Data Visualization While Determining Similarities of Medical Patterns

T. RZEŹNICZAK
tomek.rzezniczak@gmail.com

Institute of Computer and Information Systems
Faculty of Cybernetics, Military University of Technology
Kaliskiego Str. 2, 00-908 Warsaw, Poland

The article presents the concept of using the theory of similarity in the recognition of medical patterns. The aim of the work is to construct a graphical model of disease entity pattern and the state of the patient's health in such a way as to use natural human ability of perception to identify similarities between them. With this approach, the representation of medical patterns can be used to support the diagnosis process of disease entities.

Keywords: data visualization, similarity models, similarity relation, medical diagnostic.

1. Introduction

The job of physicians is based on processing large amounts of medical information describing the patient's health status, on which physicians make decisions and direct the treatment. With the development of science and technology, the number of information sources continues to grow. Physicians now have to their disposal, apart from medical interviews and physical examinations, much more specialized tests. Despite the advanced technological developments, it is still the logical thinking of the physician in conjunction with information collected in many kinds of tests that is the basis of diagnosis. By examining, the physician wants to gather as much information since any of them may affect the diagnosis. At the same time the amount of data increases the difficulty of their analysis, which is the basis for identifying the syndrome.

To set the initial or the final diagnosis, the doctor must critically evaluate the information collected and match them with known disease entities. Given the amount of possible disease entities and the amount of medical data, this task in fact is not an easy one. In addition, many factors that can cause errors, have an impact on its outcome.

Erroneous medical decisions are frequently cognitive – they are errors of reasoning, which are caused by emotions that influence the perception of physicians and their activity [12]. An example could be the expected confirmation of a diagnosis by carefully selecting information or diagnosing just the easier to associate diseases, and forgetting about the rare cases. Besides, no one is able to master the entire medical knowledge, so some errors are due to ignoring the uncertainty associated with the lack of knowledge.

Given the above issues, this work is to initially verify the applicability of visualization methods to support the recognition of disease entities and to conduct treatment. Visualization methods have already been applied in many areas of life such as science, business and media. Examples of their use can be noticed when watching the weather forecast, tracking stock market results or using maps. Visualization is common since the explosion of information forced the search for more effective methods of their processing, and thanks to the innate abilities of human perception, visualization is a very powerful tool. Well-designed visual representation allows to quickly receive information and to analyze more amounts of data by a human than with other methods.

Returning to the medical diagnosis process, the work focuses on the possibility of developing a graphic form of the patient's health condition and patterns of disease entities in order to use the natural abilities of perception to recognize the similarities between them. In addition to the methods of visualization, theory of similarity plays a key role in the work [16], [24], [20], [13]. Similarity models are the basis for the construction of a visualization, they serve as guidelines for the constructing a graphical representation and its evaluation. The tasks rely on finding the optimal visualization, it is one that most effectively supports the diagnosis by comparing the patient's health condition with the disease entity.

2. Similarity Models

As already mentioned, the theoretical basis of the work are based on the analysis of similarity relations. Similarity is the foundation of cognition, it allows the activation of memory according to what we see [14], the categorization of objects [22], decision making or solving new problems based on similar, previously known situations [21]. In the context of psychological similarity between objects we can define it as the mental representation proximity of these objects.

Many models have been created describing the relationship of similarity, among which geometric models dominate. This type of model represents each object as a point in space (usually Euclidean space), and the distance between points corresponds to the similarity of objects. An example of such a model is MDS (Multidimensional Scaling) [11]. Part of the model is a statistical technique, for which the input data are evaluations of similarities or differences between all objects in the model under consideration. The result of the technique is a geometric model representing objects as points in an n-dimensional space.

Formally, MDS can be described as follows: let *k* be the number of all objects under consideration and *n* is the number of attributes of each object. Matrix X with a dimension of $k \times n$ will contain the spatial coordinates of the objects, where the row *i* indicates the coordinates of the object *i*. However, the difference between objects *i* and *j*, will be described by δ_{ij}. The distance in the Euclidean space between objects *i* and *j* is defined as the shortest line connecting *i* with *j* and takes on the form:

$$d_{ij}(X) = \left(\sum_{s=1}^{n} (x_{is} - x_{js})^2 \right)^{\frac{1}{2}} \tag{1}$$

The purpose of MDS is to find such a matrix X that $d_{ij}(X)$ corresponds δ_{ij}. This assumption can be presented in various forms, including in the least squares MDS model proposed by Kruskal [16]:

$$\sigma^2(X) = \sum_{i=2}^{k} \sum_{j=1}^{i-1} w_{ij} (\delta_{ij} - d_{ij}(X))^2 \tag{2}$$

where w_{ij} is a non-negative weight. For example, many MDS implementations take $w_{ij} = 0$ for the missing differences.

