Standardized vectorial representation of medical data in patient records

Wolfgang ORTHUBER [a,1] and Efthymios PAPAVRAMIDIS [a]
[a]*Department of Orthodontics at Universitaetsklinikum Schleswig-Holstein Campus Kiel, Germany*

Abstract. In this paper standardized vectorial (quantitative) representation of medical data is suggested for use in patient records. Vectorial representations are (as sequences of numbers) language independent, precise, directly comparable, and they allow advanced evaluation, e.g. similarity calculation using well defined distance functions. It is possible to search for a patient with a certain combination of diagnostic parameters on the Web records of patients with similar parameters. The information about chosen treatments and treatment outcome at these patients can be used anonymously or pseudonymously for decision support. Because patient records from all countries can be compared, in the long run this could open systematic access to a very large wealth of clinically relevant information. Here the technical principle is described and illustrated by examples (e.g. similarity search of heart sounds). Previously published material is integrated in parts for explanation of the motivation and background.

Keywords. Vectorial Resource Descriptor, VRD, Vector Space Identifier, VSI, Vector Space Descriptor, VSD, Quantifiable Resource, QR, QRI, Feature vectors, task sharing, Patient record databases, metric databases, similarity comparison, decision support

1. Introduction

Step by step it is becoming common to store information about anamnesis, diagnostics, treatment and treatment result in electronic patient records. Together these records summarize a huge and permanently increasing wealth of valuable information which by far exceeds the knowledge of any human doctor. It is clear that this should be used for decision support, according to the wishes of the patients (anonymously or pseudonymously). Until now, however, patient records are usually stored in separated databases in incompatible formats. To alleviate exchange of health data, much effort is invested in building standards for electronic interchange of clinical, financial, and administrative information among health care oriented computer systems [1], and huge vocabularies have been created. For example, SNOMED CT [2] is a medical ontology with increasing size that contains already today within its English version more than 300000 concepts, 900000 descriptions and 1300000 relationships. There are also large

[1] Corresponding Author: Arnold Heller Str. 16, 24105 Kiel, Germany; E-mail: orthuber@kfo-zmk.uni-kiel.de
Citation:
Orthuber W, Papavramidis E. Standardized vectorial representation of medical data in patient records [Internet]. In: Medical and Care Compunetics 6. London, UK: IOSPress; 2010. p. 153-166.Available from: http://www.booksonline.iospress.nl/Content/View.aspx?piid=16922

collections of identifiers available for exchange of quantitative data, e.g. LOINC® [19]. Quantitative data can help to improve the resolution of symbolic descriptions. This is desirable and welcome for decision support. Because optimal therapeutic decisions are the aim, the representation of medical data should be even designed for decision support. Here we propose a concept for this. It is designed to support the clinician in his *cycle of decision*:

(a) The clinician makes measurements (in the broadest sense, also speaking with the patient and looking at a picture is a measurement).

(b) The clinician focuses on those measurement results which are interesting for his therapeutic decisions (feature extraction).

(c) The clinician compares these measurement results with experience. At this he may use rules or models which are derived from common experience.

(d) The clinician decides for therapy, and measures the effect of his decision, i.e. the cycle starts again with (a).

Usually several measurements are done before therapeutic decisions, and so (a) can be a complex process which also requires decisions. The decision for the initial measurements has to be done by the clinician, together with the patient who tells the clinician his concerns. After this initial data acquisition the clinician has a rough impression and can select a keyword (a "rough diagnosis") or already provide certain measurement results to get support when he decides for further fine measurements. In 2.5 is described how this can be used to get help for steps (a) and (b). For support of steps (c) and (d) we propose standardized vectorial representation of medical data. *Vector*s or *feature vectors* (sequences of numbers) within a vector space (metric space) provide besides high resolution also multidimensional comparability. This is especially in step (c) important and allows high resolution search within a large collection of data.

The usage of vector spaces is (of course) not new in informatics, there are already well known applications e.g. in bioinformatics, biophysics, signal processing, imaging, and vector spaces are used for data integration within the framework of the Conceptual Space approach [3][4][5]. Conceptual Spaces follow a theory of describing entities at the conceptual level in terms of their natural characteristics similar to human cognition in order to avoid the symbol grounding issue [6]. They enable representation of resources as vectors within a geometrical space which is defined through a set of quality dimensions. For instance, a particular color may be defined as vector with the dimensions hue, saturation, and brightness. This is finer and more precise than a symbolic representation. Describing instances as vectors furthermore enables the automatic calculation of their similarity, in terms of their distance, in contrast to the costly representation of such knowledge through symbolic representations. Even complex data which describe e.g. faces, sounds, fingerprints and biometric data can be processed by feature extraction to vectorial form for similarity comparison and recognition. Generally, feature extraction is an essential pre-processing step to pattern recognition and machine learning problems. The resulting feature vectors open a large spectrum of applications for vector spaces [7].

