METHOD OF DEVELOPING ARTIFICIAL PATTERNS TO ANALYZING SEMANTIC SIMILARITY OF DOCUMENTS

Tuliev Ulugbek
PhD, Senior Lecturer
National University of Uzbekistan.

Abdurakhmonova Nilufar
Doctor of science, professor
National University of Uzbekistan.

Annotation: A technique for creating artificial patterns using data from textual documents is considered. There is a need for patterns to reduce the dimensionality of the feature space and analyse the semantic similarity of thematic documents. The composition of patterns is formed according to the rules of hierarchical agglomerative grouping. The formation process is based on the selection of a latent feature space using the technology for calculating generalized assessments of objects.

Keywords: semantic similarity of documents, latent feature.

Аннотация: Рассматривается методика создания искусственных паттернов по данным из текстовых документов. Существует потребность в паттернах для снижения размерности признакового пространства и анализа семантической близости тематических документов. Состав паттернов формируется по правилам иерархической агломеративной группировки. Процесс формирования основывается на выборе латентного признакового пространства по технологии вычисления обобщённых оценок объектов.

Ключевые слова: семантической близости документов, латентный признак.

Studying the structure of object relationships is one of the sources of obtaining new knowledge in intelligent systems. In this work, an artificial pattern is understood as a stable set of entities (words, concepts) of natural language (NL) for distinguishing text documents from subject areas. It is assumed that for the stability (compactness) of sets, numerical indicators are defined, the calculation of which is based on strict mathematical principles.

One of the problems when analyzing the structure of document descriptions in problems of mathematical linguistics is the problem of the curse of dimensionality. It is believed that the number of patterns is determined algorithmically based on the rules of hierarchical agglomerative grouping.

Reducing the dimension of the feature space is carried out in order to solve the problem of blurred relationships between objects. The blurriness assessment is based on quantitative indicators of the structure of relations between class objects. As a rule, when measuring indicators, distance measures between objects are used. Due to the large number of features used to describe topical documents, this inevitably leads to the well-known phenomenon of the curse of dimensionality in data analysis. The negative effect of this phenomenon is a significant limitation of the possibilities for applying classical methods of data mining. Using methods of linear or

nonlinear mapping of object descriptions into spaces of lower dimensions also does not answer the questions:

- What dimension of space should be after transformation?
- How to measure display quality for decision making?
- By what properties can we distinguish between document descriptions in different languages?

The selection of noisy features is considered as a constrained optimization problem. A separate problem is the choice of restrictions for the implementation of numerical solution methods, taking into account the availability of classification of thematic documents. Criteria [Ignatyev 2018, 590; Ignatev, Tuliyev 2022, 1185;]given below is used to find optimal boundary of the values of features

$$\left(\frac{\sum_{p=1}^{2}\sum_{i=1}^{2} u_i^p \left(m - |K_i| - \sum_{j=1}^{2} u_j^p + u_i^p \right)}{2|K_1||K_2|} \right) \left(\frac{\sum_{p=1}^{2}\sum_{i=1}^{2} u_i^p \left(u_i^p - 1 \right)}{|K_1|\left(|K_1|-1\right) + |K_2|\left(|K_2|-1\right)} \right) \rightarrow \max_{c_1 < c_2 < c_3} \quad (1)$$

where u_i^j-number of objects which belong to K_i class in j-interval, m-number of objects in data set, $|K_i|$-number of objects which belong to K_i class.

The problems of reducing the dimension of the feature space, semantic relatedness of documents and taking into account polysemy in their various manifestations are solved [Tuliyev 2019, 102]. With the help of patterns, it becomes possible to automate the process of generating dictionaries for presenting documents in tabular form. It is relatively easy to interpret polysemy in scientific texts through glossaries and thesauri. The problems are aggravated in works of art in which the authors, through polysemy, seek to show the richness of the language.

The stability of the feature $x_c \in X(n) \backslash Z$ is calculated as []

$$\varphi(c) = \frac{1}{m}\sum_{r=1}^{m} \begin{cases} b_{rc}, b_{rc} > 0.5, \\ 1 - b_{rc}, b_{rc} < 0.5. \end{cases} \quad (2)$$

The set of admissible values (2) belongs to (0.5;1]. Stability $\varphi(c)$=1 if, according to the boundary [Ignatev, Tuliyev 2022, 1185], objects are divided into classes K_1 and K_2 without errors. The ordering of words according to (2) in general does not depend on dimensions of space for describing a collection of documents.