The main assumptions of geometrical models is to meet the following d_{ij} distance conditions:

- Non-negativity

$$d_{ij} > d_{ii} = 0 \text{ (for } i \neq j\text{)} \tag{3}$$

- Symmetry

$$d_{ij} = d_{ji} \tag{4}$$

- Triangular condition

$$d_{ij} \leq d_{ih} + d_{hj} \tag{5}$$

These assumptions have been criticized by Tversky [24], [25], as affecting the empirical observations of similarity. Simultaneously Tversky proposed a different model of similarity, defined by the characteristics of objects. Each object is described by a set of features, and the similarity between objects *a* and *b*, is expressed by the function of common and distinctive features.

Denoted by the *A* set of features of object *a* and the *B* set of features of object *b*, the $s(a,b)$ similarity is a function of three arguments, measuring the level that two sets of features fit together (Fig. 1):

- $A \cap B$ – common features *a* and *b*
- $A - B$ – features of *a* not occurring in *b*
- $B - A$ – features of *b* not occurring in *a*

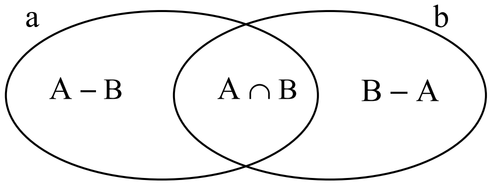


Fig. 1. Graphical representation of sets of features of objects *a* and *b*

Interval similarity scale $S(a,b)$ (*contrast model*), preserving the order of similarity [$S(a,b) > S(c,d)$ if $s(a,b) > s(c,d)$] is expressed as a linear combination of the measures of common and distinctive features:

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \tag{6}$$

where: $\theta, \alpha, \beta \geq 0$, and f is a function representing the contribution of different features of objects in their similarity.

This model does not define a unique index of similarity, but their family, as defined by parameters θ, α, β, thereby allowing the introduction of various relations of similarity between the same objects, such as:

- if $\theta = 1, \alpha = \beta = 0$,
  then $S(a,b) = f(A \cap B)$
- if $\theta = 0, \alpha = \beta = 1$,
  then $-S(a,b) = f(A-B) + f(B-A)$

Several hypotheses were defined concerning human perception of similarity in terms of *contrast model*, which then were tested in empirical research [24], [25], where it was confirmed that for man:

I. More important are the common features in determining similarity than in determining difference – focusing attention hypothesis;
II. More important are the features of the compared object (subject) rather than the object with which the comparison is made (reference) - asymmetry hypothesis;
III. More important are the features that are relevant for classification – context hypothesis.

Let us assume that $s(a,b)$ and $d(a,b)$ will be respectively measures of similarity and difference. From the focusing attention hypothesis (I) results that $s(a,b)$ grows along with $f(A \cap B)$ and decreases with the increase of $f(A-B)$ and $f(B-A)$, however $d(a,b)$ decreases with the increase of $f(A \cap B)$ and increases along with the increase of $f(A-B)$ and $f(B-A)$. *Contrast model* weights θ, α, β associated with common and distinctive features will also vary when changing the centre of interest. In the case of evaluating similarities we focus more attention on the similar features, and in the case of a difference we focus more attention on the distinctive features, resulting in weight θ of common features is greater for assessing similarity than for assessing the difference and vice versa.

The asymmetry hypothesis (II) implies that the relation of similarity should not be treated as symmetric (as is the case of geometric models). We cannot assign equivalence to claims such as ”*a* is similar to *b*” and ”*b* is similar to *a*”. Selection of the subject and reference depends largely on the relative importance of objects. We are inclined to choose objects more important as the reference and less important as the subject. For $s(a,b)$, *a* is the subject, *b* – the reference. Naturally, we focus attention on the subject, therefore the subject features are more important than features of the reference $(\alpha > \beta)$, and the similarity is more reduced by the distinguishing features of the subject than the reference. For example, a toy train resembles a real train more, because most of the attributes of the toy train is in a real train. On the other hand, a real train is not as similar to the toy, since many of its attributes are not included in the toy [25]. In the *contrast model:* $s(a,b) = s(b,a)$ if and only if $f(A-B) = f(B-A)$ *or* $(\alpha = \beta)$. This means that the symmetry is preserved only when the objects are equally important and their comparison is non-directional, that is, we evaluate the level at which *a* and *b* are similar, and not the level at which *a* is similar to *b* and vice versa.

The context hypothesis (III) tells us, however, that the significance of individual features may vary depending on the considered set of objects and methods of evaluation. An example might be to assess the similarity between two countries bordering close to each other, while having different political systems, such as North Korea and South Korea [24]. Both countries will be judged as more similar to each other in context of European countries or African countries than Asian. Weight of features changes as follows:

- a feature may become more important in some context, if it is the basis for classification in this context
- features that are shared by all objects under consideration do not have a classification value
- when we expand the context, some features may take on the classification value, because they cannot be divided by new objects, so that increases the similarity of objects from their original context
- therefore the similarity of a pair of objects in the original context will be usually smaller than in the extended.

