Similarity Metric of Categorical Distributions for Topic Modeling Problems with Akin Categories

Serhiy Shtovba[a], Mykola Petrychko[b] and Olena Shtovba[b]

[a] *Vasyl' Stus Donetsk National University, 600-richchia str., 21, Vinnytsia, 21021, Ukraine*
[b] *Vinnytsia National Technical University, Khmelnytske Shose, 95, Vinnytsia, 21021, Ukraine*

Abstract

Estimating a level of similarity of two objects is a common problem in pattern recognition, clustering, and classification. Among these problems can be reviewer recommendation, similar text documents analysis, human pose detection in video, species distribution clustering, recommendation in internet-shops etc. In case of categorical attributes an object is described as a distribution of membership degrees over categories. Similarity metrics of such distributions are usually defined as a superposition of objects' similarities for each category. Most often it is a sum of similarities in separate categories. In addition to that each category is considered independently and in isolation from the others. Some practical problems have categories that are akin. Therefore, it is expedient to consider objects' similarity not only directly, as a similarity between equivalent categories, but it is also necessary to consider an indirect similarity, cross-similarity through akin categories. It is such similarity metric of two categorical distributions that accounts for the kinship of different categories is proposed in this paper. The metric has two components. The first component is defined as Czekanowski metric. It defines a direct similarity of categorical distributions as a sum of intersection of distributions' membership degrees of two objects. After the intersection the remaining residuals are accounted for in the second component of the metric. The second metric's component is defined as element-wise product of two matrices: matrix of residuals composition from memberships of two categorical distributions and matrix of categories' paired kinship. It is assumed that kinship indices for each pair of categories are known. As a result, with a large number of categories the overall noisy contribution from weakly akin categories is prominent. Therefore, it is proposed to filter the noise and account only for contribution from strongly akin categories.

Keywords 1

Categorical distribution, akin categories, kinship coefficient, similarity metric, Czekanowski metric, topic modeling, reviewer recommendation, generalized Pareto distribution

1. Introduction

Topic modeling is a machine learning technology for summarization of huge volumes of information [1]. A popularity of topic modeling is increasing drastically over the last decade. Annual number of research papers is doubling every 3-4 years.

Estimating the level of similarity between two objects is a common task in topic modeling. For estimating the level of similarity it is necessary to describe each object as a vector of attributes. If objects are defined in a metric space, then each attribute is defined in a numerical scale. For example, in Fisher's Iris dataset, each flower is described with four attributes, namely the petal width, the petal length, the sepal width, and the sepal length. Object's attributes can be categorical as well, in this case they are defined as a distribution of membership degrees over categories. Such a categorical

The Sixth International Workshop on Computer Modeling and Intelligent Systems (CMIS-2023), May 3, 2023, Zaporizhzhia, Ukraine
EMAIL: s.shtovba@donnu.edu.ua (S. Shtovba); petrychko.myckola@gmail.com (M. Petrychko); olenashtovba@vntu.edu.ua (O. Shtovba)
ORCID: 0000-0003-1302-4899 (S. Shtovba); 0000-0001-6836-7843 (M. Petrychko); 0000-0003-1418-4907 (O. Shtovba)

CEUR Workshop Proceedings (CEUR-WS.org) Proceedings

representation of objects is often used in topic modeling problems. For the Fisher's Iris dataset, the result of flower classification can be represented as a categorical distribution, for example, some flower is classified as *Iris Setosa* with a membership degree of 0.7, as *Iris Virginica* with a membership degree of 0.1, and as *Iris Versicolor* with a membership degree of 0.2.

Depending on the type of object representation different metrics of similarity are used. For objects defined in a metric space, similarity is defined as a quantity that is inversed to the distance between two points. Coordinates of each point are numerical values of the object's attributes. The smaller the distance between the analyzed objects the more similar they are. In paper [2] around 50 different metrics are analyzed, the most popular among them are particular cases of Minkowski's metric: Euclidean distance, City Block distance, and Chebyshev metric. The cosine similarity metric is often used as well. It calculates the cosine angle between two vectors that start at the origin and end at the analyzed objects.

In categorical space, the similarity between two objects is usually defined as similarity superposition of objects over each category. Most often it is the sum of similarities over each separated category. For this approach, each category is considered independently and in isolation from others. There is also an inversed approach when the object divergence is first found for each category and then aggregates to find the overall similarity. One of the popular variants of such a metric is proposed in [3] to find the similarity of fuzzy sets. The authors define the divergence of objects as a module of membership degree difference. All metrics from the survey paper [2] and other relevant publications, for example, [4, 5, 6, 7], do not consider the kinship between categories. But for some practical problems, the categories are akin. This leads to the fact that it is better calculate the similarity between objects not only directly as the similarity between equivalent categories, but also consider indirect cross-similarity through akin categories. Developing of such a metric that additionally considers the similarity of objects through akin categories is, therefore, the purpose of this paper.

2. Objects representation in the space of akin categories

We present here an example of problems in which objects are described in the space of akin categories.