The point of calculating the coefficient of content authenticity is to find the optimal number of clusters that determine the division of documents into topics in the latent feature space [Ignatyev 2018, 590; Tuliyev 2021, 2365]. To form a latent space and a common dictionary based on it, it is proposed:

- order the features with indices from D in non-decreasing order of their stability values (2) as
$$\varphi(\varepsilon_1), \dots, \varphi(\varepsilon_j), \dots, \varphi(\varepsilon_{dim}), \varepsilon_j \in D, dim = |D|; \quad (3)$$
- generate a set of latent features $Y(t)$= (y_1, . . ., yt) according to (3);
- select $Y(\sigma) \subset Y(t)$, $\sigma \leq t$ and the set of raw features related by D to $Y(\sigma)$ as a common dictionary.

The number of words for the general vocabulary from (3) is, in general, a free parameter. This parameter can be set for heuristic reasons, or based on the results of hierarchical agglomerative grouping when forming a set of latent features.

Let us denote by TYPLAM the set of indices of raw features included in the group by the hierarchical grouping algorithm, lugat the limit on the number of words in the general dictionary, guruh the number of latent features. The implementation of the algorithm in steps will be as follows [Ignatev, Tuliyev 2022, 1185].

Step 1. j=0. *guruh*=0.

Step 2. Computing $j=j+1$. *crit*=10. $u=\varepsilon_j$. *TYPLAM* = $\{u\}$. *guruh*=*guruh*+1.

Loop for $t\in\{1,...,m\}$ $R(S_t)=\eta_u(a_{tu})$. End of loop;

Step 3. $u=\varepsilon_{j+1}$. Loop for $t\in\{1, ..., m\}$ $b_t = R(S_t) + \eta_u(a_{tu})$. End of loop;

$M_1 = \sum_{S_t\in K_1} b_t$. $M_2 = \sum_{S_t\in K_2} b_t$. $M_1=M_1/|K_1|$. $M_2=M_2/|K_2|$. θ=0. γ=0. Loop for $t\in\{1, ..., m\}$ If $S_t\in K_1$, то $\theta=\theta+|b_t-M_1|$, $\gamma=\gamma+|b_t-M_2|$. Else $\theta=\theta+|b_t-M_2|$, $\gamma=\gamma+|b_t-M_1|$. End of loop;

Step 4. If θ/γ < *crit*, then *crit*= θ/γ , *TYPLAM* =*TYPLAM* $\cup\{u\}$, $j=j+1$, go to 3.

Step 5. Print $\{R(S_t)\}_{t\in\{1,...,m\}}$, *TYPLAM*.

Step 6. If $j<$*lugat*, then go to 2; Else print *guruh*.

Step 7. End.

To demonstrate the hierarchical algorithm described above in Table. 1 we present the results of its implementation on text documents presented by abstracts of scientific dissertations. Class K_1 is represented by documents in physics and mathematics, K_2 - from 10 different subject areas.

Table 1. Latent features and their composition from raw features

№	Number of features	Features	Value of criteria (1)
1	267	179, 55, 171, … , 414, 297, 299	0.6360
2	87	423, 66, 450, … , 186, 36, 265	0.5061
3	66	70, 67, 192, …. , 290, 15, 21	0.3917
4	10	393, 61, 75, 253, 163, 405, 53, 24, 44, 490	0.2793
5	23	322, 408, 295, … , 311, 364, 316	0.3493
6	1	89	0.2442
7	16	139, 189, 166, … , 305, 199, 304	0.3492
8	8	50, 158, 223, 195, 427, 302, 433, 369	0.2372
9	2	98, 197	0.2127
10	1	177	0.1852
11	2	57, 279	0.2344
12	2	120, 135	0.2460

References:

1. Ignatyev N. A. Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis. 2018. V. 28. № 4. P. 590–597.
2. N. A. Ignatev, U. Y. Tuliyev, “Semantic structuring of text documents based on patterns of natural language entities”, Computer Research and Modeling, 14:5 (2022), 1185–1197 http://doi.org/10.20537/2076-7633-2022-14-5-1185-1197
3. N. Abdurakhmonova, U. Tuliyev and A. Gatiatullin, "Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz," 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 1-4, http://doi.org/10.1109/ICISCT52966.2021.9670043

4. Tuliyev U. (2021). Space formation for the description of thematic documents. AIP Conference Proceedings. 2365. 070007. http://doi.org/10.1063/5.0056963
5. Тулиев У. Ю. Кластерный анализ текстовых документов по отношению их связности // Проблемы вычислительной и прикладной математики. — 2019, No 6(24). — С. 102–109.
6. N.A. Ignatyev, Sh.F.Madrakhimov, D.Y.Saidov. Stability of object classes and selection of the latent features // International journal of engineering technology and sciences, 2017, Malaysia, Vol. 7, pp. 1-10.
7. Игнатьев Н.А., Саидов Д.Ю. Анализ данных и принятие решений с помощью логических закономерностей в форме полуплоскостей // Известия СамНЦ, 2017, Том 19, № 4(2), С. 294-300.