Previous hypotheses were associated with parameters θ, α, β, hypothesis (III) concerns the issue to what degree does *f* change depending on the context of the features.

The *contrast model* also has some gaps, and in many cases the psychological representation is better characterized by structured hierarchical systems. For this reason, *structural models of similarity* [20] were created, which assume that the process of assessing the similarity of a pair of objects must take into account the relations between attributes, and not just attributes. The following example explains this assumption: to model the representation of the ”red square on a blue triangle” attributes of ”red” and ”square” should be combined with the object at the top as well as ”blue” and ”triangle” with the object at the bottom, and adjust the upper and lower object in the ”over” relationship [20]. Thus, the comparison of objects requires a structural adjustment process, what geometric models lack and those based on features. Therefore, by using them, distinguishing ”red square on a blue triangle” from ”red triangle on a blue square” will be unsuccessful. Capturing these properties requires a structural representation.

According to research, the definition of similarity involves the same process of structural mapping, which is used in inference through analogy [20], [10], [17]. Because of this, structural models of similarity introduce an additional division for common and distinctive attributes of objects. We have two kinds of common attributes [10]:

- MIP (*Match In Place*) is a match between common attributes
- MOP (*Match Out of Place*) is a match between differing attributes.

For example, by comparing a bird with a gray head and red wings with a bird with a gray head and a red tail, the colours of heads are MIP, while the red wings and red tail are MOP. MIP has a greater impact on the similarity than MOP. There are also two kinds of differences between the compared objects [20]:

- agreeable differences – differences between common attributes of objects
- non agreeable differences – differences between attributes that do not match or differences between an attribute in one representation, which does not correspond to any attribute in the other representation.

An example of the *agreeable difference* for a car and motorcycle is the number of wheels, which they possess. However, an example of a *non agreeable difference* for the same objects may be seatbelts, because a motorcycle does not have a device that corresponds to seatbelts. Similar objects tend to have more agreeable differences than differing objects. *Agreeable differences* are easier to determine, they are more important for similarities than *non agreeable differences*.

For completeness of considerations the *transformation models* should be mentioned, which are based on the theory of the Kolomogorov complexity [17], [13]. According to this point of view the measure with which object *a* is similar to *b* is the number of steps in which *a* can be transformed to *b*. Thus the similarity is a function of transformational complexity. Transformation steps may be different in nature and depend mainly on the representation of objects. For sentences of a given language may be, for example, linguistic operations at the level of words, syntaxes and semantics, for structures in the form of a tree – tree transformation operations, etc.. For example XXXOOXO is more similar to OXOOXXX than OXOOXX because OXOOXXX requires only a reflection of XXXOOXO, and OXOOXX requires a reflection plus the removal of X from the right side OXOOXXX [15].

It should be noted that the structural representations, that pose problems for spatial models or feature based models, are easily carried out in the transformation model. Larkey and Markman found some evidence against the transformation model, showing that the number of steps needed to transform colours and shapes of geometric objects is not relevant to human assessment of their similarity [12].

3. Process Characteristics of Data Visualization

Analyzing the data visualization process, you should pay attention to the aspect of the data type and its dimensionality, which directly affects the available forms of presentation. For purposes of visualization, there are three types of data: *nominal*, *ordinal*, *quantitative* [6]. *Nominal* data type is one that can only be equal to or different from other nominal values, *ordinal* data have an additionally defined order, while with *quantitative* data, we can perform arithmetic operations. For example, attributes describing a car, such as manufacturer and model, are nominal data, the segment is of *ordinal* type, and the distance driven is *quantitative* data type. By analyzing dimensionality, we can distinguish the following types of data [23]:

- 1D – linear or sequential data, such as text or program source code (sets and sequences)
- 2D – flat data, such as a floor plan (maps)
- 3D – physical objects, such as molecules, buildings or the human body (shape)
- Temporal – data that include time lines, such as patient records, project data, historical data
- Multidimensional – with more than three variables, like most relational or statistical databases (*case-by-variables*)
- Trees/Hierarchies – each node (except the root) has its unique parent
- Networks/graphs – structures composed of nodes and connections between them.

