", "can", etc.

This allowed the object to be represented as a tuple based on the object name, descriptors (and their values) and keywords related to descriptors. To obtain such structures, the toolkit of context-relative text mining was used. The ability to represent the entities in question as a formal structure allowed the attribute space formation algorithm to be developed and complex metrics to be constructed to solve the stated clustering problem. The developed algorithm permits development of various services for a faster and more objective decision making process in educational and professional sectors.

As a result of the work, about two thousand vacancies were obtained and transformed into descriptor entities.

Based on the error matrix, it can be judged that the resulting descriptor entities have been clustered with a sufficient level of quality. This demonstrates the applicability of the model for presentation and analysis for elements from the subject area.

1. Introduction

The relevance of developing a clustering algorithm for concepts from such subject areas like education and the labour market is due to the introduction of new educational standards in Russian education.

The educational core of a direction is given by universal general professional competences. Specialisation is set by professional competences whose formulation is the responsibility of the developers of educational programs.

The process of detecting professional competences is complex and ambiguous. It suggests a link between the education core and the employers' needs. Analysis of the employers' needs is possible, for example, through vacancies. Having learned to compare vacancies, we can cluster and select groups of similar employers' requirements. Having then determined which of these groups are closest to the educational core of the direction for which the educational program is being developed, we use them as the foundation of professional competences [1].

This clustering-based scheme based assumes the presence of tools to compare such concepts as competence, vacancy, educational direction, discipline, teacher's specialization, and etc. The

main problem is that all these entities are described in a natural language and their analysis reveals the subjectivity of researchers in their interpretation. To make the comparison process more formal, we need a basis (space) in which they would become comparable. Such a basis can be their representation through descriptors "know", "can", etc. In the following, for the purpose of brevity, the concepts that can be represented through the above descriptors will be called descriptor entities[2].

2. The formalisation of the subject area

2.1. Formalised representation of a descriptor entity

The following tuple will be understood as a descriptor entity $K =< S, D, T >$, wherein:

- S^k is the formulation of the descriptor entity K in a natural language, i.e. text string.
- $D^k = \{D_1, D_2, ..., D_m\}$ – is a set of descriptor entity's descriptors K;
  $D_j =< S_{D_j}, V^j >,\ j = \overline{1,m}$, where: S_{D_j} – is the formulation of the descriptor via the infinitives such as "know", "can", etc.;
  $V^j = \{v_1^j, v_2^j, ..., v_{p_j}^j\}$ – is a set of descriptor values;
  $v_{p_j}^j$ is an expression in a natural language representing the explication of the formulation S_{D_j} (for example, "know discrete mathematics").
- $T^k = \{T_1, T_2, ..., T_n\}$ — is a set of terms for the descriptor entity K;
  $T_i =< S_{T_i}, W^i >,\ i = \overline{1,n}$, where: S_{T_i} is the formulation of the term in natural language (can be both a word and a word combination);
  $W^i = \{W_1^i, W_2^i, ..., W_m^i\}$ – is a set of term's weights for each descriptor from D;
  $W_j^i = (w_{j1}^i, w_{j2}^i, ..., w_{jp_j}^i)$, where $w_{jp_j}^i$ is term's weights for descriptor's values. (i.e. for $v_{p_j}^j$).

2.2. The algorithm for building descriptor entities

(i) Extract text from a document.

(ii) Extract the necessary block of information.

(iii) Split a piece of text into sentences.

(iv) Convert sentences to descriptors.

(v) Match keywords to descriptors.

2.3. The algorithm for finding semantic similarity

Similarity coefficient (similarity index) is a dimensionless index of similarity for compared objects.

It is also known as "measure of association," "measure of similarity," etc.

Similarity coefficient is used in biology to quantify the degree of similarity of biological objects (sites, areas, individual phytocenoses, zoocenoses, etc.)[3, 4, 5]. They are also used in geography, sociology, pattern recognition, search engines, comparative linguistics, bioinformatics, chemical informatics, string comparison, etc.

In a broader sense, there are measures of proximity such as diversity measures, concentration (homogeneity) measures, inclusion measures, similarity measures, differentiation measures (including distance), event compatibility measures event incompatibility measures, interdependence measures, inter-independence measures, etc. There are many different ideas about the formalisation of relations of similarity.

Most coefficients are normalised and range from 0 (no similarity) to 1 (complete similarity). Similarity and difference complement each other.

Similarity coefficients can be divided into three groups, depending on the number of objects considered:

- Unary – one object is considered. This group includes diversity and concentration measures.
- Binary – two objects are considered. This is the most known group of coefficients.
- Multinary – a set of objects are considered. This group is the least known.

