Identification of information objects from various sources for forming an information resource of dynamic objects monitoring system

Volodymyr Yuzefovych[1,*,†], Yevheniia Tsybulska[1,†] and Nikolai Stoianov[2,†]

[1] *Institute for Information Recording of NAS of Ukraine, Shpak Str. 2 03113 Kyiv, Ukraine*
[2] *Bulgarian Defense Institute, "Professor Tzvetan Lazarov" Blvd. 2 1592 Sofia, Bulgaria*

Abstract

An approach to identifying information objects (IOs), data about which is received by the monitoring system from independently operating sources, is presented. It considers a situation where data about the same physical object can be entered multiple times into an information resource, as about different objects. At the same time, the values of such IOs features do not completely coincide, since the data sources introduce some operation errors. The proposed approach for object identification is based on a new proximity (similarity) measure of information objects, which takes into account the existing probabilistic uncertainty regarding the values of quantitative features and the uncertainty of the possibility type for qualitative features.

Keywords

identification, information object, monitoring system, proximity (distance) measure, quantitative features, qualitative features, probability distribution law, fuzzy set

1. Introduction

The majority of information systems have a functionally, and sometimes organizationally dedicated monitoring subsystem, the aim of which is to extract (obtain) data about the external environment and the state of the system as a whole. This work focuses on one of the problematic tasks of forming an information resource that is populated with data from monitoring the external environment, shaped by the actions of dynamic objects in the surrounding space. As the number of monitored objects increases, along with expanding the means of monitoring, or when multiple monitoring systems are integrated into a higher-level system, there is an increased probability that data about the same object can independently be entered into the common information resource of the monitoring system. This situation is typical for cases when the monitoring system (or subsystem, if the monitoring system is hierarchical) simultaneously analyzes objects in the surrounding space in overlapping spatial areas. Figure 1 schematically shows several overlapping areas monitored by different data sources (DS), and, accordingly, a few sources can observe the same physical (real) objects. The set of features of such objects, determined by the data source, will be called an information object (IO). Essentially, the IO, formed by a suitable data source, is an information representation of a real object in the system's information resource in the form of a finite set of features and their values.

ITS-2024: Information Technologies and Security, December 19, 2024, Kyiv, Ukraine
* Corresponding author.
† These authors contributed equally.
uzefv71@gmail.com (V. Yuzefovych); evts68@gmail.com (Ye. Tsybulska); n.stoianov@di.mod.bg (N. Stoianov)
0000-0002-6336-9548 (V. Yuzefovych); 0000-0003-3342-4507 (Ye. Tsybulska); 0000-0002-4953-4172 (N. Stoianov)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4068/paper7.pdf
CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

In the case described above, there is a need to solve the problem of identifying IOs, that is, attributing them to a single physical object, with the subsequent unification (aggregation) of the characteristics of such (identified) IOs. In the simplest case, the values of IO features that refer to the same real object should completely coincide, even if they are determined by different data sources. Such a coincidence would make it quite easy to classify a certain number of IOs as those describing the same physical object, provided that in the area of the monitoring system operation, there are no different objects with completely identical observational characteristics. Regarding the latter condition, we would additionally note that if the monitoring system allows the formation of completely identical IOs without additional (marking) data that refer to different physical objects, this indicates its insensitivity to certain differences in the external environment, which can be seen as a flaw.