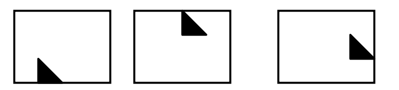
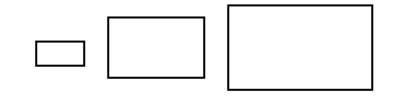
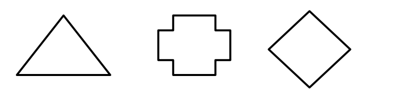
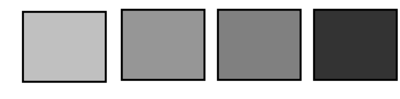
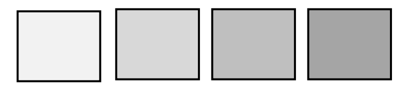
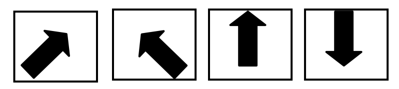
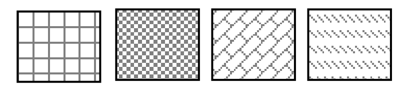
The object of our interest are mainly multi-dimensional data, since they correspond to the information about the patient’s health condition and disease entities. While the human eye is well adapted for cases of 1D, 2D and 3D, beyond this limit, we cannot easily map data on graphics structures.

There are many techniques for multidimensional data visualization to deal with the above-mentioned problem. The base for creating visualizations are basic units of information representation called marks recognized by Bertina [4], [5]. Featured symbols are:

- points – denoting the position in space
- lines – representing information of a certain length
- areas – having a length and width (2D)
- surfaces – areas in 3D without thickness
- volume – having a length, width and depth.

In addition, methods of modifying *symbols* were defined called *visual variables* [5]. These include: shape, size, texture, intensity/value, colour, orientation, position. According to Bertin's theory, the human eye is sensitive to these variables, so they are received by the eye effortlessly and automatically (Tab. 1).

Tab. 1. Examples of *image variables*

Position	
Size	
Shape	
Intensity/ Value	
Colour	
Orientation	
Texture	

It was observed that the *visual variables* have a different information transfer efficiency for quantitative data depending on the type of graphical coding. On this basis, the Cleveland & McGill ranking was created, which was later expanded and supplemented by Mackinlay [23], [8], [19]. Mackinlay's ranking (Tab. 2) also includes *nominal* and *ordinal* data types, moreover it expands the image *variables*.

Tab. 2. Mackinlay's ranking of information transmission efficiency of *visual variables* – in order from most to least efficient [19]

Quantative	**Ordinal**	**Nominal**
Position	Position	Position
Length	Intensity/Value	Colour (hue)
Angle	Colour (saturation)	Texture
Slope	Colour (hue)	Connection
Area (Size)	Texture	Containment
Volume	Connection	Intensity/Value
Intensity/Value	Containment	Colour (saturation)
Colour (saturation)	Length	Shape
Colour (hue)	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area (Size)
Shape	Shape	Volume

According to [18], [7] the most significant aspect of visualization is the use of space. It is treated in a particular way in relation to other image attributes – it is the basis, on which other elements are then distributed. So, the empty space of the image is a container with a metric structure that can be described by axis:

- unstructured axis (no axis);
- nominal axis (the region is divided into a sub-region);
- ordinal axis (the order of sub-regions is considerable);
- quantitative axis (region with the metric)

Axes can be linear or radial shape.

Moreover, techniques have been developed that allow increasing the amount of coded information on each axis [7]:

- *Composition* – orthogonal arranging the axis, which allows direct modeling of the relation between data
- *Alignment* – repeating the axis in different places in space, for example, side by side
- *Folding* – continuation of the axis in the dimension perpendicular to it (when there is no place for it)
- *Recursion* – repeated division of space (e.g. reflecting the directory structure)
- *Overloading* – reuse the same space many times (worlds in worlds), based on the fact that the data covers only a portion of space, which allows for the development of the remainder of it.

It should also be mentioned of the importance of the use of symbols and lines to present certain topology [7]. They allow the representation of relations between objects without geometric constraints (mapping data on the axes of space): *Connections*, that is the indication of relations between objects by drawing links between them, and *Containment*, consisting of representing relations by drawing objects contained within themselves. The form of presentation of the structure may have an impact on its reception by the observer. Bertin distinguished five forms of structural representation of the image (Fig. 2) [5]:

- linear structure – organizes the elements without the use of position
- circular structure – simple in design and allows presenting relations with straight lines
- pattern – the form, in which just the position does not carry any information, but the created pattern can present symmetry or similarity in the structure
- ordered pattern – two-dimensional representation, where one dimension is in order
- stereogram – uses arrangement to present volume and enables observation of 3D patterns.