It is possible to make precise diagnostic measurements on a patient, and to make feature extraction on decision relevant findings. The resulting vectorial representation can be standardized. This would allow to search worldwide (within all accessible data collections) the records of patients with similar diagnostic findings, to study possible therapies and their long term consequences with probabilities of failure and success, in order to support the clinician in making better decisions.

This motivated us to write this paper which proposes standardized vectorial representation of medical data. First we introduce the basics and describe the application of vector spaces for representation of quantitative properties and data integration in general. Then a framework for vectorial representation of medical data and efficient implementation of vectorial similarity search is presented.

2. Material and Methods

2.1. Appropriate resources for vectorial representation

The numbers which represent quantitative data can be regarded as components of a vector. Quantification is precondition for vectorial representation. To be suitable for vectorial representation, a resource must have one or several quantitative properties, i.e. one or several attributes which each have an inherent order (from "little" to "great"). We will call such a resource "*Quantifiable Resource"* (QR). If two QRs are represented by instances of the same quantitative property (or properties in case of many dimensions), these QRs are *comparable*. Their vectorial description belongs to the same vector space.

For example measurement results and numeric results of feature extraction are QRs. Important for similarity comparison is that small changes of the original are mapped to small changes of the representing numbers.

2.2. Vector space

The mathematics of *vector spaces* or *linear spaces* is focused in linear algebra and used in many areas. There are well known applications in nature sciences, e.g. in physics. Also in computer science vectors are used for representation of concepts or real world objects, e.g. in metric databases. All vectors which represent the same sort of data are comparable and belong to the same vector space. They can be added, multiplied and subtracted (which is important for comparison). Let n denote the dimensionality of a vector space V^n, then the coordinate vectors $\mathbf{e}_1 = (1, 0, ..., 0)$, $\mathbf{e}_2 = (0, 1, 0, ..., 0)$, to $\mathbf{e}_n = (0, 0, ..., 0, 1)$, form a *basis* of V^n, called the *standard basis*, so that any vector $(x_1, x_2, ..., x_n) \in V^n$ can be uniquely expressed as linear combination of these vectors:

$(x_1, x_2, ..., x_n)$
$= x_1(1, 0, ..., 0) + x_2(0, 1, 0, ..., 0) + ... + x_n(0, ..., 0, 1)$
$= x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + ... + x_n\mathbf{e}_n.$

It is possible to define a norm [8] or "length" function on a vector space. Then the vector space can be called *metric space* [9]. Let $\| \|$ denote the norm and A, B two vectors, then the distance d between A and B is the length of the difference vector:

$$d(A,B) = \|A - B\|$$

With A and B also the difference $A - B$ is a vector. There are many norms and associated metrics (distance functions). Well known are the Manhattan metric and the Euclidean metric. For clarification we write the Euclidean metric here explicitly - without loss of generality. Let A, B denote two vectors with $A = (a_1, a_2, ...a_n)$, $B = (b_1, b_2, ...b_n)$, $a_i, b_i \in \Re$, then the unweighted Euclidean distance between A and B is

$$d_u(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad (1)$$

Different dimensions can have different variance and importance. Therefore it is adequate to define a weighting vector $W = (w_1, w_2, ...w_n)$, $w_i \in \Re \setminus \{0\}$ and the *weighted* Euclidean distance:

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (w_i a_i - w_i b_i)^2} \quad (2)$$

If $w_i = 1$ for every dimension $i \in (1, 2, ...n)$, the unweighted distance (1) is identical to the weighted distance (2). Here by "distance" we always mean the weighted distance. The greater w_i, the more contributes dimension i to the distance. Formula (2) shows that $d(A,B) \geq 0$; the distance $d(A,B)$ is the greater, the more A and B differ, and $d(A,B) = 0$ if and only if $A = B$, i.e. $a_i = b_i$ for every dimension. So we see that (2) can be used to quantify similarity between two vectors.

2.3. Vectorial Resource Descriptors (VRDs)

The proposed data structure for representation of a QR is called *Vectorial Resource Descriptor* (VRD). We suggest a design with full integration into the Web of Linked Data, to achieve maximal efficiency within the current Web Infrastructure. The Resource Description Framework (RDF) can be used, with HTTP URIs (and not other URI schemes as URNs and DOIs) as identifiers, in agreement with the recommendations for Linked Data [10]. HTTP URIs provide as Web Addresses a simple way to create globally unique names[2] and task sharing without centralized management, and they work not just as a name but also as a means of accessing information about a resource over the Web. This is in many applications even necessary, in connection with the vectorial approach it is necessary to get immediate access to the definition of the valid distance function of the vector space for calculation of similarity.