Let us choose the Jaccard binary similarity coefficient as the basis for our algorithm. Thus, the similarity between two descriptor entities can be expressed by the formula:

$$SIM_K(K_1,K_2)=\frac{\sum\limits_{p_j^{K_1}}\sum\limits_{p_j^{K_2}}\left[SIM_w(v^j_{p_j^{K_1}},v^j_{p_j^{K_2}})\times\frac{\sum\limits_z\min(\overline{W}_{zp_j^{K_1}},\overline{W}_{zp_j^{K_2}})}{\sum\limits_z\max(\overline{W}_{zp_j^{K_1}},\overline{W}_{zp_j^{K_2}})}\right]}{\sum\limits_{p_j^{K_1}}\sum\limits_{p_j^{K_2}}SIM_w(v^j_{p_j^{K_1}},v^j_{p_j^{K_2}})},\quad z=\overline{1,|T^{K_1}\cup T^{K_2}|} \tag{1}$$

Consider the algorithm:
Input: two descriptor entities K_1 and K_2. Output: number $\in$ [0..1] that characterises the similarity of two descriptor entities.

Requirements:

(i) Descriptors are the same and lexicographically sorted.
(ii) Terms are lexicographically sorted.

Algorithm process:

(i) Take the descriptors that have the same indices from K_1 and K_2. Compile for them $\overline{W}_{zp_j^{K_1}}$ and $\overline{W}_{zp_j^{K_2}}$
   (a) Make a matrix of terms coefficients for the descriptor values of the first descriptor entity. The lines are the terms, the columns are the descriptor values, the intersection is the term weight for the value.
   (b) In the same way, make a matrix for the descriptor from the second descriptor entity.
   (c) In case of discrepancy of the term sets, complement each matrix with the terms from another matrix, and put 0 at the intersection.
   (d) Sort the strings in the lexicographical order.
(ii) Calculate the value of the formula for two descriptors.
   (a) Calculate the sum of the SIM multiplications and the ratios of the sums of the minimum and maximum weights from the matrices $\overline{W}$ for each with each descriptor value.
   (b) Divide by the sum values of SIM for each with each descriptor values.
(iii) If there are remaining descriptors, then return to step 1.
(iv) Do a convolution of the results.

With this algorithm, we will be able to apply classical clustering methods to descriptor entities[6, 7, 8].

3. Numerical experiment

The experiment consists of several stages:

(i) Loading and preprocessing data.
(ii) Processing by clustering algorithm.
(iii) Clustering quality assessment.

The results of each step are presented in Fig. 1, Fig. 2 and Fig. 3

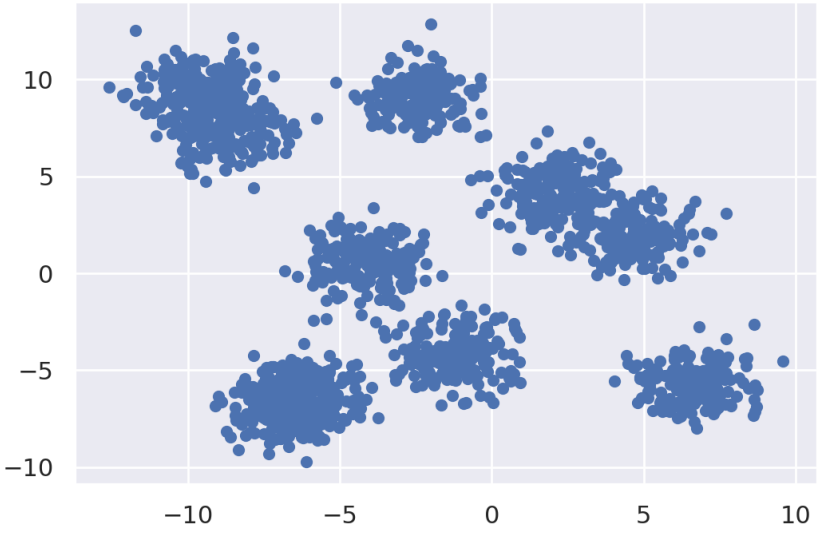

Figure 1. Data visualisation.

3.1. Loading and preprocessing data.
The data source is one of the most popular job search sites. Each vacancy was converted to a descriptor entity using a conversion algorithm.

By transforming the matrix of distances into coordinates, we were able to represent each

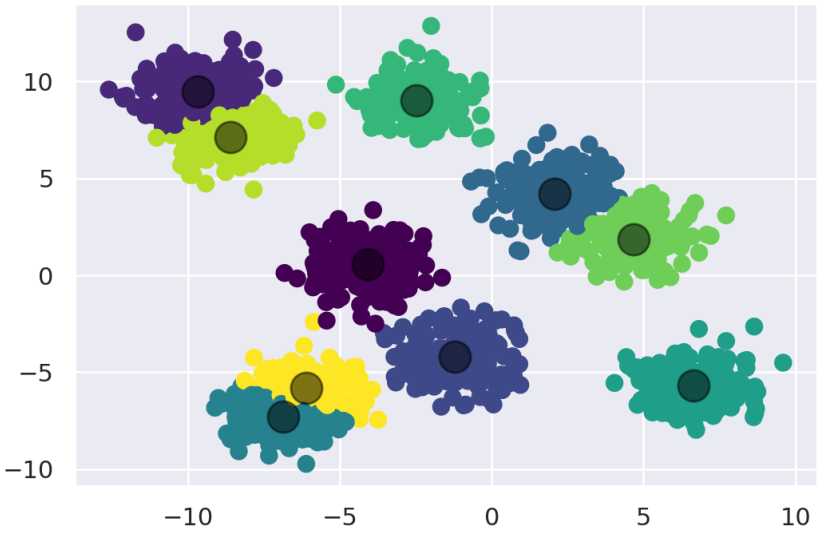

Figure 2. The result of the K-MEAN algorithm.

vacancy on a two-dimensional plane (see Fig. 1). In total, about two thousand vacancies were uploaded and preprocessed.

