Proceedings of the 2013 Federated Conference on
Computer Science and Information Systems pp. 257–260

Similarities in Spaces of Features and Concepts: Towards Semantic Evaluations

Wladyslaw Homenda
Faculty of Mathematics and Information Science
Warsaw University of Technology
ul. Koszykowa 75, 00-662, Warsaw, Poland
Web page: wwww.mini.pw.edu.pl/∼homenda

Agnieszka Jastrzebska
Faculty of Mathematics and Information Science
Warsaw University of Technology
ul. Koszykowa 75, 00-662, Warsaw, Poland
Email: A.Jastrzebska@mini.pw.edu.pl

***Abstract*—The article discusses abstract spaces of concepts and features. Concepts correspond to real-world objects. Concepts are described by their features. The study is devoted to relations in the space of concepts and in the space of features. Of greatest interest is similarity of structures in the concepts and features spaces. There is a direct link between features and concepts. Therefore, similarity may be analyzed through structures of both concepts and features. Authors propose generalized similarity relation, applicable to the developed framework. In addition, similarity of nested sets of the space of features and concepts is discussed. Authors introduce an algorithm, which calculates similarity of two structures of nested structures. Developed semantics leads to the set-theoretic model, which allows to flexibly describe abstract information.**

I. Introduction

SIMILARITY is one of the most important dependencies in our environment. In this article developed framework for constructing generalized similarity relations is presented.

Similarity estimation has to involve narrowing the focus on chosen aspects of compared objects. In our nomenclature object is called concept, object's attribute is a feature. We chose these names intentionally, to highlight that our model may be applicable to modeling in various areas of science. In sections II-A and III the core of developed framework of phenomena description is depicted. We start from the space of features - pieces of information, that describe concepts. In section III-B our own approach to similarity relations modeling in the fuzzified space of concepts and features is introduced. Similarity relations for features vectors are discussed. We introduce also an algorithm for computing similarity of linear orders in the space of features.

The goal of the paper is to present the research on concepts and features spaces descriptions and similarity modeling. Valuation mappings and similarity relations able to process fuzzified descriptions of real-world objects are introduced.

II. Preliminaries

In our approach a start point are features - descriptions of concepts. Features are gathered in vectors. We operate in the namespace of features and in the space of their evaluations rather than on the concepts (objects) themselves. We are interested in similarity of descriptions of concepts, i.e. in vectors of features' evaluations. We assume, that each hypothetical or real concept can be described with qualitatively the same set of features, but evaluated differently.

Due to space limitations we do not present literature review on this topic. Interesting research on similarity can be found not only, but also in: [2], [5], [6], [8] and [9]. It is important to mention, that approaches present in the literature are suitable for similarity based on features. Our model is concepts' oriented in its nature. Therefore, it is necessary to include relations between concepts and features.

We have developed a framework for describing the space of concepts and the space of features. We have also proposed similarity measures dedicated for this model, which we present in the next paragraphs.

A. The space of concepts and the space of features

1) The space of features: A concept corresponds to a real-world object. Usually, due to various constraints and complexity of real-world phenomena, we do not operate directly on concepts. Instead, we describe them with their features. In the developed model the space of features is defined as follows:

$$\mathcal{D} = \{(\mu_1, \ldots, \mu_n) : \mu_i \in [0,1],\ i = 1, \ldots, n,\ n \in \mathcal{N}\} \quad (1)$$

Under our assumptions features are imprecise. Concepts correspond to real-world objects. One of many imprecise information representation models known in literature may be applied, for example: [1] or [3]. Authors treat imprecise information analogically to the uncertainty in the sense of Zadeh. We are aware that there are other frameworks (i.e. probability theory) that are able to describe uncertainty, which we are not recalling here.

Features evaluations are expressed through their degree of membership as a single numerical value from the $[0,1]$ interval. Features vectors belong to the namespace of features. There is unlimited amount of features vectors evaluations, but in our application there is finite amount of features. Of interest is possibility of features space structuring by introducing certain relations, like inclusion, exclusion and overlapping.