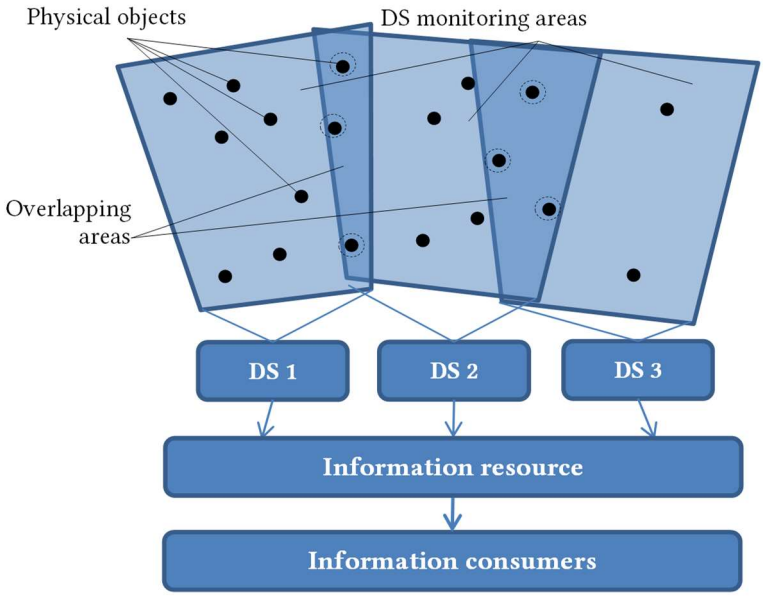

Figure 1: Observation zones with overlapping areas related to different data sources of the monitoring system.

In our case, given that among the key object features in the surrounding space are the static or time-varying coordinates of their location, we can reasonably assume that the condition above is satisfied. However, even if the feature values for different IOs referring to the same real object are equal, there would be a need to identify such IOs due to the fact that not all sources are able to determine the full list of features. Therefore, even in such a simple case, to solve the problem of identifying the IO, it would be necessary to introduce and analyze a certain IO proximity measure, such as Rao coefficients [1]. Additionally, the situation is further complicated by the fact that the monitoring system tools determine any features of objects with errors, which makes the chance of an exact match of even some feature values random and unlikely. This prompts a transition from searching for an exact equality of IO features to analyzing the proximity of IOs across the full set of features available for observation, taking into account the errors in their determination. Therefore, solving the problem of identifying IOs requires solving the problem of formally defining the proximity measure between the feature values and between IOs as a whole.

2. Related works

The problem of determining which observations or descriptions correspond to the same object (object instance) exists in various fields. These can be image recognition and analysis, natural

language generation and processing, text processing, integration of information resources, etc. For example, to track a moving object in computer vision systems, it is necessary to identify whether two shapes in different frames of a video stream are actually the same object. Creating a distributed information system with a common information space involves merging separate databases. As a result, it is necessary to determine which records belong to the same entity and solve the integration task when incomplete matching of their attributes. When preparing reference lists in articles, it is necessary to find which citations refer to the same papers to avoid duplication. In natural language processing, a key task is to determine which phrases (word combinations) refer to the same entity. An object identification for databases data merging/cleansing, record linking and duplicate removal was first formulated as a separate problem by Newcombe et al. [2] and solved by Fellegi and Sunter [3], whose method became the basis for further developments.

There are currently numerous developments in this area, including Wang and Ji [4], Nagarajan and Grauman [5], Singla and Domingos [6], Cohen and Richman [7], Sarawagi and Bhamidipaty [8], Pasula et al. [9], etc. Most existing approaches are the development and improvement of the original Fellegi and Sunter model, in which object identification is considered as a classification problem. That is, it defines a vector of similarity scores between the attributes of two observations, based on which the classification into "Match" or "Not Match" is performed. Each candidate pair is assessed separately, and a matching decision is formed. Then a transitive closure is constructed to eliminate inconsistencies. At the same time, the development of new methods is ongoing in two directions: improving measures and metrics for assessing the proximity of research objects and improving methods for group processing of multiple comparisons.

3. Problem formulation

The set of information objects that will be used by an information system is first defined during its design phase. Later, over the life of the system, this set is supplemented and edited in accordance with users' information needs. IOs can describe:

- Single entities (material objects, persons)
- Abstract entities (concepts)
- Group entities (homogeneous or heterogeneous)
- Static composite entities (situation description)
- Dynamic composite entities (processes).

An information object can be formally specified by a tuple

$$IO = \langle m, S, D, K_S \rangle, \quad (1)$$

where m – IO unique identificator

$S = \{s_j | j = 1 \dots l\}$. – a set of IO features (attributes)

$D = \{d_i | i = 1 \dots p\}$. – a set of constraints on the object attribute values

K_S:$S \to D$. – a mapping to set constraints for each attribute.