Let us consider task of finding similar researchers, for example, for reviewer recommendation system. Based on research papers, each researcher can be categorized into a few research specialties (research fields) according to some research classification system. For example, researcher A is categorized to *System Analysis* with a membership degree of 0.4 and to *Information Systems and Information Technologies* with a membership degree of 0.6. Researcher B is categorized to *System Analysis* with a membership degree of 0.7 and to *Computer science* with a membership degree of 0.2. Researcher C is categorized to *System Analysis* with a membership degree of 0.4 and to *Marketing* with a membership degree of 0.6.

Finding the similarity between each pair of these researchers using known metrics will only take into account their membership degrees to *System Analysis*. Membership degrees in other categories are not taken into account because they are different for each researcher. The similarity between researchers A and B is defined only by their membership degrees to *System Analysis* category, which equals 0.4 and 0.7, respectively. If the similarity is defined as the common part of membership degrees using *min* operation, then between researchers A and B, it equals to $Fit(A, B) = \min(0.4, 0.7) = 0.4$. By the same reasoning, the similarity between researchers A and C equals to $Fit(A, C) = \min(0.4, 0.4) = 0.4$, and between researchers B and C equals to $Fit(B, C) = \min(0.7, 0.4) = 0.4$. This shows that the similarity of all pairs of researchers is the same. But the research domains are such that *Information Systems and Information Technologies* is significantly closer to *Computer Science* than to *Marketing*. Also, *Computer Science* is significantly closer to *System Analysis* than to *Marketing*. Therefore, the similarity between researchers A and B has to be stronger than the similarity between researchers A and C, or between researchers B and C. But the well-known similarity metrics do not take into account the kinship of categories; therefore, it is impossible to take into account such peculiarities using them.

3. The proposed metric for akin categories

Let us denote the number of categories as *m*. Then, objects *X* and *Y*, the similarity of which is to be estimated, we represent with the following membership distributions to categories: $(\mu_1(X), \mu_2(X), ..., \mu_m(X))$ and $(\mu_1(Y), \mu_2(Y), ..., \mu_m(Y))$. The distributions are considered to satisfy the following conditions:

$$\mu_i(X) \in [0; 1],\ i = \overline{1, m};$$

$$\mu_i(Y) \in [0; 1],\ i = \overline{1, m};$$

$$\sum_{i=\overline{1,m}} \mu_i(X) = 1;$$

$$\sum_{i=\overline{1,m}} \mu_i(Y) = 1.$$

The problem is that the level of similarity is to be calculated for objects *X* and *Y*. The domain research's peculiarity lies in the fact that some categories are akin. Therefore, it is necessary to consider not only the similarity of identical categories but also the similarity of akin categories. Below we propose a metric that accounts for the semantic kinship of categories.

The similarity between two objects *X* and *Y* is proposed to be calculated in the following way:

$$Fit(X, Y) = F(X, Y) + \Delta F(X, Y), \tag{1}$$

where $F(X, Y)$ denotes addend that assesses the direct similarity between the objects *X* and *Y* over identical categories;

$\Delta F(X, Y)$ denotes addend that considers the similarity of objects *X* and *Y* over akin categories.

We calculate the first addend of the formula (1) by the simplified version of Chekanowski metric for the case when membership degrees are in the $[0; 1]$ and their sum equals 1. The resulting form of the first addend in (1) is as follows:

$$F(X, Y) = \sum_{i=\overline{1,m}} \min(\mu_i(X), \mu_i(Y)) \tag{2}$$

where $\mu_i(X)$ denotes membership degree of object X to the *i*-th category, $i = \overline{1, m}$;

$\mu_i(Y)$ denotes membership degree of object Y to the *i*-th category, $i = \overline{1, m}$.

The formula (2) can be interpreted as a sum of membership degrees of intersection of fuzzy sets X and Y. In formula (2), it is implied that the overall similarity of two objects is a sum of their similarities by each category. The similarity by category is defined as both objects' minimum memberships to the category. Thereby, one of the objects contributes the entire value of the membership degree to a category, and the other one – only a part of it.

After applying the formula (2) we get the following residuals of membership degrees:

$$r_i(X) = \max(0, \mu_i(X) - \mu_i(Y));$$

$$r_i(Y) = \max(0, \mu_i(Y) - \mu_i(X)), i = \overline{1, m}.$$

Let us consider the contribution of residuals to the similarity of two objects using kinship of categories. We assume that the information about categories' pair kinship is known, and denote it with the following binary relation:

$$\mathbf{K} = \left\| k_{ij} \right\|,$$

where $k_{ij} \in [0; 1]$ denotes the kinship coefficient of the *i*-th and *j*-th categories, $i = \overline{1, m}$, $j = \overline{1, m}$.

The categories are more akin, the higher the kinship coefficient. Kinship relation is symmetric and reflexive, therefore, $k_{ij} = k_{ji}$ and $k_{ii} = 1$.