The VRD structure is shown in Figure 1. It contains:

(e) The *identifier of the QR* (QRI). It is a HTTP URI which points to the resource.

[2] These can integrate also names of existing definitions, e.g. codes of LOINC® [19] or DOIs, so systematic adoption of existing work is possible.

(f) The *Vector Space Identifier* (VSI). It is a HTTP URI which points to the *Vector Space Descriptor* (VSD).

(g) The *feature vector* (usually a sequence of numbers [3]). It represents the quantitative properties of the resource.

Additionally it can contain:

(h) Auxiliary data, e.g. date, keywords.

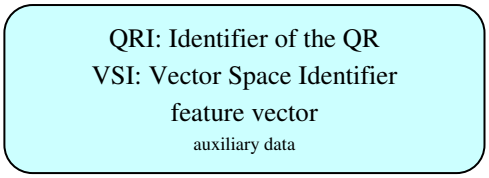

Figure 1. The Vectorial Resource Descriptor (VRD). It is the data structure for representation of a QR. The components QRI and VSI are both HTTP URIs.

The feature vectors of all VRDs with the same VSI are elements of the same vector space and comparable as described in 2.2. Similarity search is done within this space. As HTTP URI the VSI not only identifies the content of the feature vector, it simultaneously points to the *Vector Space Descriptor* (VSD, Figure 2) which provides all necessary information about the vector space, particularly about the metric for comparison, definitions of dimensions, templates for human readable representation of instances and links to further related Web content.

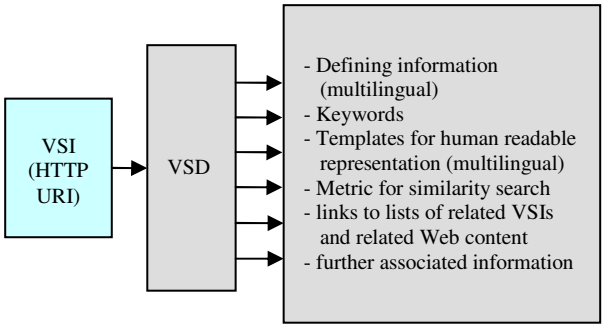

Figure 2. The *Vector Space Identifier* (VSI) points to the *Vector Space Descriptor* (VSD) which provides important information about all VRDs with this VSI.

[3] This sequence can generally represent numeric mathematical objects (also e.g. matrices for conversion of vectors, tensors). The search process also allows an expanded definition: More generally it can be defined as a data structure designed for efficient comparison. Usually the result of the comparison is a number.

VRDs are machine readable and can be embedded into the semantic Web as Linked Data [11]. They are uniformly comparable and searchable, and the usage of HTTP URIs makes it possible to share the work for definition of vector spaces and generation of VRDs among all domain name owners.

2.4. Vectorial Web Search

Subsequently we assume that the patient records are distributed over the Web and accessible according to the wishes of the patient. So the total Web is the database of patient records, therefore we use the term "search engine" for the tool which allows similarity search of patient records according to the wishes of the user. The VRDs are the fundament of Vectorial Similarity Search over the total Web (Vectorial Web Search). Due to the standardized structure of the VRDs one and the same search engine can be used for all search queries. Vectorial Web Search consists of the following steps:

- User provides a VRD or only its feature vector F with VSI.
- User confines search by a regular expression and/or by a conventional word based search string S (optional)
- Search engine selects all VRDs
  - with the chosen VSI
  - optionally with string S at the associated resource (identified by QRI)

  via conventional word based search
- If a regular expression is given in step two, the collection is confined so that it fulfills this expression.
- Using the metric provided in the VSD (figure 2) the search engine calculates distances between the feature vector F and the feature vectors of the collected VRDs and sorts them according to distance.
- In the search result the rank of collected VRDs and associated resources is the higher, the smaller the distance is.

As it is apparent in the above list, Vectorial Web Search starts with word based search (for the VSI): After the user has provided a VSI and feature vector F, among all VRDs those with this VSI are selected and (after optional further confinement) used for comparison, i.e. the distances between their feature vectors and the searched feature vector F are calculated. In the search result the rank of VRDs and (via QRI, figure 1) associated resources is the higher, the less the distance is, i.e. most similar resources are listed first.