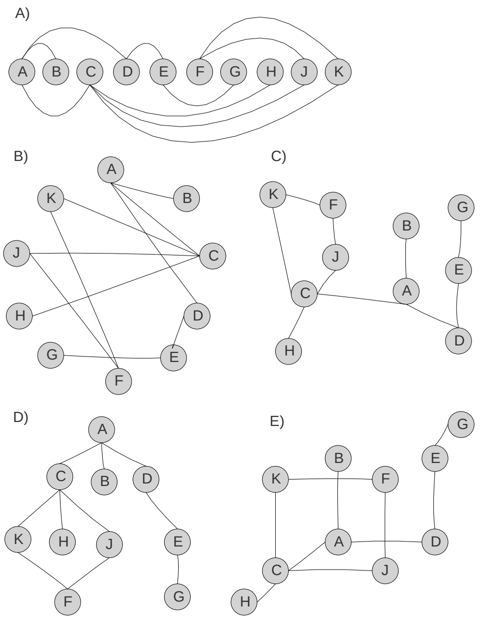


Fig. 2. Forms of structural representation of Bertin's image [5]: A – rectilinear, B – circular, C – pattern, D – orderly pattern, E – stereogram

Among the techniques for multidimensional data visualization, we have a large collection of ready solutions at our disposal, such as: Charts, Treemaps, Scatterplot matrix, Reordable matrix, Parallel coordinates, Glyphs, Spiderweb Chart, Pixel-oriented Technique [23], [3]. Analysis of individual solutions goes beyond the scope of this work, so there will not be further developed of the topic here.

4. Models of Medical Patterns

Now let us consider how the medical data look like that will be visualized and compared. In this paper we will use a simplified model of the description of the patient's condition and the disease entity model proposed by Ameljańczyk [1], [2]. In full version, the model is based on two elements: a set of symptoms and a set of risk factors. Symptoms are all signs of illness identified during the visit to the doctor and as a result of conducted specialized tests, such as high body temperature, swollen glands, coughing, runny nose, fluid in the sinuses, etc. Risk factors are occurrences, which allow to predict the likelihood of the development of the disease entity, for example: obesity, smoking, alcohol abuse, lack of physical activity, etc. To simplify the discussion in the paper we will consider a model limited only to symptoms.

If we assume that the set $S\subset \mathbf{N}$ is a set of numbers of all the symptoms, which can describe the disease entity, then the disease entity model $m\in M=\{1,...,\mathrm{M}\}$ can be formally written as:

$$M(m)=(S^m,C^m), \qquad (7)$$

where:

S^m – set of symptoms numbers of disease $m\in M$

C^m – set of disease value ranges of symptoms of disease $m\in M$.

Additionally, S^m and C^m:

$$S^m=\{s_1^m,...,s_k^m,...,s_{K(m)}^m\}\subset S, m\in M \qquad (8)$$

$$C=\{C_1^m,...,C_k^m,...,C_{K(m)}^m\}\ \ m\in M \qquad (9)$$

$K(m)$ – number of symptoms of disease entity $m\in M$.

$C_k^m=\lfloor \underline{c}_k^m, \overline{c}_k^m \rfloor$ – disease value range of symptom *k* in disease $m\in M$.

The model of the patient's condition $x\in X$, built based on the disease entity model, will be presented in the form of:

$$P(x)=(S_o(x),W(x)) \ , \qquad (10)$$

where: set $S_o(x)\subset S$ is a set of disease symptoms numbers occurring in a patient, and $W(x)$ is a set of levels of severity of different symptoms:

$$S_o(x)=\{s\in S|w(x,s)>0\} \qquad (11)$$

with $w(x,s) \in W(x)$ − level of severity of symptom $s \in S$.

It should be noted that if $s = s_k^m$ and $w(x,s) \in C_k^m$, then the symptom severity is in the range of the disease values for the disease entity $m \in M$.

5. Medical Data Visualization Space

Changing the disease entity model and the model of the patient's health condition to a graphical representation will be started by introducing the *Information Visualization Design Space* [6]. We will use a narrowed down description proposed by Mackinlay, i.e. *Space Visualization*, which is based on:

- *Symbols:* Point, Line, Area, Surface, Volume
- *Visual variables:* Color, Size, Shape, Intensity, Orientation, Texture, Connection, Inclusion, Position (X, Y, Z).

For comparison reasons of individual visualizations, Mackinlay presented principles of visualization in the form of a table, which, adapted to our needs, will contain columns as in Tab. 3. Their explanation is presented in Tab. 4.

Tab. 3. Visualization description in the form of a table [6]

Variable	**D**	**F**	**D’**	**R**	**X**	**Y**	**Z**

Tab. 4. Explanation of symbols for the tabular description of the visualization

Symbol	**Definition**
Variable	Name of the represented information
D	Data type: N (Nominal), O (Ordinal), Q (Quantative)
F	Function re-coding data, e.g.: f (unspecified), > (filter), s (sorting), mds (MDS)
D’	Data type re-coded
R	Visual variable: C (Color), S (Size), F (Shape), (V) Intensity, O (Orientation), T (Texture), -- (Connection), [] (Inclusion)
X,Y,Z	Position in space represented by a symbol: P (Point), L (Line), S (Surface), A (Area), V (Volume)

For example, information that will be presented in our visualization derive from a disease entity model and a model of a patient's health condition. Visualization attributes of the disease entity (limited only to symptoms and their standard disease values), written in the form of a table could look like Tab. 5.