We denote the composition of residuals in the form of a matrix as follows:

$$\mathbf{E} = \left\| e_{ij} \right\|,$$

where $e_{ij} = \min\left(r_i(X), r_j(Y)\right)$, $i = \overline{1, m}$, $j = \overline{1, m}$.

The contribution of residuals to metric (1) using paired kinship of categories is calculated as follows:

$$\Delta F(X, Y) = \sum_{i=\overline{1, m}} \sum_{j=\overline{1, m}} e_{ij} \cdot k_{ij} . \tag{3}$$

Example. Two objects are defined with the following membership degrees to categories $\{A, B, C, D\}$: $X = (0.4 \quad 0.3 \quad 0.2 \quad 0.1)$ and $Y = (0.7 \quad 0.1 \quad 0.1 \quad 0.1)$. Kinship of the categories are described by the following matrix: $\mathbf{K} = \begin{Vmatrix} 1.0 & 0.5 & 0.0 & 0.0 \\ 0.5 & 1.0 & 0.1 & 0.0 \\ 0.0 & 0.1 & 1.0 & 0.3 \\ 0.0 & 0.0 & 0.3 & 1.0 \end{Vmatrix}$.

Let us calculate the similarity between objects X and Y using the proposed metric (1).

To calculate the first addend of the similarity metric (1) let us find the intersection of two distributions. The intersection is shown in Figure 1 with a hatch.

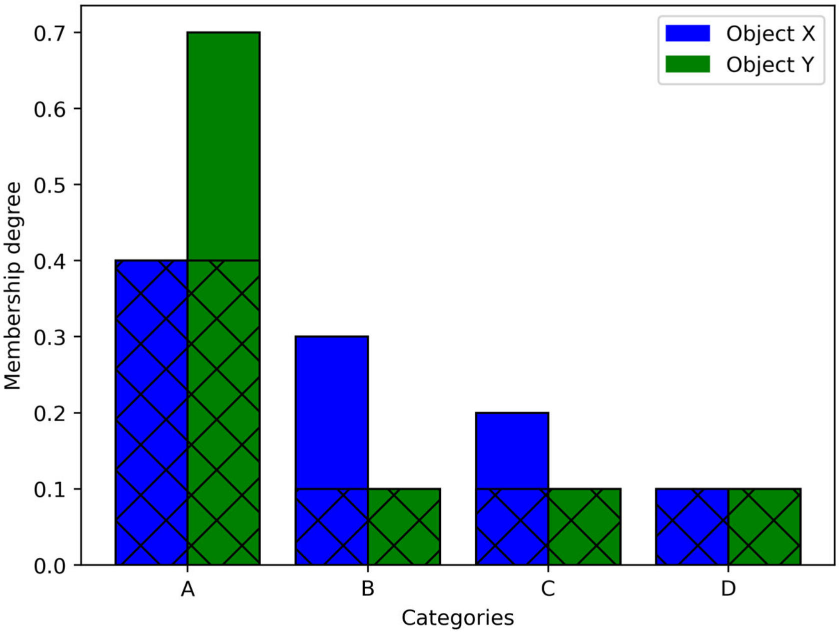

Figure 1: Intersection of two categorical distributions for calculating $F(X, Y)$

The numerical value of the first addend is as follows:

$$F(X, Y) = \min(0.4, 0.7) + \min(0.3, 0.1) + \min(0.2, 0.1) + \min(0.1, 0.1) = 0.4 + 0.1 + 0.1 + 0.1 = 0.7 .$$

Residuals after the intersection are:

$e(X) = (0 \quad 0.2 \quad 0.1 \quad 0)$;

$e(Y) = (0.3 \quad 0 \quad 0 \quad 0)$.

The residuals composition is calculated as $\mathbf{E} = \begin{Vmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{Vmatrix}$.

By having performed element-wise product of matrices $\mathbf{E}$ and $\mathbf{K}$, we get the following matrix of contributions through akin categories: $\begin{Vmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{Vmatrix}$. From this matrix it is observed, that the contribution of kinship of the second and first category equals to 0.1. The kinship contribution of other categories is zero.

The summed contribution of akin categories equals to: $\Delta F(X,Y)=0.1$. The resulting similarity value of objects X and Y using formula (1) equals $F(X,Y)=0.7+0.1=0.8$.

4. Computational Experiments

In the above example, taking into account the kinship of categories increased the similarity metric by 0.1, which is 14% of the initial value obtained by the Chekanowski metric. Let us conduct computational experiments to establish how sensitive the proposed metric is to taking into account the kinship of categories.