2.4.1. Combination of VRDs

It is possible to combine several VRDs in a search to find resources which are simultaneously associated with (resp. in the QRI field of) all VRDs. At this it is necessary to use a combined distance d_{comb} . It can be set to the weighted sum of the single distances:

$$d_{comb} = \sum_{i}^{n} m_i d_i$$

in which n is the number of combined VRDs, m_i the weighting factor and d_i the distance of VRD *i*. One cannot expect that the user provides appropriate weighting factors m_i, so there must be the option for automatic weighting. A possible solution is to set

$$m_i = \frac{1}{s_i}$$

in which s_i is the standard deviation of the distances of VRD *i*. This is a reasonable choice because it takes into account that the statistical significance of a fluctuation is the greater, the less the standard deviation of the concerned variable is.

2.5. Application in medicine

The here proposed application in the medical domain is the representation of decision relevant anamnestic and diagnostic (and past therapeutic, e.g. medication) data in vectorial form (as VRDs). As mentioned in the introduction (see 1), in the *cycle of decision* the clinician makes (unconsciously) feature extraction from all available measurements (also from complex measurements like MRI scans) and compares the result with experience. Important parts (of course not all) of this procedure can be mapped to database algorithms.

First the scientific community has to start a *diagnosis-VSI-database*: A list of important keywords or "rough diagnoses" must be defined or derived from existing ontologies like SNOMED CT [2]. Then to every [4] rough diagnosis standardized measurement combinations must be associated, which provide the most important data necessary for therapeutic decisions and for control of therapeutic success, with standardized numeric representations of the results - as combinations of elementary VRDs with appropriate VSIs[5], with combined distance functions (default see 2.4.1). The combination of elementary VRDs with minimal dimensionality can be appropriate to facilitate flexible queries. After creation of the diagnosis-VSI-database the following workflow is possible:

- Select a rough diagnosis in the diagnosis-VSI-database.
- Get from the diagnosis-VSI-database the information about the combination of VSIs which is appropriate for this rough diagnosis, and make all necessary measurements to get the VRDs with these VSIs which are connected with this rough diagnosis.
- Optionally make further measurements which are interesting (relevant for therapeutic decisions.) and convert them to VRDs.
- Select all decision relevant VRDs and send them to the search engine to search patient records which have within a maximal date range these VRDs with minimal distance (2.4.1).
- Look in the found patient records whether additional diagnostic measurements have been useful. If appropriate, make these measurements, convert these to VRDs and continue with the previous step.

[4] It makes sense to start with most important and most expensive rough diagnoses.

[5] If the medical community defines a 3D reference coordinate system on an anatomical reference model, it is possible to use a VSI for localization of most diagnostic findings and the "rough diagnosis" need not contain localization information.

- Look in the found patient records for the most successful therapeutic decisions. Statistical evaluation and mathematical modeling is possible.

3. Examples

3.1. VRD combination for vertebral (osteoporotic) compression fractures

Let us take as example "vertebral compression fracture" as rough diagnosis. We will reflect on a VRD combination, which could be associated to this rough diagnosis in the diagnosis-VSI-database. Aim is to pack most important decision relevant information in a sequence of numbers as short as possible. In case of orthopedic findings the geometry is usually important, for an osteoporotic compression fracture we can define the numbers *v*, *c* and *d* which show the relative remainders of the vertebral body ventrally (*v*), centrally (*c*) and dorsally (*d*). Furthermore the number *a,* which shows the relative narrowing of the vertebral channel can be important, and the number *n* of the vertebra (numbered consecutively from head to sacrum). Let *d1, d2, d3, c2, v1, v2, v3, a1, a2, a3* denote the scalar lengths of the lines as drawn in figure 3. Then we can calculate the numbers *v, c, d, a* as follows:

*v=2*v2/(v1+v3),*
*c=4*c2/(v1+v3+d1+d3),*
*d=2*d2/(d1+d3),*
*a=2*a2/(a1+a3).*

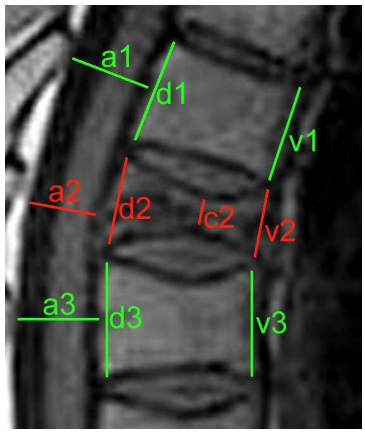

Figure 3. MRI scan of a vertebral compression fracture in the area of maximal compression. The lengths *a1,a2, a3, d1, d2, d3, c2, v1, v2, v3* are used for construction of the feature vector of this fracture as described in the text.