Tab. 5. Example of the visualization description in a tabular form for the disease entity

Variable	**D**	**F**	**D’**	**R**	**X**	**Y**	**Z**
Symptom	N				A	A	
Standard disease symptom value	Q	f	O	C			

Tab. 5 presents only one of the possible variants of visualization, in which the symptoms as a *nominal* (N) data type are mapped to areas (A) in space (X,Y), while the standard disease symptom value (quantitative type) (Q) is transformable using the function *f* to the *ordinal* (O) set of data and mapped to a color (C).

The description in the form presented above can be used only to compare the properties of different visualizations. It is certainly insufficient to construct a target image. There are a few missing aspects, among others there are no precisely defined methods of arranging symbols or a method of using information coding techniques. It is therefore necessary to supplement it with transformation rules that will transform an object into a particular image. A full set of rules generating a unique image on the basis of object attributes will be called the *visualization model.*

For the purpose of further consideration, let us assume that any object *o* belonging to the universe $o \in U$, is represented as a set of attributes $o = \{c_1, c_2, ...\}$ and denoted by *g* of any image belonging to the set of all possible images $g \in G$ to generate. Our projection, converting the object to an image based on the visualization model, can be written in the form of $v : U \to G$, i.e. $v(o) = g$.

6. Study of Visualization Model

Let us denote by $V = \{v_1, v_2, ...\}$, the set of all possible *visualization models* of disease entities and patient health condition, where $v \in V$ is the *visualization model* defined as in the shown above description. With $v(m)$ we will be denoting the application of model *v* for the disease entity $m \in M$, which is a specific graphical representation (image) of a disease entity *m*. For simplification reasons we will assume $p = P(x)$ as patient *x*'s health condition. Its graphical representation generated by the visualization model *v* will be indicated $v(p)$. We assumed that *m* and *p* are objects, for which the same *visualization model v* can be applied. This can be accepted since both types of objects are made up of symptoms and respectively – the standard disease value and the current level of symptom intensification.

Our task is to find an optimal *visualization model* $v^* \in V$ that maximizes the similarity assessment of the disease entity and the patient's health condition, if the medical condition corresponds to the given unit. This means that when a physician compares the graphical representation of the patient's health condition with examples of graphical representations of various disease entities, then the most similar disease entity for him/her will be the one that the patient is suffering from.

Previously discussed similarity models can be used as a basis for evaluation and formulation of constraints for verified solutions. In accordance to the *contrast model*, we can define a similarity scale projecting a natural (perceived by the observer) order of compared visualizations of disease entities and health condition in the form of $s(a,b) = s(v(p), v(m))$. The value v^* that is search by us must meet condition:

$$s\left(v^*(p), v^*(m)\right) \ge s\left(v\ (p), v\ (m)\right) \qquad (12)$$

for every $p = P(x)$ and for every $m \in M$, when patient *x* is ill with disease entity *m*, where $v \in V_m \setminus \left\{v^*\right\}$.

Another condition, which we will defined is:

$$s\left(v^*(p), v^*(m_1)\right) > s\left(v^*(p), v^*(m_2)\right) \qquad (13)$$

for every $p = P(x)$ and for every $m_1, m_2 \in M$, when patient *x* is ill with disease entity m_1 and is not ill with m_2.

The final condition written as follows:

$$s\left(v^*(p_1), v^*(m)\right) > s\left(v^*(p_2), v^*(m)\right) \qquad (14)$$

corresponds to the case, in which $p_1 = P(x_1)$ and $p_2 = P(x_2)$ represent the health condition of patients x_1 , x_2 ill with the same disease entity $m \in M$, where patient x_1 has more symptoms corresponding to disease entity *m* than patient x_2, that is, the following inequality occurs between cardinalities of sets of common symptoms of health condition and disease entity:

$$card\left(S_o(x_1) \cap S^m\right) > card\left(S_o(x_2) \cap S^m\right) \quad (15)$$

On the basis of (12), among all visualization models of the disease entity with the application of visualization model *v** for each case of the patient's health condition and disease entity (which corresponds to this case), the biggest similarity is observed. From (13) results that within the visualization model *v**, the biggest similarity will always concern the disease entity similarity, which corresponds to the case of the patient's health condition. However, the last condition (14) is fulfilled, if with the increasing number of symptoms corresponding to the disease entity, the similarity increases of the patient's health condition and the entity.