We perform 18 series of experiments. All the experiments within a series is carried out with the same number of categories. The number of categories (m) from series to series increases from 2 to 70 with a step of 4. In each series, the similarity of 5000 pairs of objects X and Y is calculated. Attributes of objects X and Y as well as the kinship matrix of categories are randomly generated using the generalized Pareto law [8]:

$$f(x)=\frac{1}{\sigma}\left(1+\frac{k(x-\theta)}{\sigma}\right)^{\left(-\frac{1}{k}-1\right)}.$$

The choice of this law is due to the fact that in topic modeling, the categories' similarity distribution often is similar to the Pareto distribution (see, for example, [9, 10]). For each experiment, we firstly randomly generate parameters of the distributions. The shape parameter k is generated from the range $[0.15; 0.5]$, the scale parameter σ is generated from the range [0.001; 0.01]; we set the bias as zero: $\theta=0$. In each experiment, using synthesized distributions, we randomly generate the coordinates of the vectors X and Y, as well as kinship matrix $\mathbf{K}$. In matrix $\mathbf{K}$ the maximum value of the elements is limited to 0.4; vectors X and Y are normalized by 1. Then, we calculate the similarity between X and Y using the proposed metric (1).

The results of the experiments in the form of box-plot diagrams are shown in Figure 2. It can be seen from the figure that the similarity values are in the range $[0;1]$. As the number of categories increase, the spread of results decreases. Median of distributions for $m>6$ does not depend on the number of categories.

On Figure 3 box-plot diagrams of the distribution of the term $\Delta F(X,Y)$ are shown, which takes into account the similarity of objects X and Y over akin categories. The median of this value increases from 0.001 to 0.066 with an increase in the number of categories. The spread of the distributions also increases, while the number of outliers decreases from 550 to 7. The increase in the median probably indicates that as the number of categories increases, a large number of weak ties between akin categories make a significant contribution to $\Delta F(X,Y)$. As the number of categories increase, the number of pairs of akin categories increases quadratically. The contribution of many pairs of akin categories is tiny. It looks as a noise. But the sum of huge number of noise contributions turns out to be large. The decrease of the number of outliers is also a negative factor. The need for a new metric is primarily due to the need to identify specific cases with strong cross-over effects due to akin categories. Also, the decrease in the number of outliers indicates that noise kinships make it difficult to detect such pairs of objects.