2) The space of concepts: $\mathcal{C}$ is the set of all concepts fulfilling some logical conditions, e.g. consumers from a city, pixels from certain images, musical symbols of some score.

$$\mathcal{C} = \{c_1, c_2, \ldots, c_r\} \quad (2)$$

978-1-4673-4471-5/$25.00

such that $r \in \mathcal{N}$ is the number of concepts in the space $\mathcal{C}$. The source of information about concepts may be for example measurement devices or questionnaire surveys.

The space of concepts is limited. We are interested in structuring the space of concepts through the space of features, the space of their evaluations and dependencies in these spaces as well as in their subsets. Concepts' space analysis will be performed using relations of similarity, inclusion, exclusion and others.

In the next section similarity relations customized for the developed model are proposed. Presented technique aims at mimicking human way of how similarity is estimated. Concepts are described with their features, comparison happens in a feature-wise fashion. The strength of belongingness of a particular feature to the concept (corresponding to the degree of membership) influences similarity measure.

III. Similarity in the space of concepts and in the space of features

In this section we introduce similarity measures adjusted for spaces of concepts and features.

A. The valuation mapping

Firstly, let us discuss the valuation mapping. It is a relation, that for every vector of features assigns a set of concepts.

$$V : \mathcal{D} \to 2^{\mathcal{C}} \tag{3}$$

For instance, the simplest valuation assigns all concepts with a given features' vector $(\mu_1, \ldots, \mu_n)$ to this vector of features:

$$V(\mu_1, \mu_2, \ldots, \mu_n) = \{c \in \mathcal{C} : d(c) = (\mu_1, \mu_2, \ldots, \mu_n)\} \tag{4}$$

where the mapping $d : \mathcal{C} \to \mathcal{D}$ defines features for concepts. Valuation mapping V generates a subset of the space of concepts, such that each i-th feature of selected concepts was evaluated exactly the same, as the respective i-th feature from analyzed vector $(\mu_1, \mu_2, \ldots, \mu_n)$. Valuation mapping defined in formula 4 is called *pointwise valuation mapping*.

Alternatively, we propose generalized approach to valuation mappings called *filling up valuation mapping* defined as follows:

$$V_I(\mu_1, \ldots, \mu_n) = \big\{c \in \mathcal{C} : d(c) = (\mu_{c1}, \ldots, \mu_{cn}) \text{ and } \mu_{ci} \leqslant \mu_i \text{ for } i = 1, \ldots, n\big\} \tag{5}$$

Filling up valuation mapping generates a subset of the space of concepts, that groups objects, which each i-th feature was not evaluated as greater than respective feature in given features vector. In other words, filling up approach is a generalization of pointwise valuation mapping V, which translates features vectors into groups of concepts. As a result, we extract objects, which description satisfy certain conditions to some point, but not beyond that point. In this study conditions are between 0 and 1, but they may be different if we assume other information representation model (for example: balanced fuzzy sets defined in [3] utilize $[-1, 1]$ interval, intuitionistic fuzzy sets defined [1] employs doubled unit interval $[0, 1]$).

Valuation mappings V and V_I are semantic mappings. They allow transformation from the namespace of features and the space of their evaluations to subsets of the space of concepts. In practice, we use valuation mappings to generate subsets of the space of concepts, in which we are interested in. The key, by which subsets are generated, are features - and that was our initial goal, to describe concepts and groups of concepts by their features.

B. Similarity relations

In this section we introduce similarity relation for features' vectors. Let us assume that we have two vectors of features:

$$\mu_A = (\mu_{A1}, \mu_{A2}, \ldots, \mu_{An})$$
$$\mu_B = (\mu_{B1}, \mu_{B2}, \ldots, \mu_{Bn})$$

Below we introduce a generalized similarity measure $s_{G(eneralized)}$ of two vectors of features.