The VRD for the geometry of the compression has the 5 dimensional feature vector *(v,c,d,a,n)*. Additionally a representative measurement of the bone density *t* like the DXA T-score (at representative vertebra, e.g. L1) may be important, so that we could combine (2.4.1) the two VRDs with feature vectors *(v,c,d,a,n)* and *(t)* to search for patient records which contain similar parameters *v, c, d, a, n, t,* and for the most successful therapeutic decisions at these patients. It is clear that purely symbolic

approaches cannot provide the resolution and precision of this method. For decision support, however, such precision is desirable.

3.2. Similarity search of heart sounds

We have built a software prototype to demonstrate the functionality of the approach also in case of nontrivial vectorial representations. Complex original data like sounds and pictures usually require an appropriate transformation as first step for calculation of feature vectors. In case of heart sounds a wavelet transformation proved to be useful, because it allows analysis of the signal at different scales and times [12][13]. This is implemented in our software. It consists of two program modules:

The first module (used for generation of searchable data) reads a sequence of heart sounds from a .wav file and displays it on the screen (figure 4). One period is reproducibly selected (figure 5) and a Daubechies wavelet transformation is performed. The smoothened absolute values of the transformation coefficients (figure 6) are used for building a searchable feature vector of the sound. This is stored as a VRD in a local database in an internal format.

The second program module (used for similarity search) calculates a VRD from the upload[6] and compares it with the VRDs of the database by calculating the distances to them. It generates the search result which is an ordered list of references to these VRDs (figure 7). The less the distance, the higher is the rank of a VRD in the search result list. This means that VRDs which represent the most similar sounds in the database have upper position in the search result. The main parts of the second program module could be used also for similarity search of other data which allow a vectorial description. This means that one and the same search engine can be used for similarity search in very different applications, if the searchable data are stored in an appropriate way.

Technical remarks: The finest scale for the wavelet transformation has been 2^{11} (figure 8). This resulted in $\sum_{i=1}^{11} 2^i = 4094$ coefficients. This is also the dimensionality of the feature vector, because all coefficients have been used in this application. As distance function we used

$$d(X,Y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

in which n=4094 is the dimensionality, X, Y are the compared vectors, x_i, y_i their coordinates, i.e. the smoothened absolute values of the wavelet coefficients, and $\bar{x}$ the mean of the x_i resp. $\bar{y}$ the mean of the y_i. It is apparent that the distance function contains the correlation coefficient and ensures that the greater the correlation is, the smaller the distance becomes. So the rank of a heart sound is the greater in the search

[6] The module allows upload of original heart sounds as .wav files in compressed form. Because it is specialized to one kind of original data, it can convert these to VRDs before comparison without needing a VSI.

result, the more its wavelet decomposition correlates with the wavelet decomposition of the searched sound.

We want to underline that this implementation is designed for illustration purposes. All 4094 coefficients have been used, so that the time frequency distribution becomes well visible (figure 6). For more efficiency there would be much potential for compression and dimension reduction, because heart sounds have by far not the variability which 4094 coefficients provide. Moreover the process could be fully automated [14]. In contrast to this the workflow of figures 4, 5 and 6 shows the principle and demonstrates a typical half automatic feature selection. Especially in the initial state after introduction of VRDs the half automatic selection of features (or even manual selection like done in figure 3) can have advantages because it can help to keep overview, and introduction of software for feature vector and VRD generation can be accelerated. Later step by step more and more automation is possible, if wished.

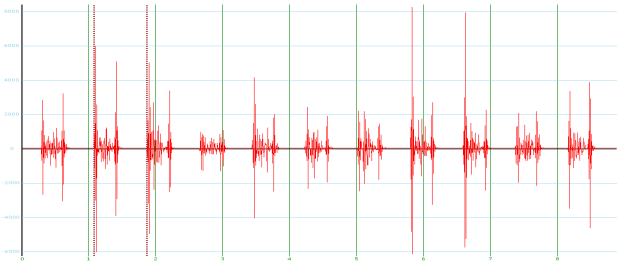

Figure 4. Sounds of a heart with aortic valve stenosis; vertical axis: relative amplitude, horizontal axis: time in seconds. The brown dashed lines have been set reproducibly by a catching algorithm after rough manual prepositioning (semi-automatic process). They show the bordering of a representative period.