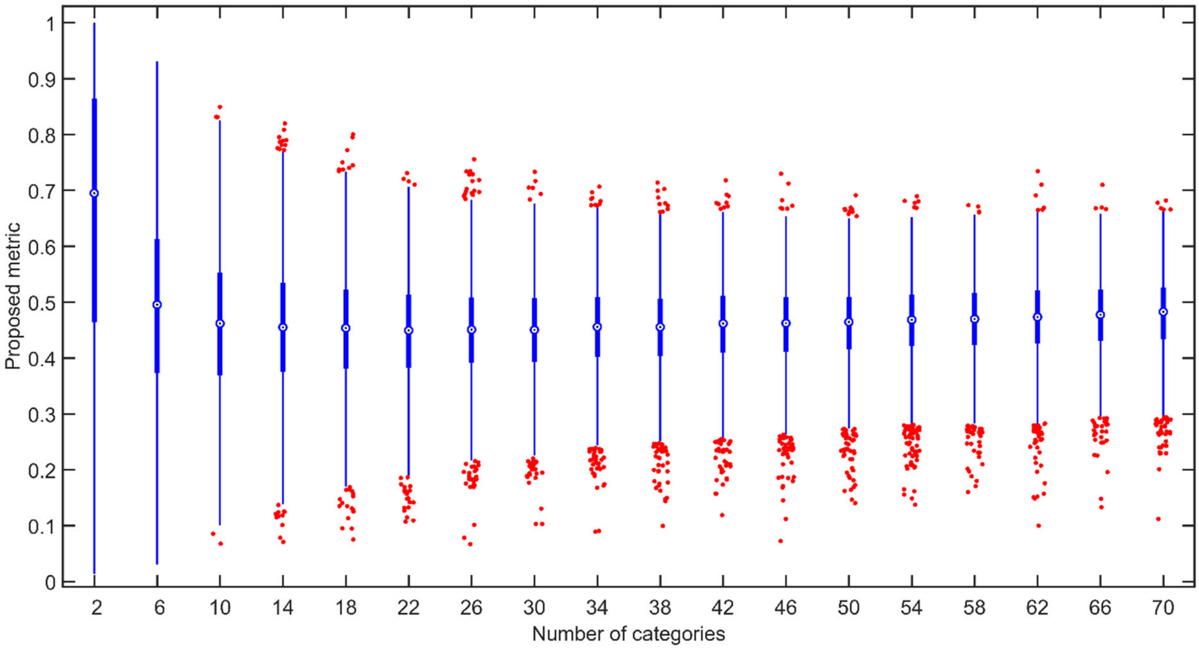

Figure 2: Similarity distributions according to the proposed metric

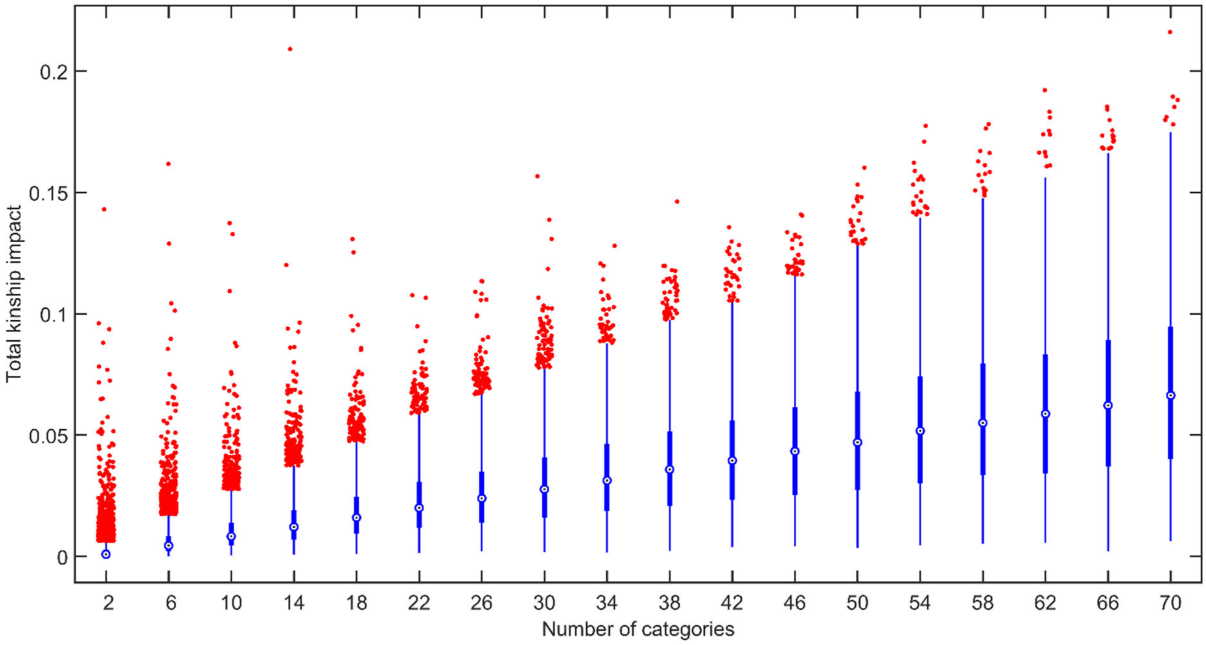

Figure 3: Distribution of the akin categories' contribution

Let us reduce the noise contribution of categories with weak kinship. We assume the kinship to be noisy if the kinship coefficient is less than 0.05. Box-plot diagrams of the noise kinship distribution's contribution are shown in Figure 4. It shows that the median noise contribution increases linearly and reaches a value of 0.056 for 70 categories. The maximum value of the contribution from noise kinship is fixed at the level of 0.122.

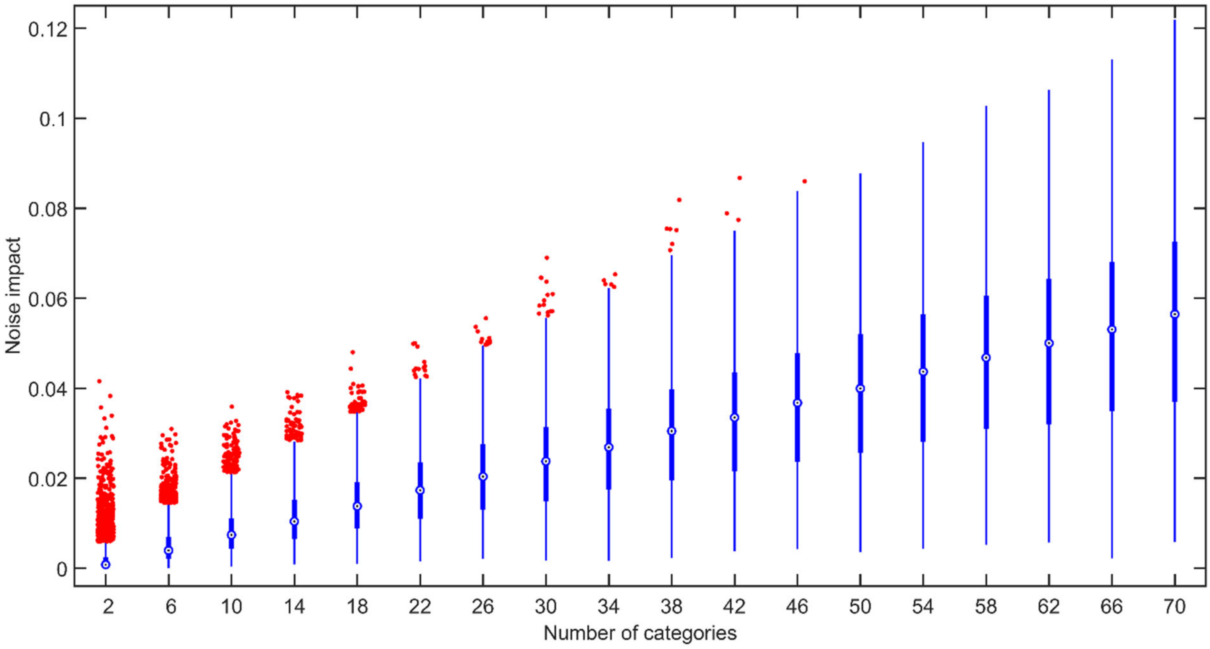

Figure 4: Distributions of noise contribution from consideration of akin categories

As for the relative noise contribution (shown on Figure 5), its median exceeds 10% in the series of experiments for a large number of categories. In each series of experiments, there are numerous cases when the noise contribution exceeds 15%. The noise contribution exceeds 25% in 194 cases out of 90,000 (Figure 6).