In general, information objects are interconnected and interdependent. Let us define the set of relations between IOs as follows:

$$R = \langle R_1, R_2, R_3 \rangle, \quad (2)$$

where R_1 – inheritance relationship ("class-subclass") $R_1(IO_1, IO_2)$, where IO_1 is an upper class for IO_2

R_2 – aggregation relationship ("included in") $R_2\left(IO_1, \{IO_j\}\right)$, where IO_1 features are included in the set of features of information objects set $\{IO_j, j = 1 \dots l\}$

R_3 – association relationship (semantic relations).

The total set of possible IO features can be divided into features of a quantitative and qualitative nature. The values of quantitative features are determined using certain measuring instruments and, accordingly, are characterized by uncertainty of the type "probability", since any means of measurement has limited, albeit defined, accuracy. The values of qualitative features are determined by the active participation of a person, and therefore contain a subjective component, which today is usually described by uncertainty of the type "possibility" [10, 11, 15]. It is obvious that in this case, the proximity (or distance) measure between the IOs described by a set of features should be a combined quantitative-qualitative one. Few such proximity measures are known. In particular, these include the Voronin approximation proximity measure [12], the Mirkin similarity measure [13], and the most "physically transparent" Zhuravlev measure [1, 14], which for two IOs i and j is determined as follows

$$\rho_{ij} = \sum_{l=1}^{L} \alpha_{ij}^{l}, \tag{3}$$

where l – an index of an object feature ($l = \overline{1,L}$), L – total number of features;

$\alpha_{ij}^{l} = \begin{cases} 1, & if\ |x_i^l - x_j^l| \le \varepsilon^l, \\ 0, & in\ other\ cases; \end{cases}$ (for quantitative features)

$\alpha_{ij}^{l} = \begin{cases} 1, & if\ the\ feature\ is\ present\ and\ its\ value\ matches, \\ 0, & if\ the\ feature\ is\ absent\ or\ its\ value\ doesn't\ match; \end{cases}$ (for qualitative features)

ε^l - quantitative proximity threshold for the l-th feature.

As can be seen from expression (3), Zhuravlev's measure for quantitative features allows the possibility of some slight difference in their values, within which it is assumed that the features still coincide. That is, a threshold analysis of the proximity of such feature values is used. It should be noted that determining the threshold value ε^l when solving a specific problem is up to the researcher. Such a possibility is not provided for qualitative features, and only complete coincidence/difference of their values is allowed.

Let us consider the acceptability of the approach to accounting for the possible difference in the quantitative feature values $x_i^l - x_j^l = r_{ij}$ through defining some admissible value of it - ε^l. It is known that measurement errors of various quantities are distributed according to a certain law of probability distribution. This distribution is characterized by the mathematical expectation of the error (the average error value, which is equal to zero in the absence of a systematic component), the standard deviation from its mathematical expectation (mean square error - MSE), and other higher-order moments.

It is generally accepted that measurement errors are most often distributed according to a normal law. Let us assume this statement is true for all our cases. Then the measurement error distribution for quantitative features is completely determined by the first order moment (mathematical expectation) and the second order moment - the dispersion, or the standard deviation of the random variable. It is quite obvious that when the linear distance r_{ij} between the measured feature values decreases, the probability that the obtained measured values actually refer to the same true value will increase nonlinearly, in accordance with the distribution laws of their measurement errors with two different means (which are generally characterized by different MSE). In addition, if the measured feature values coincide, but the values were obtained with an error, then such a coincidence cannot be guaranteed to mean a coincidence of the true values. Therefore, using the constant ε^l is a fairly rough approximation to reality.

Therefore, this work aims to construct a proximity (similarity) measure to compare information objects, which takes into account the possibility of errors for both types of IOs features – quantitative and qualitative.