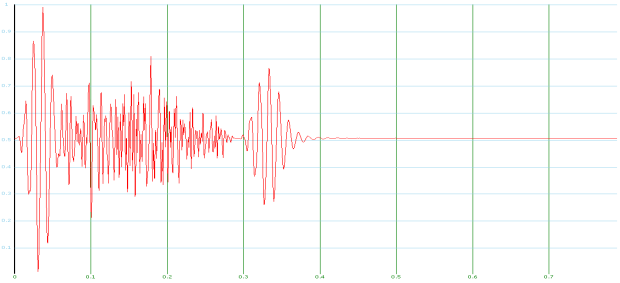

Figure 5. The bordered part of figure 4 is stretched so that one period remains, i.e. the length of the period is standardized. So heart sounds with different period length can be compared.

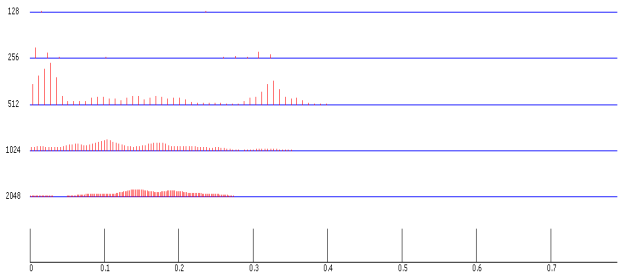

Figure 6. The sound of figure 5 after Daubechies wavelet transform; smoothened absolute values of the wavelet transform coefficients for five different scales. They can be directly used for building a searchable feature vector of this heart sound. Note the delayed appearance of the high frequency part which results from blood flow through the narrowed aortic valve during systole.

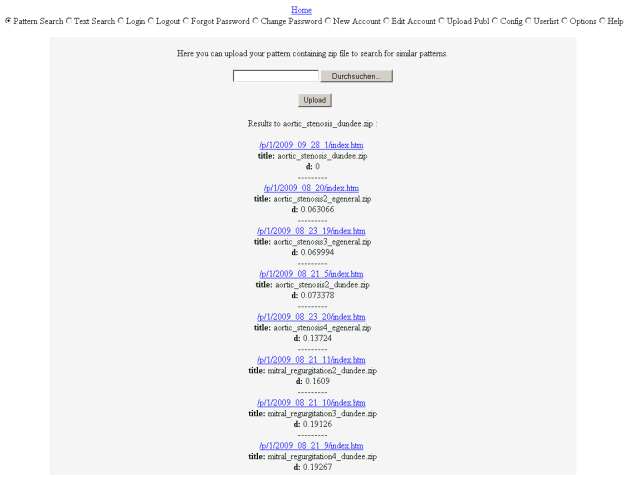

Figure 7. Exemplary output of our prototype. The VRD of the uploaded file represents the heart sound in case of aortic valve stenosis after wavelet transformation as shown in figure 6. Links to resources with most similar VRDs are listed first. The distance d quantifies the deviation, the first link points to a VRD which represents the same sound like the uploaded file, therefore the distance is zero.

4. Discussion

Focus and aim of the approach is *multidimensional decision support*, i.e. support for decisions about the therapy of multidimensional findings ("findings" in the broad sense, meant is the complete state with anamnesis, pretreatment) which depend on several measurable parameters (e.g. medication, laboratory results). The clinician needs just this: He can focus single parameters individually with great flexibility (more flexible than software) and treat these more or less independently of each other, but he cannot reproduce the complex interaction of several findings, i.e. he cannot reproduce complex reality conform multidimensional spaces and their association with certain "optimal"

therapeutic decisions. The proposed approach helps just for this, because it allows representation and localization of multidimensional findings as VRDs and with this the localization and grouping of patients with similar findings, to check therapies and therapeutic results within this group.

According to 2.4 such grouping of patients bases on similarity comparison and similarity search of VRDs (Vectorial Web Search) which is the sequence of two well known workflows:

1. Conventional word based search (for VRDs with the same VSI; this can be simplified using an index: To every VSI an index with all VRDs and their URLs can be collected beforehand).
2. Similarity comparison (of the VRDs' feature vectors, as done in metric databases or by our prototype)

Both workflows are established, so it has to be expected that also the sequence "one after the other" works. Of course the hardware requirement of the global Web application must be considered. For larger amounts of data we made a small experiment to get a hint about the necessary time for comparison of vectors using a C++ compiler on a single PC (Pentium® 4 CPU). The time for calculating 1 million weighted Euclidean distances of vectors with dimensionality 10 in double accuracy was between 0.20 and 0.21 seconds. Such calculation would be necessary if similarity search is done within VRDs with altogether 1 million 10 dimensional feature vectors with serial comparison. Depending on implementation, disk access would require additional time. To minimize the time for comparison, the total dimensionality (the sum of dimensions of all feature vectors) of the searched VRD combination should be minimized. This should be also done because the sparsity increases exponentially with the total dimensionality in case of a constant amount of data, with points tending to become equidistant from one another (“Curse of Dimensionality” [15] [16]). Parallel processing can substantially reduce the time for comparison because it is not necessary to calculate the distances sequentially.