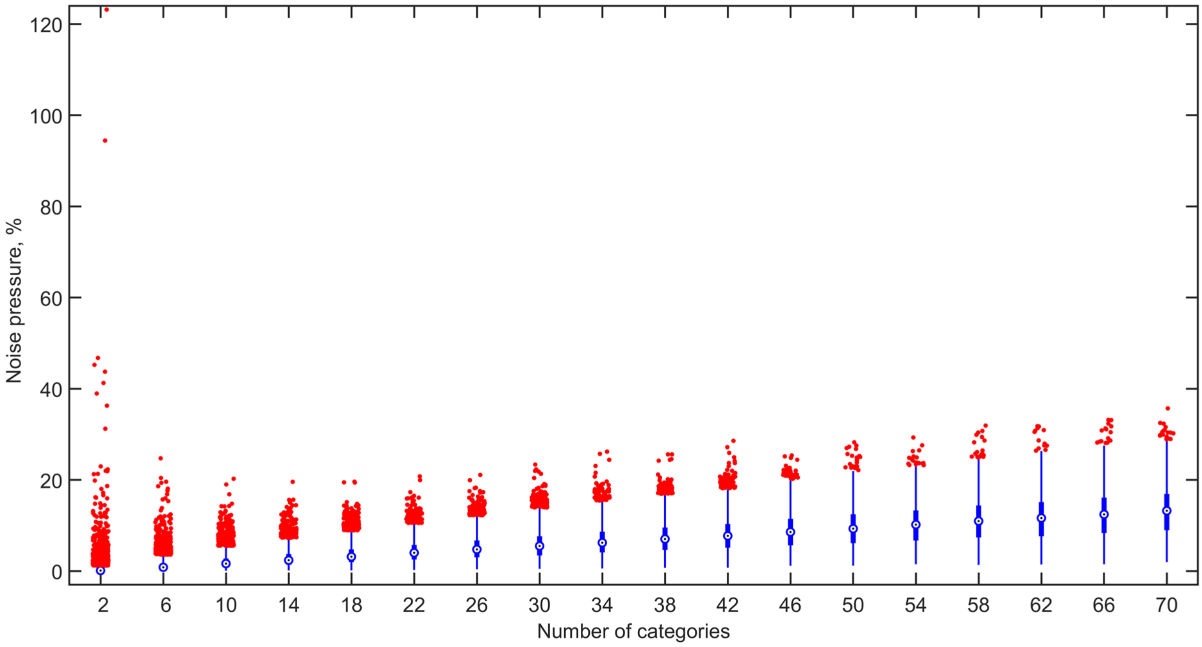

Figure 5: Distributions of the relative value of the noise contribution from akin categories

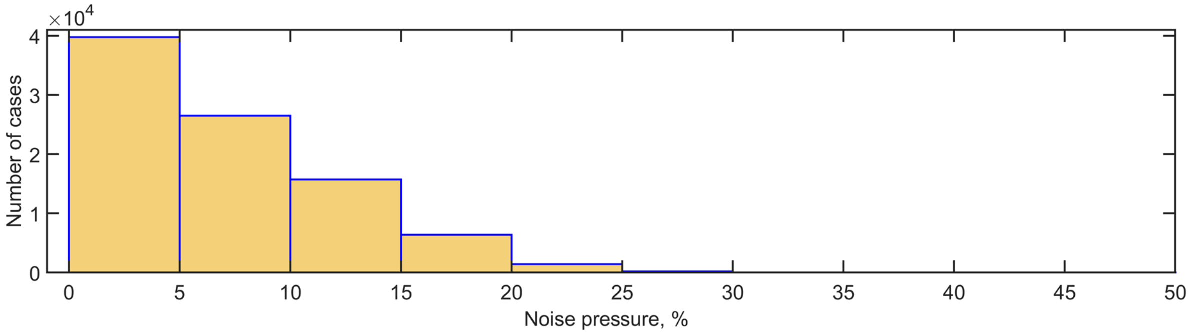

Figure 6: Distribution of noise contribution

After noised component elimination, the box-plot diagrams of the akin categories contribution's distribution are shown in Figure 7. The median of this value increases only from 0 to 0.005 as the number of categories increases. This is 13 times less than without noise elimination. At the same time, a large number of outliers are observed – their number is from 5-20%. This indicates that the metric is able to identify cases of strong interaction across akin categories. The box-plots of the proposed metric after removing the contribution from the noisy kinship of the categories are shown in Figure 8. Median of the refined metric lies in range [0.43; 0.69]. It is still high, especially for cases with large number of categories. Probably, it is caused by the noised memberships to many categories in source distributions in for X and Y. Hence, it is better to eliminate the noised memberships before assessing the similarity of two categorical distributions. It is reasonable to do so for such topical modeling problems when it is known that any object may belong just to a small number of categories.