4. Determining the proximity measure for quantitative features

This paper proposes an alternative method for calculating the degree of proximity between the quantitative feature values, which takes into account the probabilistic nature of the process of their definition by different sources. The currently known proximity measures between two measured (quantitative) feature values require calculating the probability that the true value of the feature for both measurements is actually the same value. Calculating such a probability requires full knowledge of the probability distribution laws for measured values (not just the measurement errors) and, therefore, the true value of the measured quantity. In our problem, this value is unknown. Taking into account the above, it is proposed to formalize the quantitative feature values by the normal distribution law of their determination errors, where the feature value obtained from the source is considered as the mathematical expectation. Data on the standard measurement error is expected to be obtained from the data source or determined based on its characteristics as a measurement means. Then the coincidence of the feature values obtained from two sources can be considered dependent on the probability of finding the true value in the intersection area of the two distribution laws, which can be calculated based on the Laplace function, the probability multiplication theorem for independent events and the well-known "three sigmas" rule. The expression for calculating the probability of finding a random variable x in the interval (c, d) with its normal distribution has the following form

$$P(c \le x \le d) = \Phi\left(\frac{d-m}{\sigma}\right) - \Phi\left(\frac{c-m}{\sigma}\right), \quad (4)$$

where $\Phi(\cdot)$ – Laplace function

m – mathematical expectation of a random variable.

Given the independence of the two measurements, the probability that the measured quantity is actually within the range of values ($c \le x \le d$): $P_{ij} = P_{i(c,d)} \cdot P_{j(c,d)}$, where $P_{i(c,d)}$ and $P_{j(c,d)}$ - the probabilities that the feature values for each measurement are within the interval (c, d).

Consider the example shown in Figure 2. Let the value of the attribute x be measured by two different sources. The obtained measurement results are: $X_1 = 12$ and $X_2 = 18$ units. In this case, the MSE for measurement errors for each source are: $\sigma_{X_1} = 3$, $\sigma_{X_2} = 2$, and $r_{X_1X_2} = 18 - 12 = 6$.

The boundaries of the overlapping regions of the probable value's areas (which determine the specified probability) for two variables - $\delta_{X_1X_2}$ (taking into account the "three sigmas" rule) are determined as the difference between the smaller value from the pair $m_{X_1} + 3\sigma_{X_1}$ and $m_{X_2} + 3\sigma_{X_2}$ and the larger value from the pair $m_{X_1} - 3\sigma_{X_1}$ and $m_{X_2} - 3\sigma_{X_2}$, where $m_{X_1} = X_1$ and $m_{X_2} = X_2$. For the case shown in Figure 2: $m_{X_1} + 3\sigma_{X_1} = 21$; $m_{X_2} + 3\sigma_{X_2} = 24$; $m_{X_1} - 3\sigma_{X_1} = 3$; $m_{X_2} - 3\sigma_{X_2} = 12$, so the interval $\delta_{X_1X_2} = (12{,}21)$.

Next, we calculate the probability that the true value of each measurement is in the range $\delta_{X_1X_2}$. For the given example, we get

$$P_{X1}(12 \le x \le 21) = \Phi\left(\frac{21-12}{3}\right) - \Phi\left(\frac{12-12}{3}\right) = 0{,}49865,$$

$$P_{X2}(12 \le x \le 21) = \Phi\left(\frac{21-18}{2}\right) - \Phi\left(\frac{12-18}{2}\right) = 0{,}43319 + 0{,}49865 = 0{,}93184.$$

Finally, for a given case, the probability that the measured quantity is within the common range of measurement values from two data sources is: $P_{X_1X_2} = 0{,}49865 \cdot 0{,}93184 \approx 0{,}46$.