We expect the most important problems in the beginning. Initially investment is necessary for creating the infrastructure, and initially there won't be enough patient records available for decision support. When the number of patient records grows, the situation becomes better, because the concept is designed for large data collections. The resolution of the numeric description grows exponentially with the dimensionality - it is well known that if there are 10 possible values per number, there are 10^k possible values for k numbers. Because usually there are much more than 10 possible values per number, it becomes soon clear that similarity search has a high resolution and can be highly selective even within huge amounts of data - at last a great advantage for decision support when many patient records are available.

As long as there are not many patient records with the selected rough diagnosis, one may be restrained to the most important parameters (low dimensional search), because in this way it is more probable to find patient records with "similar" parameters. In the above example 3.1 one may only provide the number *n* of the vertebra, to find compression fractures near this location. When the database returns more cases near this *n* than one can survey, it is possible to provide additional numbers (dimensions) for a more selective search.

Additionally to direct decision support the precise vectorial descriptions can be useful for skill enhancement (e.g. study of selected diagnostic constellations) and scientific research. Because the feature vectors of VRDs with the same VSI represent points within the same metric space, there is the possibility for "local statistics", i.e. statistics among VRDs whose vectors lay "near" (within a given distance from) a given point. For example it is possible to calculate within a given radius the average of a further dependent variable or the frequency of some incident. Such local statistics is not difficult to interpret, moreover it has the advantage that it can be quickly computed. But especially if there are only a few points (vectors) within the investigated radius, the result could be imprecise due to coincidental fluctuations and it can be more accurate to consider all available points (defined by feature vectors of all VRDs with appropriate VSI) for immediate regression analysis or for calculation of a mathematical model [17] (blue curve in Figure 8).

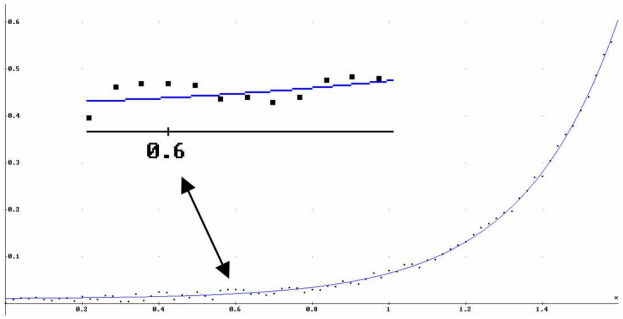

Figure 8. Local statistics can be used if there are many points „near“ to the area of interest. If not, coincidental fluctuations, as shown in the enlarged detail, can become relevant and a mathematical model (blue curve), which takes into account all points, is more accurate.

We show a practical example of how such modeling can be used to "interpolate" at places where data are missing. Suppose patient records are available which contain VRDs whose feature vector describe the medication of a patient in the form (*t, x*) resp (*t, y*), in which *t* is the duration of the medication in days, and the numbers *x* resp *y* represent the daily dosage (e.g. in mg) of medicament X resp. medicament Y. Furthermore VRDs with another VSI provide the blood concentration *z* of a liver enzyme (which indicates damage of the liver). For modeling it is possible to select all patients with $t > t_{min}$ (e.g. $t > 30$ days, to ensure enough medication time), and to calculate from the found data points a function [18] which approximates the concentration *z* of the liver enzyme in dependence of the medication dosages *x* and *y*. A possible resulting function is shown in Figure 9. Because polymedication and drug interaction is an increasing problem (especially in geriatrics), the practical relevance of the approach is obvious.