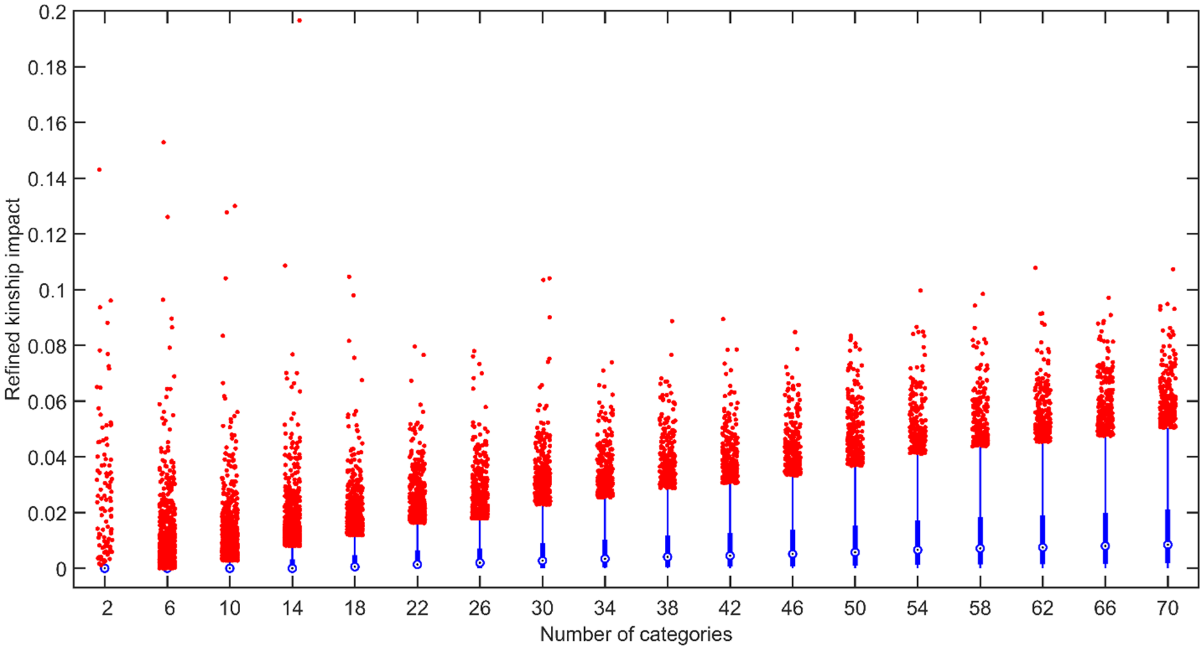

Figure 7: Distributions of the akin categories contribution after noise filtering

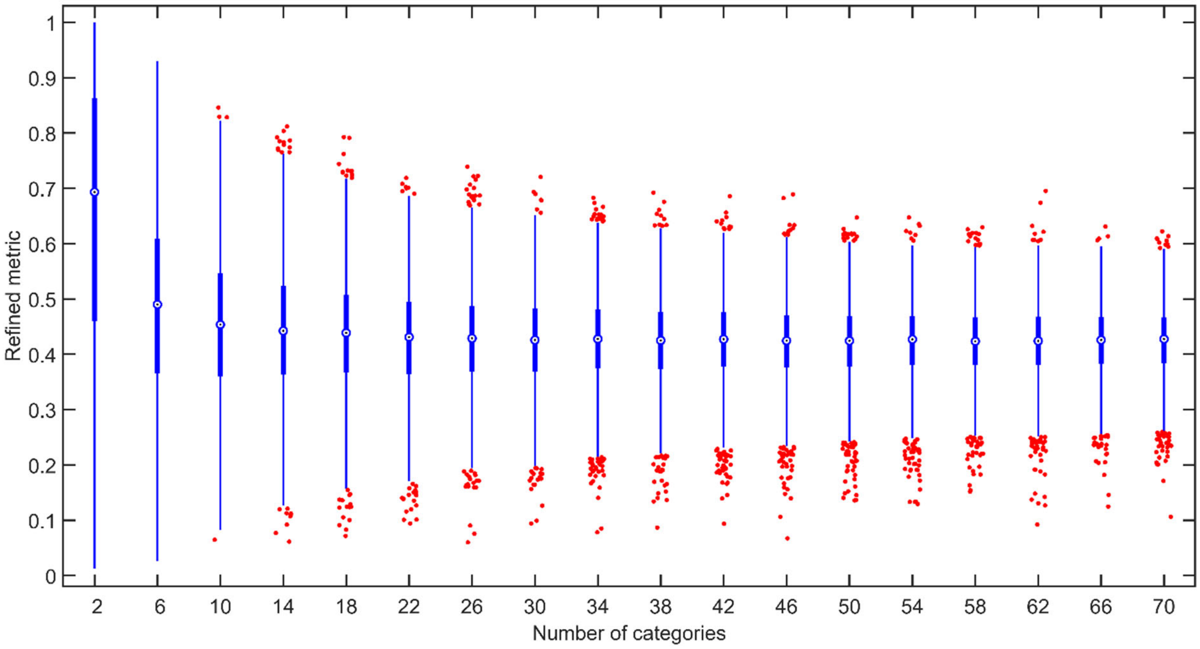

Figure 8: Similarity distributions according to the proposed metric with noise filtering