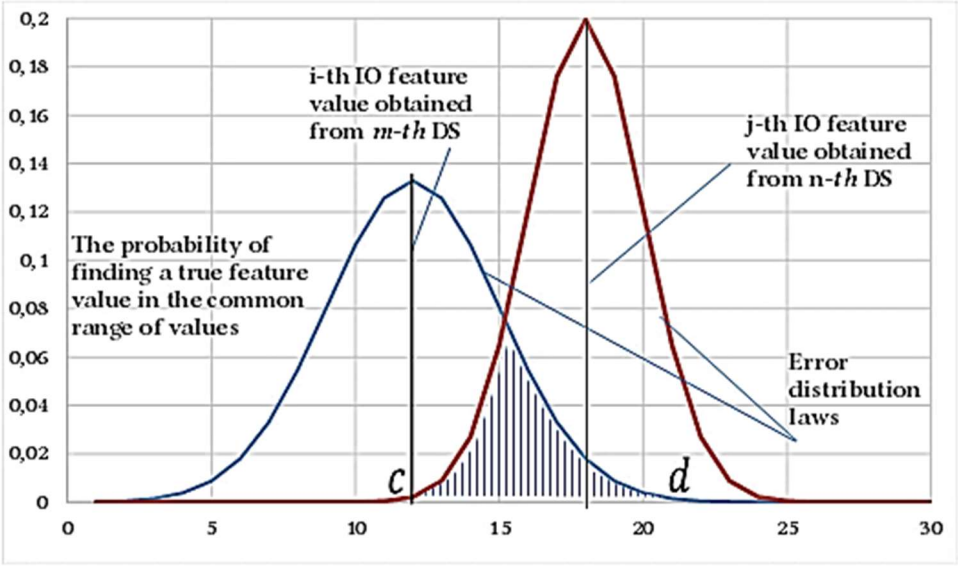

Figure 2: Error distribution laws for measured feature values ($X_1 = 12$ and $X_2 = 18$), obtained from two data sources.

We will perform similar calculations for other measured values of the same feature $X_1 = 14$ and $X_2 = 17$ units (other parameters are the same as in the previous case), $r_{X_1X_2} = 17 - 14 = 3$.

For this case: $m_{X_1} + 3\sigma_{X_1} = 23$; $m_{X_2} + 3\sigma_{X_2} = 23$; $m_{X_1} - 3\sigma_{X_1} = 5$; $m_{X_2} - 3\sigma_{X_2} = 11$ and, respectively, $\delta_{X_1X_2} = (11,23)$. Next we calculate

$$P_{X1}(11 \le x \le 23) = \Phi\left(\frac{23-14}{3}\right) - \Phi\left(\frac{11-14}{3}\right) = 0{,}84,$$
$$P_{X2}(11 \le x \le 23) = \Phi\left(\frac{23-17}{2}\right) - \Phi\left(\frac{11-17}{2}\right) = 0{,}9973.$$

Therefore $P_{X_1X_2} = 0{,}8377$.

Let's perform another calculation to analyze the change nature in the proposed proximity measure and set $X_1 = 15$ and $X_2 = 17$ units ($r_{X_1X_2} = 2$). Then: $m_{X_1} + 3\sigma_{X_1} = 24$; $m_{X_2} + 3\sigma_{X_2} = 23$; $m_{X_1} - 3\sigma_{X_1} = 6$; $m_{X_2} - 3\sigma_{X_2} = 11$ and, respectively, $\delta_{X_1X_2} = (11,23)$. So

$$P_{X1}(11 \le x \le 23) = 0{,}905,$$
$$P_{X2}(11 \le x \le 23) = 0{,}9973.$$

And finally, $P_{X_1X_2} = 0{,}9025$.

We also note that if the distribution laws and the measured values completely coincide, the probability calculated by the expression (1): $P_{ij}^{l} \approx 1$. If they do not intersect within 3σ, the value P_{ij}^{l} will be equal to zero.

Comparing the calculation results, it can be stated that as the difference between the two feature value measurements r_{ij} obtained from different data sources decreases, the value of the proximity measure increases. Additionally, the measure value changes nonlinearly in relation to the linear change of r_{ij} in accordance with the distribution laws of measurement errors of the feature value. Testing the proposed measure for compliance with known conditions for its acceptability and validity (non-negativity, symmetry, maximum similarity of an object to itself, and the "triangle inequality" [1]) shows the possibility of non-fulfilment of the last condition while meeting the first three conditions. At the same time, the latter condition is considered additional and optional [1].

The obvious advantage of using probability to calculate a proximity measure for quantitative features is that the probability (and therefore the measure) changes from 0 to 1, i.e., it is

immediately normalized. That is, calculating proximity values for different features does not require their transformation when determining the common proximity for all features.