Keeping in mind the form of the linear combination of *contrast model* (4), and the hypotheses of similarity (I, II, III), we can consider the task of finding the optimal visualization model. Let us assume that:

$$attr(g) = \{a_1, .a_2, \ldots\} \qquad (15)$$

where $g \in G$ and $n \in N$, is the operator determining the set of image attributes identified by the observer. Then for weights θ, α, β, scale f and defined $G^p = attr(v(p))$ and $G^m = attr(v(m))$ the model has a form of:

$$\begin{aligned} S\left(G^p, G^m\right) &= \theta f(G^p \cap G^m) \\ &- \alpha f(G^p - G^m) \\ &- \beta f(G^m - G^p) \end{aligned} \qquad (16)$$

Therefore increasing the observed similarity will consist of:

- $\theta \to \max$
- $\alpha, \beta \to \min$
- $f\left(G^p \cap G^m\right) \to \max$
- $f\left(G^p - G^m\right) \to \min$
- $f\left(G^p \cap G^m\right) \to \min$

It should be noted that the set of visualized object attributes, in our case – a set of symptoms, does not translate easily to the attributes of the visualization itself. The visualization may contain attributes that are an indirect result of its structure. According to the creators of *Gestalt psychology*, human image perception should be treated as a whole, without decomposing into smaller entities [9]. According to this theory, only the global relation between all elements determines the main aspects of perception, e.g. perception of a circle created out of single points, which are equally distant from the center. Therefore, for the visualization of the disease entity size of the sets of the disease entity

$$card\left(G^p \cap G^m\right) \qquad (17)$$

and

$$card\left(S_o(x)\cap S^m\right) \tag{18}$$

for $m\in M, p=P(x)$, they do not have to be equal. Thus, in order to try to calculate the similarity, we have to complete the task of determining all the characteristics that are recognized by the observer, i.e. constructing the operator *attr*.

As previously discussed the visualization model is defined as a set of rules that generate the image for a certain object. By Analyzing the various hypotheses of similarity, we consider their impact on the optimal construction of such rules. On the basis of (I), we can assume that visualization models will be preferred that display common attributes. This is because in our case, the task put forward to the observer is to assess the similarity and not assessing the difference. The relation directionality resulting from hypothesis (II) has already been implicitly introduced in our consideration by defining the task as an "assessment of the similarity of the patient's health condition with the disease entity." In this situation, the subject is the patient's health condition and the disease entity is the reference. Thus, a greater impact on the assessment of similarity are the distinguishing attributes of the patient's health condition than the attributes distinguishing the disease entity, which is another indication affecting the construction of the visualization model.

Using hypothesis (III) also becomes a source of guidance. First of all attributes should be preferred in the presentation that are relevant to the classification; secondly, an important role in the process may have the presentation of each image to the observer. For example, we can imagine that the presentation of the patient's health condition in the context of a few different images of disease entities will change the assessment of similarity in relation to the presentation in the context of a single entity.

7. Conclusion

This paper presents preliminary results in the scope of assessing the role of data visualization in determining the similarity of medical patterns. Described similarity models can serve as a guide to the evaluation of visualization methods. The emphasis has been on the *contrast model*, however, during further research other models should not be forgotten. For example, it may be interesting to apply the MDS model in the first stage of selecting disease entities, with which the patient's health condition will be matched with. Visualizing all disease entities at the same time seems impossible, therefore, reduction of this set at the initial stage should be applied. For this purpose we could use another type of image created by MDS. The graphics would present the patient's health condition and disease entity as a single point on the plane. Therefore, the use of MDS requires knowledge of the value of difference/similarity δ_{ij} between individual objects, for our case they could be calculated on the basis of the number of corresponding symptoms in individual objects. Such a proposed visualization would present in the immediate neighbourhood of the point representing the patient's health condition points representing disease entities having the most common symptoms with it. The second stage would be to present the observer the visualization of the patient's health condition in context of a few selected disease entities from the first stage – many simultaneously, or sequentially, with each individually.

The main issues for further research include the development of the subject of visualization space and the creation of description principles of a full visualization model. Another area is to formulate a complete optimization problem, which solution would be the best visualization model (in accordance with pre-defined understanding of this concept).

Closely associated with this is also the construction of a method of models assessment. This is a nontrivial task, if we assume to try to at least partially automate it. It is related to the issue of identifying attributes of objects, i.e. building the *attr* operator. It will probably be useful here to refer to publications dealing with the human perception of the image, such as the work of Colin Ware [26]. At the same time, because of ready algorithms such as: SME (*Structural Mapping Engine*) [20], SIAM (*Similarity as Interactive Activation and Mapping*) [10], CAB (*Connectionist Analogy Builder*) [10], the use of *structural* and *transformational models* can be an area of additional research opportunities, particularly in the scope of visualization model assessment.

In conclusion, the presented work attempts to define the basis for further research of medical data visualization. The expected result is to find a visualization model that allows for the creation of a new tool to support physicians in diagnosis and treatment, thus contributing to the elimination of some popular medical malpractice.