5. Conclusions

A new similarity metric of categorical distributions is proposed. A feature of the proposed metric consists of taking into account the kinship of categories. The proposed metric has two components. The first component is defined as Chekanowski metric. It calculates the direct similarity of distributions by categories as the sum of the intersection of distributions' membership degrees of two objects. The second component of the metric takes into account the similarity of objects through akin categories. It is assumed that the kinship coefficients of each pair of categories are known.

Computational experiments have shown that in the case of Pareto distributions of objects memberships to categories and Pareto distributions of kinship coefficients of categories, the proposed metric takes values from the interval [0; 1]. It was established that with an increase in the number of categories, the contribution to the proposed metric of the term, which takes into account the kinship of the categories, strongly increases. This is due to the fact that the number of weak ties between akin categories increases quadratically, each of which adds some contribution to the value of the metric. And although the contribution from many pairs of akin categories is tiny, roughly speaking - noisy, but the sum of the contributions turns out to be large.

To eliminate the noise impact, we suggest ignoring the noise kinship of the categories. A simple filter with kinship coefficient threshold at the level of 0.05 eliminated this drawback. After such noise filtering, the distributions of the second component of the metric, which takes into account the kinship of the categories, have a significant number of outliers. A significant number of outliers is observed in all the series of experiments, both with a small number of categories and with a large one. This fact indicates that the proposed metric makes it easy to identify pairs of objects whose similarity is largely determined by membership to akin categories.

The proposed metric can be used for topic modeling tasks, in which, when evaluating the similarity of two objects, it is necessary to take into account their membership to akin categories. Such tasks can be the selection of reviewers of research papers and theses, the analysis of the similarity of text documents, the identification of poses of people in a video stream, the clustering of species distribution, the formation of recommendations in online shops, etc.

6. References

[1] A. Abdelrazek, E. Yomna, G. Ema, M. Walaa, H. Ahmed, Topic modeling algorithms and applications: A survey, Information Systems 112 (2023) 102131. doi: 10.1016/j.is.2022.102131.

[2] N. Sebe, J. Yu, Q. Tian, J. Amores, A new study on distance metrics as similarity measurement, in: Proceedings of IEEE International Conference on Multimedia and Expo, Toronto, 2006, pp. 533–536. doi: 10.1109/ICME.2006.262443.

[3] W.-J. Wang, New similarity measures on fuzzy sets and on elements, Fuzzy Sets and Systems 85 (1997) 305-309. doi: 10.1016/0165-0114(95)00365-7.

[4] S.-H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, International Journal of Mathematical Models and Methods in Applied Sciences 1 (2007) 300-307.

[5] J. Yu, Q. Tian, J. Amores, N. Sebe, Toward robust distance metric analysis for similarity estimation, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, 2006, pp. 316–322, doi: 10.1109/CVPR.2006.310.

[6] K. Louisa, A. Colubi, New metrics and tests for subject prevalence in documents based on topic modeling, International Journal of Approximate Reasoning 157 (2023) 49-69. doi: 10.1016/j.ijar.2023.02.009.

[7] D. J. Weller-Fahy, B. J. Borghetti, A. A. Sodemann, A survey of distance and similarity measures used within network intrusion anomaly detection, IEEE Communications Surveys & Tutorials, 17 (2015) 70-91. doi: 10.1109/COMST.2014.2336610.

[8] H. T. Davis, M. L. Feldstein, The generalized Pareto law as a model for progressively censored survival data, Biometrika, 66 (1979) 299-306. doi: 10.1093/biomet/66.2.299.

[9] S. Shtovba, M. Petrychko, An algorithm for topic modeling of researchers taking into account their interests in Google Scholar profiles, in: Proceedings of the Fourth International Workshop on Computer Modeling and Intelligent Systems, Zaporizhzhia (2021), CEUR Workshop Proceedings, 2864 (2021) 299-311. URL: https://ceur-ws.org/Vol-2864/paper26.pdf.

[10] S. Shtovba, M. Petrychko, Jaccard index-based assessing the similarity of research fields in Dimensions, in: Proceedings of the First International Workshop on Digital Content & Smart Multimedia, Lviv (2019), CEUR Workshop Proceedings, 2533 (2019) 117-128. URL: https://ceur-ws.org/Vol-2533/paper11.pdf.