$$s_G(\mu_A, \mu_B) = \frac{\big|V_I(\mu_A) \cap V_I(\mu_B)\big|}{\big|V_I(\mu_A) \cap V_I(\mu_B)\big| + \mathcal{V}_{A\setminus B} + \mathcal{V}_{B\setminus A}} \tag{6}$$

where:

$$\mathcal{V}_{A\setminus B} = \int_0^{\rho_{max}} \alpha(x) \cdot \mathcal{V}_{A\setminus B}(x)\, dx$$
$$\mathcal{V}_{B\setminus A} = \int_0^{\lambda_{max}} \beta(x) \cdot \mathcal{V}_{B\setminus A}(x)\, dx \tag{7}$$

and

α and β are real nonnegative functions

and

$$\mathcal{V}_{A\setminus B}(x) = \Big|\big\{c \in V_I(\mu_A) \setminus V_I(\mu_B) : \max_{i=1}^{n}\{\mu_{Ai} - \mu_{ci}\} = x\big\}\Big|$$
$$\mathcal{V}_{B\setminus A}(x) = \Big|\big\{c \in V_I(\mu_B) \setminus V_I(\mu_A) : \max_{i=1}^{n}\{\mu_{Bi} - \mu_{ci}\} = x\big\}\Big|$$

and

$$\rho_{max} = \min\big\{\rho \geqslant 0 : (\forall\, i = 1, 2, \ldots, n)\ \mu_{Bi} + \rho \geqslant \mu_{Ai}\big\}$$
$$\lambda_{max} = \min\big\{\lambda \geqslant 0 : (\forall\, i = 1, 2, \ldots, n)\ \mu_{Ai} + \lambda \geqslant \mu_{Bi}\big\}$$

and

$$\mu_c = \big(\mu_{c1}, \mu_{c2}, \ldots, \mu_{cn}\big)$$

is the vector of features of the concept c.

Valuation mapping V_I transforms vectors of features into subsets of the space of concepts satisfying certain conditions (as in formula 5). Generalized similarity s_G is calculated as a fraction. Similarity is enlarged as the size of intersection of compared subsets of the space of concepts grows. Similarity becomes smaller as the number of elements that do not belong to sets' intersection grow. The decreasing effect of features, which are not shared, is conditioned on the difference between these features and vectors μ_A and μ_B. We integrate on the space of features. Integration is done separately for concepts in $V_I(\mu_A) \setminus V_I(\mu_B)$ and in $V_I(\mu_B) \setminus V_I(\mu_A)$. The domain of integration spans from 0 to ρ_{max} or to λ_{max} respectively for these two cases. Functions α and β allow to introduce more

punishing effect of $V_I(\mu_A) \setminus V_I(\mu_B)$ and $V_I(\mu_B) \setminus V_I(\mu_A)$ on the similarity value. α and β may be also used to enhance nonsymmetry of relation s_G.

Let us also introduce a discretized version of similarity relation s_G (named $s_{D(iscretized)}$). For any $0 < \rho$ and $0 < \lambda$ the value of similarity is based on estimations of integrals from the formula 6 in a following manner:

$$\mathcal{V}_{A\setminus B}(x) = \sum_{j=1}^{\rho_{max}} \Big(\alpha(j) \cdot \mathcal{V}_{Aj}\Big)$$
$$\mathcal{V}_{B\setminus A}(x) = \sum_{j=1}^{\lambda_{max}} \Big(\beta(j) \cdot \mathcal{V}_{Bj}\Big) \tag{8}$$

where:

$$\mathcal{V}_{Aj} = \left|\Big\{c \in V_I(\mu_A) \setminus V_I(\mu_B) : \tau_{\rho j}\big(\max_{i=1}^{n}\{\mu_{Ai} - \mu_{ci}\}\big)\Big\}\right|$$
$$\mathcal{V}_{Bj} = \left|\Big\{c \in V_I(\mu_B) \setminus V_I(\mu_A) : \tau_{\lambda j}\big(\max_{i=1}^{n}\{\mu_{Bi} - \mu_{ci}\}\big)\Big\}\right|$$

and

$$\tau_{\rho j}(\text{exp}) \;\equiv\; \rho \cdot (j-1) < \text{exp} \leqslant \rho \cdot j$$
$$\tau_{\lambda j}(\text{exp}) \;\equiv\; \lambda \cdot (j-1) < \text{exp} \leqslant \lambda \cdot j$$

and

$$\rho_{max} = \min\Big\{j{=}1,\ldots : (\forall\, i{=}1,\ldots,n)\ \mu_{Bi} + \rho \cdot j \geqslant \mu_{Ai}\Big\}$$
$$\lambda_{max} = \min\Big\{j{=}1,\ldots : (\forall\, i{=}1,\ldots,n)\ \mu_{Ai} + \rho \cdot j \geqslant \mu_{Bi}\Big\}$$