Thus, the proposed proximity measure for quantitative features can be considered acceptable, especially since it is based on a probability analysis that has a specific physical meaning. Also, this proximity measure is simply transformed into a distance measure by its inversion

$$\rho^l_{\text{K}ij} = 1 - P^l_{ij}. \quad (5)$$

5. Determining the proximity measure for qualitative features

Let us consider the problem of calculating the possibility that two specified values of a certain qualitative feature are actually the same value. At the same time, it is important to remember that a qualitative characteristic can be expressed in numbers and still retain its qualitative character, since this feature character is determined not by the form of its expression (reflection), but by the method of its acquisition.

To determine the proximity measure between the values of qualitative features, it is proposed to use their formalization in the form of fuzzy sets by constructing a triangular membership function for each obtained feature value on a clear set of its possible values (an example is shown in Figure 3). The number of feature values in the support sets is determined by the possibility of other feature values in reality (which can be considered a certain analogue of measurement error for quantitative attributes). Then the distance between the qualitative feature values can be defined as the maximum value of the set, which is the intersection of two fuzzy sets (formalized feature values). So the expression for the proximity measure is

$$M_{G_1G_2} = max\{\mu_{G_1 \wedge G_2}(x)\} = max\{min[\mu_{G_1}(x), \mu_{G_2}(x)]\}, \quad (6)$$

where G_1 and G_2 – clear values of the feature x

$\mu_{G_1}(x)$, $\mu_{G_2}(x)$ – membership functions of fuzzy sets constructed for both values of the qualitative feature.

Figure *3* shows the results of the formalization of two qualitative feature values expressed in the form of a number: $G_1 = 12$ and $G_2 = 18$. The region of clear feature values, which are covered by non-zero values of the membership function, is specified by the maximum possible errors in the feature determination. As a result, the proximity measure is between the obtained values $M_{G_1G_2} =$ 0,67. It is obvious that if the clear feature values approach each other, the value of the proximity measure approaches 1. Otherwise, it approaches zero.

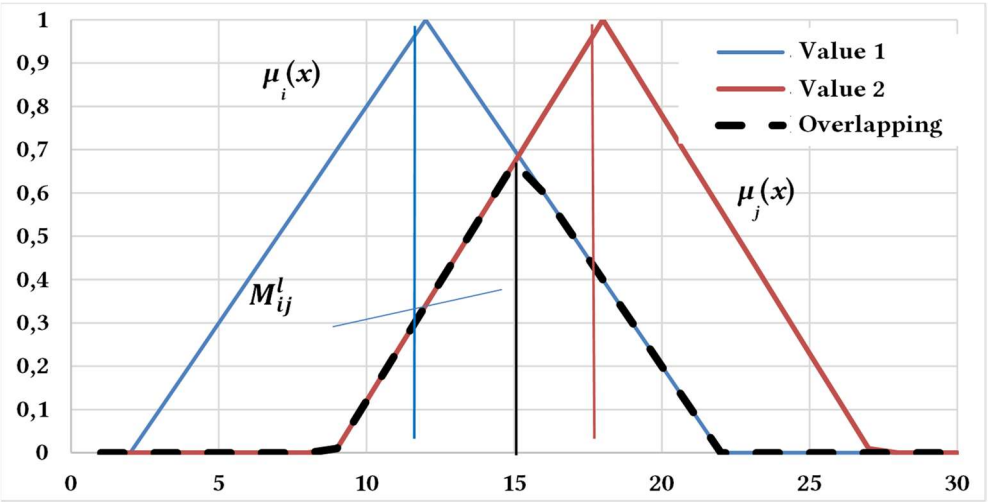

Figure 3: Determination of the proximity measure for qualitative features by formalizing them in the form of fuzzy sets

To obtain a distance measure, by analogy with quantitative features, it is also necessary to invert the obtained value, since as the obtained feature values approach each other, the value M_{ij}^{l} will increase: the proximity measure is

$$\rho_{Zij} = 1 - M_{ij}^{l}. \tag{7}$$

The problem is solved similarly in relation to qualitative IO features, given by linguistic concepts on an ordinal scale. Then the fuzzy set is formed based on the term-set. In this case, to form fuzzy sets, it is necessary to sort the feature values by increasing (strengthening) of the object property that it characterizes.