3.2. Processing by clustering algorithm.
This set of descriptor entities was processed by the K-means algorithm. The processing results can be seen in Fig. 2.

3.3. Clustering quality assessment.
To assess the quality of the error matrix has been drawn up. The results of the algorithm evaluation are presented in Fig. 3.

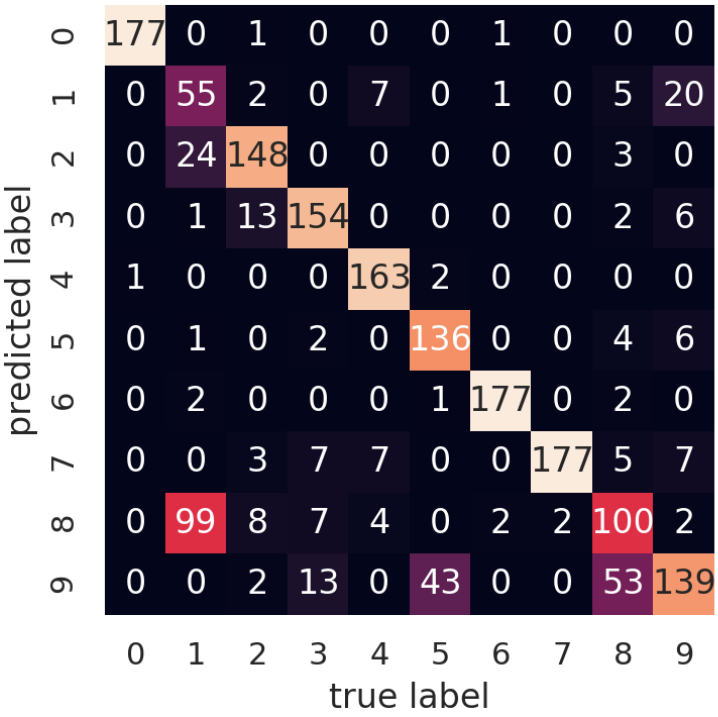

Figure 3. Error Matrix for Clustering Results.

4. Experiment Results

As a result of the work, about two thousand vacancies were obtained and transformed into descriptor entities.

Based on the error matrix, it can be judged that the resulting descriptor entities have been clustered with a sufficient level of quality. This demonstrates the applicability of the model for presentation and analysis for elements from the subject area.

5. Conclusion

The article presents the formalisation of such concepts as competence, vacancy, educational direction, discipline, and teacher's specialisation through the proposed descriptor entities. The

task of clustering the proposed descriptive entities is considered. Metrics for already existing clustering algorithms are proposed.

The result of the work are the developed approaches to define the groups of similar employers' requirements.

An approach and tools have been developed for comparing such concepts as competence, vacancy, educational direction, discipline, teacher's specialisation, etc.

Numerical investigations of the constructed models are of practical importance: these investigations have shown the possibility of an automated approach using cluster analysis to identify groups of similar employers' requirements.

References

[1] Zykina A.V., Kaneva O.N., Munko V.V. 2018 *Algorithm of determining the competence of educational programs of higher education* (Informational bulletin of Omsk Scientific and Educational Center of OmSTU and SB RAS in the field of mathematics and computer science vol 2 no 1) pp 74–77

[2] Kaneva O.N., Akimova K.A., Sharun I.V. 2017 *Finding semantic similarity between the competences algorithm development* (Informational bulletin of Omsk Scientific and Educational Center of OmSTU and SB RAS in the field of mathematics and computer science vol 1 no 1) pp 155–157

[3] Jaccard P. 1901 *Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines* (Bull. Soc. Vaudoise sci. Natur. vol 37) pp 241—272

[4] Sörensen T. 1948 *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content* (Kongelige Danske Videnskabernes Selskab. Biol. krifter. vol 4) pp 1-34

[5] Simpson G.G. 1947 *Holarctic mammalian faunas and continental relationship during the Cenozoi* (Bull. Geol. Sci. America. vol 58) pp 613—688

[6] Steinhaus H.1956 *Sur la division des corps materiels en parties* (Bull. Acad. Polon. Sci., C1. III vol 4) pp 801—804

[7] Lloyd S. 1957 *Least square quantization in PCM's.* (Bell Telephone Laboratories Paper)

[8] Hastie T., Tibshirani R., Friedman J. 2001 *The EM algorithm. The Elements of Statistical Learning* (New York : Springer) pp 236–243.