Analogously to the formula 6, we use filling up valuation mapping V_I. It produces a subset of the space of concepts, that satisfies given conditions to the extent not greater than the conditions stated in the input features vector. By analogy, in the s_D compared sets intersection and outlying parts ($V_I(\mu_A) \setminus V_I(\mu_B)$ and $V_I(\mu_B) \setminus V_I(\mu_A)$) are accounted. Concepts not present in sets' intersection are lying in one of the two remaining parts. They decrease the value of similarity in a nonsymmetrical fashion (through functions $\alpha(j)$ and $\beta(j)$). The decreasing impact of concepts lying in sets' differences is conditioned on the difference between the particular concept μ_c and μ_A or μ_B. We may visualize such „outlying" concepts as crescent-shaped hulls around sets' intersection. These crescent-shaped hulls are divided into up to ρ_{max} and λ_{max} segments. The greater amount of concepts fall to the furthest part of such crescent, the more decreasing effect there is on the similarity value.

We may simplify formula s_D further on. Instead of functions α and β under the sum in formulas 8, we may multiply coefficients by parameter λ or ρ and by j in a following way:

$$\mathcal{V}_{A\setminus B}(x) = \sum_{j=1}^{\rho_{max}} \Big(\rho \cdot j \cdot \mathcal{V}_{Aj}\Big)$$
$$\mathcal{V}_{B\setminus A}(x) = \sum_{j=1}^{\lambda_{max}} \Big(\lambda \cdot j \cdot \mathcal{V}_{Bj}\Big) \tag{9}$$

The proposed idea relies on relation built around sets' intersection and concepts lying beyond this intersection. Alignment of the outlying concepts influences similarity value. Similarity relation's codomain is $[0, 1]$. Note, that it is reflexive. The relation s_G was also intentionally constructed to be nonsymmetric, but they can be adjusted to be symmetric. Asymmetry of similarity relation is a highly desirable property from the applicational point of view. Due to space limitations we do not elaborate on similarity relation properties.

In given definition of valuation mapping V_I, what strikes immediately, is that the similarity relation induced by this mapping may also create linear orders (chains) in the space of subsets of $\mathcal{D}$. In this article we discuss linear orders only. Of interest is similarity of such nested structures. In the next section we present this nontrivial modeling problem to a greater extent.

C. Similarity of linear orders in the space of concepts

In this paragraph we investigate such subsets of the space of features $\mathcal{D}$, that this mapping computes the same value for all features' vectors included in such subset, i.e. $\mathcal{A} \in \mathcal{D}$ is such subset if $(\forall\, \mu_1, \mu_2 \in \mathcal{A})\, V_I(\mu_1) = V_I(\mu_2)$. The structures of such subsets can be formally described as linear orders. Let us recall that linear order is a pair $(X, \leqslant)$, where X is a set of elements and $\leqslant$ is a binary relation satisfying axioms of: antisymmetry, transitivity and totality.

Let us introduce a similarity relation able to compare subsets of the space of features nested in the sense explained above, i.e. $(\forall\, \mathcal{B}, \mathcal{A} \in \mathcal{D})\ \mathcal{B} \leqslant \mathcal{A} \equiv V_I(\mathcal{B}) \subset V_I(\mathcal{A})$. Also, whenever $\mu_A = (\mu_{A1}, \ldots, \mu_{An}) \in \mathcal{A}$ and $\mu_B = (\mu_{B1}, \ldots, \mu_{Bn}) \in \mathcal{B}$ and $\mu_{Bi} \leqslant \mu_{Ai},\ i = 1, \ldots, n$, then $\mathcal{B} \leqslant \mathcal{A}$.