If the feature is nominal, the membership function will be characterized by one extreme value and some constant value Δ for all other members of the set, which will characterize the possibility of false determination of the feature. Therefore, if the IO feature values obtained from two sources do not coincide, their proximity will be determined by this value Δ, regardless of the feature values themselves.

The proposed proximity measure for qualitative features of fuzzy sets meets all four conditions for the measures' validity.

Note, if, instead of the error distribution laws, we use their triangular approximation for quantitative characteristics, the triangle inequality condition will also be met.

6. Determining the proximity measure (metric) for IOs on the set of their features

To determine the metric (function) of IOs similarity by all features, we can use one of the known additive functions with normalization by the number of values of quantitative and qualitative features and, if necessary, different weight values for each feature, or subsets of quantitative and qualitative features as a whole. It is desirable that the sum of the weight coefficients be equal to 1. Considering the different natures of uncertainty for quantitative and qualitative features, the most acceptable expression for determining the distance measure between IOs on the set of their features is

$$\rho_{Y}{}'_{ij} = w \sum_{l=1}^{L_1} \rho_{\mathrm{K}}{}_{ij}^{l} \Big/ L_1 + (1-w) \sum_{l=L_1+1}^{L} \rho_{Z}{}_{ij}^{l} \Big/ (L - L_1), \tag{8}$$

where L_1 – the number of quantitative features

w – the weight coefficients for quantitative features.

7. Analysis of the IOs proximity metric over time to improve the quality of identification

The final stage of solving the identification problem may be the analysis of the IOs' behavior (actions) over time from the point of view of possible changes in their feature values. This analysis requires setting a criterion by which IOs are considered to be identified as a single object, given the variable distance (proximity) between them over time. Such a criterion will be determined by the characteristics of the chosen distance metric and the specific application problem. For example, if the same physical object is observed by two data sources over several cycles of information update, increasing the probability of correct object identification can be achieved by analyzing the linear trend (α_i) of the change in the distance metric (proximity). For this purpose, the well-known least squares method (LSM) can be applied. Let us denote the distance metric value between two IOs at

time i as ρ_{Y_i}. Then, using the expression obtained based on LSM to calculate the trend line slope and substituting successive time stamps from the interval $i = \overline{1,N}$ into it, we obtain

$$\alpha_i = \left(N \sum_{i=1}^{N} i\, \rho_{Y_i} - \sum_{i=1}^{N} i \sum_{i=1}^{N} \rho_{Y_i} \right) \Big/ \left(N \sum_{i=1}^{N} i^2 - \left(\sum_{i=1}^{N} i \right)^2 \right). \tag{9}$$

Using expressions for the sum of the first N terms of the arithmetic progression of natural numbers

$$\sum_{i=1}^{N} i = \frac{N(N+1)}{2} \tag{10}$$

and the expression for the sum of the squares of the first N terms of the arithmetic progression of natural numbers

$$\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}, \tag{11}$$

we get

$$\theta = \left(N \sum_{i=1}^{N} i\, \rho_{Y_i} - \frac{N(N+1)}{2} \sum_{i=1}^{N} \rho_{Y_i} \right) \Big/ \left(\frac{N^2(N+1)(2N+1)}{6} - \left(\frac{N(N+1)}{2} \right)^2 \right). \tag{12}$$

We can finally write after simplifications to define θ

$$\theta = \left(N \sum_{i=1}^{N} i\, \rho_{Y_i} - \frac{N(N+1)}{2} \sum_{i=1}^{N} \rho_{Y_i} \right) \Big/ \frac{N^2(N^2-1)}{12}. \tag{13}$$

If the value θ is close to zero, the distance between IOs in time has no trend towards change. This suggests that if these IOs were the candidates to be identified as one object by separate values ρ_{Y_i}, then most likely these IOs really belong to the same physical object. If the value θ is negative, then the distance between the IOs decreases over time, and therefore, the possibility that the identification problem is solved correctly increases. If the value θ is significantly greater than zero, most likely the IOs under consideration should not be identified as a single object, and the low value ρ_{Y_i} at a particular point in time from the interval was random. Moreover, a larger value θ gives greater confidence that the IO data refers to different physical objects.