8. Bibliography

[1] A. Ameljańczyk, „Analiza wpływu przyjętej koncepcji modelowania systemu wspomagania decyzji medycznych na sposób generowania ścieżek klinicznych", *Biuletyn Instytutu Systemów Informatycznych*, Nr 4 (2009).

[2] A. Ameljańczyk, „Wielokryterialne mechanizmy wspomagania podejmowania decyzji klinicznych w modelu repozytorium w oparciu o wzorce", *Biuletyn Instytutu Systemów Informatycznych*, Nr 5 (2010).

[3] M. Ankerst, "Visual Data Mining with Pixel-oriented Visualization Techniques", *ACM SIGKDD Workshop on Visual Data Mining*, San Francisco, CA, 2001.

[4] J. Bertin, *Graphics and Graphic Information-Processing,* Walter de Gruyter, Berlin, 1981.

[5] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*, The University of Wisconsin Press, Wisconsin, 1983.

[6] S. Card, J. Mackinlay, "The Structure of the Information Visualization Design Space", *INFOVIS '97*, IEEE Computer Society Washington, DC, USA, 1997.

[7] S. Card, J. Mackinlay, B. Shneiderman, *Readings in information visualization: using vision to think*, Morgan Kaufmann Publishers, San Francisco, 1999.

[8] W. Cleveland, R. McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods", *Journal of the American Statistical Association*, 79, 531−554 (1984).

[9] E. Goldmeier, "Similarity in Visually Perceived Forms", *International Universities Press, Inc.*, 1972.

[10] R. Goldstone, "Similarity, Interactive Activation, and Mapping", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 20, No. 1, 3–28 (1994).

[11] P. Groenen, M. van de Velden, "Multidimensional Scaling", *Econometric Institute Report*, EI 2004-15, April (2004)

[12] J. Groopman, *Jak myśli lekarz*, Wydawnictwo Dolnośląskie, Wrocław, 2009.

[13] U. Hahn, N. Chater, L. Richardson, "Similarity as transformation", *Cognition*, 87, 1−32, Elsevier, 2003.

[14] D. Hintzmann, "Schema abstraction in a multiple-trace memory model", *Psychological Review*, 93, 411–428 (1986).

[15] S. Imai, "Pattern similarity and cognitive transformations", *Acta Psychologica*, 41, 433–447 (1997).

[16] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika*, 29, 1–27 (1964)

[17] L. Larkey, A. Markman, "Processes of Similarity Judgment", *Cognitive Science*, 29, 1061−1076, Cognitive Science Society, Inc., 2005.

[18] A. MacEachren, *How Maps Work*, The Guilford Press, New York, 1995.

[19] J. Mackinlay, "Automating the Design of Graphical Presentations of Relational Information", *ACM Transactions on Graphics* , Vol. 5, Issue 2, New York, USA, April, 1986.

[20] A. Markman, D. Gentner, "Structural Alignment during Similarity Comparisons", *Cognitive Psychology*, 25, 431–467, Academic Press, 1993.

[21] D. Medin, R. Goldstone, A. Markman, "Comparison and choice: Relations between similarity processes and decision processes", *Psychonomics Bulletin and Review*, 2, 1−19 (1995).

[22] R. Nosofsky, "Attention, similarity and the identification-categorization relationship", *Journal of Experimental Psychology*, 115, 39–57 (1986).

[23] H. Siirtola, "Interactive Visualization of Multidimensional Data", *Dissertations in Interactive Technology*, Vol. 7, 2007.

[24] A. Tversky, "Features of Similarity", *Psychological Review*, Vol. 84, Number 4, 327−352, American Psychological Association , July, 1977.

[25] A. Tversky, I. Gati, "Studies of Similarity", in: *Cognition and Categorization*, 79−98, E. Rosch & B. B. Lloyd (Eds.), Hillsdale, 1978.

[26] C. Ware, *Information Visualization: Perception for Design, 2nd Edition*, Morgan Kaufmann Publishers, 2004.

Wizualizacja danych w określaniu podobieństwa wzorców medycznych

T. RZEŹNICZAK

W artykule przedstawiono koncepcję wykorzystania teorii podobieństwa w rozpoznawaniu wzorców medycznych. Celem prowadzonych prac jest skonstruowanie postaci graficznej wzorca jednostki chorobowej oraz stanu zdrowia pacjenta, w taki sposób, aby wykorzystać naturalne zdolności percepcyjne człowieka do identyfikacji podobieństwa między nimi. Dzięki takiemu podejściu, reprezentacja wzorców medycznych może zostać zastosowana do wsparcia procesu diagnozowania jednostek chorobowych.

Słowa kluczowe: wizualizacja danych, modele podobieństwa, relacja podobieństwa, diagnostyka medyczna.