Comparing nested subsets of the space of features requires taking into account not only cardinality of compared subsets, but also actual elements lying inside. Therefore, measures of similarity operating only on cardinalities are not satisfactory to describe such structures. We need to compare actual content of each subset.

In order to compare nested subsets of the space of features we have developed an algorithm, which we describe here. Features nesting is understood in a following way: given features vector $\mu_A = (\mu_{A1}, \mu_{A2}, \ldots, \mu_{An})$ nests features vector $\mu_B = (\mu_{B1}, \mu_{B2}, \ldots, \mu_{Bn})$ if:

$$\mu_{Bi} \leqslant \mu_{Ai},\ i = 1, 2, \ldots, n \tag{10}$$

The space of features (see formula 1) and subsets of the space of concepts (see formula 2) generated with valuation mapping V_I (defined in formula 5) are recalled here. Nested features, through filling up type of transformation enforced by the valuation mapping, generate nested subsets of the space of concepts. Structure of nested sets in the space of concepts corresponds to features' nesting.

If sets of concepts are nested, then each bigger set contains each concept from each smaller set and optionally several more concepts. Let us analyze an exemplar order E_o, which contains following 3 subsets of some space of concepts:

$$P_1 = \{c_1, c_2\}$$
$$P_2 = \{c_1, c_2, c_4\}$$
$$P_3 = \{c_1, c_2, c_4, c_7, c_8\}$$

where P_1 is a subset generated with $V_I(\mu_{P_1})$, P_2 is a subset generated with $V_I(\mu_{P_2})$ and P_3 is a subset generated with $V_I(\mu_{P_3})$. μ_{P_1}, μ_{P_2} and μ_{P_3} are particular features vectors evaluations. For clarity of algorithm description, nomenclature used later refers only to subsets of the space of concepts named as P_m. The method of obtaining these subsets, through valuation mapping V_I, is assumed by default. In the given example of E_o following structure is observed: $P_1 \subseteq P_2 \subseteq P_3$.

In the developed algorithm, written to compare two nested structures of the space of concepts, as input data we have two such structures (denoted as $E1$ and $E2$):

$$E1 = P_{i_1}, \ldots, P_{i_k}$$
$$E2 = P_{j_1}, \ldots, P_{j_l}$$

where $P_{i_1}, \ldots, P_{i_k}, P_{j_1}, \ldots, P_{j_l}$ are subsets of the space of concepts. $E1$ and $E2$ are linear orders, so $P_{i_1} \subseteq \ldots \subseteq P_{i_k}$ and $P_{j_1} \subseteq \ldots \subseteq P_{j_l}$. We do not make any assumptions about the length of orders E1 and E2. $P_{i_1}, \ldots, P_{i_k}, P_{j_1}, \ldots, P_{j_l}$ contain concepts. $P_{i_1}, \ldots, P_{i_k}, P_{j_1}, \ldots, P_{j_l}$ are sets, so the order of appearance of concepts in each set can be omitted. Hence, we always use alphabetical order, to improve efficiency of our algorithm. Each order E can be written as a sequence of concepts, starting from the „deepest" of the nested sets.

The input data to our algorithm are two linear orders $E1$ and $E2$ written as sets in a form (convention) described above.