8. Conclusions

The paper proposes a method for solving the problem of identifying IOs that enter the monitoring system's common information resource from several data sources. For this purpose, it is proposed to use a new proximity measure (similarity) of IO, which takes into account the nature of uncertainty of the type "probability" for quantitative features and the type "possibility" for qualitative features. At the same time, it does not require the transformation of feature values, which significantly simplifies the formation of a metric – a proximity (or distance) function between IOs as a whole according to all available features. The proposed measure was checked for compliance with the mandatory conditions for the validity of measures.

Performing the IO identification procedure allows the consumer to avoid duplication or data conflict, as well as increases the accuracy of IO representation in the monitoring system. Additionally, it is proposed to analyze the linear trend of the change in time of the distance

between the IOs to be identified, which can be calculated using the least squares method, which improves the quality of solving the identification problem.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] I. Mandel, Cluster analysis, Finance and statistics, Moscow, 1988.

[2] H. Newcombe, J. Kennedy, S. Axford, A. James. "Automatic linkage of vital records." Science 130.3381 (1959): 954-959.

[3] I. P. Fellegi, A. B. Sunter. "A Theory for Record Linkage." Journal of the American Statistical Association 64.328 (1969): 1183–1210.

[4] X. Wang, Q. Ji, A unified probabilistic approach modeling relationships between attributes and objects, in: Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2013, pp. 2120-2127. doi:10.1109/ICCV.2013.4. URL: https://openaccess.thecvf.com/content_iccv_2013/papers/Wang_A_Unified_Probabilistic_2013_ICCV_paper.pdf.

[5] T. Nagarajan, K. Grauman, Attributes as Operators: Factorizing Unseen Attribute-Object Compositions, in: Proceedings of the European Conference on Computer Vision, 2018. doi:10.48550/arXiv.1803.09851.

[6] P. Singla, P. Domingos, Object Identification with Attribute-Mediated Dependences, in: A.M.◦Jorge, L. Torgo, P. Brazdil, R. Camacho, J. Gama (Eds), Knowledge Discovery in Databases: PKDD 2005, Lecture Notes in Computer Science(), vol 3721, Springer Berlin, Heidelberg, 2005, pp. 297-308. doi:10.1145/775047.775116. URL: https://alchemy.cs.washington.edu/papers/pdfs/singla-domingos05b.pdf.

[7] W. Cohen, J. Richman, Learning to match and cluster large high-dimensional data sets for data integration, in: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, Association for Computing Machinery, New York United States, 2002, pp 475-480. URL: https://www.cs.cmu.edu/~wcohen/postscript/kdd-2002.pdf

[8] S. Sarawagi, A. Bhamidipaty, Interactive deduplication using active learning, in: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02, Association for Computing Machinery, New York, 2002, pp 269-278. URL: https://www.scribd.com/document/23688804/05-InteractiveDeduplicationUsingActiveLearning

[9] H. Pasula, B. Marthi, B. Milch, S. Russell, I. Shpitser, Identity uncertainty and citation matching, Advances in Neural Information Processing Systems 15 (2003): 1401-1408. URL: https://people.csail.mit.edu/milch/papers/nipsnewer.pdf

[10] L. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1.1 (1978): 3–28. doi:10.1016/0165-0114(78)90029-5.

[11] D.Dubois, H.Prade, Possibility Theory: An Approach to Computerized Processing of Uncertainty, Springer Science & Business Media, USA, 2012.

[12] Yu. A. Voronin, Classification theory and its applications, Nauka, Novosibirsk, 1985.

[13] B. G. Mirkin, Qualitative features and structures analysis, Finance and Statistics, Moscow, 1985.

[14] Yu. I. Zhuravlev, On the application of algebraical techniques in pattern recognition and classification problems. Pattern recognition, classification, forecasting, 1st. ed., Nauka, Moscow, 1989.

[15] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed., Pearson, 2009. URL: https://api.pageplace.de/preview/DT0400.9781292153971_A27091185/preview-9781292153971_A27091185.pdf