1) Start with order E, which has last set smaller. If cardinalities of last sets of both orders are equal, choose one order at random. Denote this order as $E_{f(irst)}$ and its last subset as P_{f_u} (it is either P_{i_k} or P_{j_l}).
2) Denote the second order as $E_{s(econd)}$. Denote last (biggest) set of order E_s as P_{s_v}. Search for such set in the order E_s, which shares the biggest number of the same elements (concepts) with P_{f_u} and is the largest. We assumed inclusion (see formula 10), so the last set of E_s, which is P_{s_v} will be always chosen. It is either P_{i_k} or P_{j_l}, the one not chosen in the above point.
3) Collate E_f and E_s in a following way: set P_{f_u} corresponds to set P_{s_v}, set $P_{f_{u-1}}$ corresponds to set $P_{s_{v-1}}$ and so on. If sets from one order run out, assume $\varnothing$.
4) For each pair of sets P_{f_i} and P_{s_i} compute $s_D(P_{f_i}, P_{s_i})$. The formula for s_D is given in 6 with redefined nonoverlapping parts defined by formulas 8 (see section III-B).
5) Similarity of linear orders E_f and E_s is equal to aggregated similarities computed as in point 4, i.e. $\text{aggr}\{s_D(P_{f_i}, P_{s_i}), i = 1, 2, \ldots, \max\{f_u, s_v\}\}$, with some aggregation operator aggr. Note that for linear orders, as in discussed case, collated are simply P_{i_k} and P_{j_l}, $P_{i_{k-1}}$ and $P_{j_{l-1}}$ etc.

Properties of the developed algorithm depend on assumed information representation model, on the similarity relation applied in step 4 and on the aggregating operator calculated in step 5. Due to space constraints we do not compare here various possibilities, which may be chosen in steps 4 and 5. In our first attempt as aggregating function we took mean, as it is very intuitive and normalized measure of dependency between real numbers (and the sum of all $s_D(P_{f_i}, P_{s_i})$ gives us a real number). To maintain comparability, aggregating function should produce normalized values of similarity.

IV. Conclusions

The article discusses developed model of features and concepts spaces. Of interest is similarity between descriptions of real-world objects, which we call concepts. Such descriptions (features vectors evaluations) through valuation mapping generate subsets of the space of concepts. Valuation mapping from the namespace of features into subsets of the space of concepts can be performed in a point-wise fashion or in an filling up way. All concepts form the universe of discourse, on which we do not directly operate. Instead, we use features, which describe concepts.

Two similarity relations developed for this model are introduced. First one is a generalized similarity relation between features vectors. Second one is a discretized version of the generalized similarity relation. Presented measures take into account fuzziness of analyzed information.

In this paper an algorithm of evaluating similarity between structures of nested features vectors based on generalized similarity relation and valuation mapping is introduced.

Acknowledgment

The research is supported by the National Science Center, grant No 2011/01/B/ST6/06478, decision no DEC-2011/01/B/ST6/06478.

A. Jastrzebska contribution is supported by the Foundation for Polish Science under International PhD Projects in Intelligent Computing. Project financed from The European Union within the Innovative Economy Operational Programme (2007-2013) and European Regional Development Fund.

References

[1] K. T. Atanassov: *Intuitionistic fuzzy sets*, Fuzzy Sets and Systems 20, 1986, pp. 87-96.
[2] Belanche L. , Orozco J., *Things to Know about a (dis)similarity Measure*, in: LNAI 6881, 2011, pp. 100-109.
[3] Homenda W., *Balanced Fuzzy Sets*, Information Sciences 176, 2006, pp. 2467-2506.
[4] Hung W., Yang M., *Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance*, in: Pattern Recognition Letters 25, 2004, pp. 16031611.
[5] Julian-Iranzo P., *A procedure for the construction of a similarity relation*, in: proc. of IPMU'08, 2008, pp. 489 - 496.
[6] Klawonn F., Kruse R., *Similarity Relations and Independence Concepts*, in: G. Della Riccia, D. Dubois, R. Kruse, H.-J. Lenz (eds.): Preferences and Similarities, Springer, Wien, 2008, pp. 179- 196.
[7] Orozco J., Belanche L. ,*On Aggregation Operators of Transitive Similarity and Dissimilarity Relations*, w: FUZZ-IEEE, 2004, pp. 1373-1377.
[8] Szmidt E., Kacprzyk J., *A Similarity Measure for Intuitionistic Fuzzy Sets and Its Application in Supporting Medical Diagnostic Reasoning*, in: LNAI 3070, pp. 388-393, 2004.
[9] Tversky, A., *Features of Similarity*, in: Psychological Reviews 84 (4), 1977, pp. 327-352.