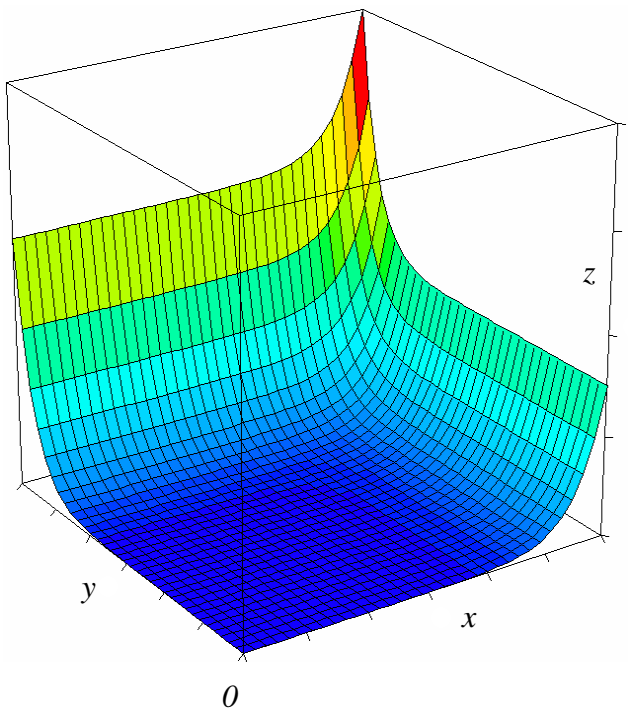

Figure 9. Example of a function which estimates the value of a dependent variable *z* (vertical axis: concentration of a liver enzyme as indicator of liver damage) in dependence of the two variables *x* and *y* (horizontal axes: dosages of two medicaments). It is visible that *z* grows strongly when *x* or *y* exceeds a toxic threshold. Data from the VRDs' feature vectors can be used for calculation of such a function (mathematical modeling).

Of course the data for such modeling could be also fetched from special databases. This requires extra programming. The standardized VRD structure, however, can be used generally by one and the same *evaluation engine* which allows to select VSIs and the numbers of the feature vectors' dimensions for specification of independent and dependent variables.

At last a possibility for anonymization should be mentioned: The patient can select the option that his VRDs can be used for decision support, but only statistically over more than e.g. 5 patients without showing the individual values. It would be possible to calculate and show the VRDs' average (about the therapeutic result) of minimal n (e.g. n=5) patients with similar diagnosis in dependence of therapy.

5. Conclusion

Vectorial representations are precise, directly comparable, and they allow advanced analysis, e.g. similarity calculation using well defined distance functions. Due to the shown advantages of vectorial representation the usage of standardized Vectorial Resource Descriptors (VRDs) is proposed for representation of medical data in patient records.

References

[1] HL7. Standards for electronic interchange of clinical, financial, and administrative information among health care oriented computer systems. http://www.hl7.org/, viewed 2010-02-09

[2] International Health Terminology Standards Development Organisation. SNOMED CT *(Systematized Nomenclature of Medicine – Clinical Terms)*, http://www.ihtsdo.org/snomed-ct/, viewed 2010-02-09

[3] I. Pratt, O. Lemon. Logical and Diagrammatic Reasoning: the Complexity of Conceptual Space. Proceedings of the Nineteenth Annual Conference of the Cognitive Science (1997), Stanford, 430-435

[4] P. Gärdenfors. Conceptual Spaces - The Geometry of Thought. MIT Press, 2000.

[5] P. Gärdenfors. How to make the semantic Web more semantic. In A.C. Vieu and L. Varzi, editors, Formal Ontology in Information Systems, IOS Press (2004), 19–36.

[6] S. Dietze, J. Domingue, W. Orthuber. Blending the Physical and the Digital through Conceptual Spaces, Workshop: OneSpace 2009 at Future Internet Symposium (FIS) (2009), Berlin, Germany.

[7] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh ed. Feature Extraction, Foundations and Applications. Studies in Fuzziness and Soft Computing, Springer, Berlin, Heidelberg, 2006.

[8] Wikipedia, http://en.wikipedia.org/wiki/Normed_vector_space, viewed 2010-02-09

[9] Wikipedia, http://en.wikipedia.org/wiki/Metric_space, viewed 2010-02-09

[10] C. Bizer, R. Cyganiak, T. Heath. How to Publish Linked Data on the Web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/20070727/, viewed 2010-02-09

[11] Linked Data. Connect Distributed Data across the Web. http://linkeddata.org/, viewed 2010-02-10

[12] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inform. Theory, **36** (1990), 961-1005.

[13] B. El-Asir, L. Khadra, A.H. Al-Abbasi, MMJ. Mohammed. Time frequency analysis of heart sounds. Proc IEEE. **2** (1996), 553–558.

[14] C. Faloutsos, M. Ranganathan, Y. Manolopoulos. Fast Subsequence Matching in Time-Series Databases. ACM SIGMOD Int.Conf. on Management of Data (1994), 419-429.

[15] S. Berchtold, C. Böhm, H.-P. Kriegel. The Pyramid-Tree: Breaking the Curse of Dimensionality. SIGMOD Conference (1998), 142-153.

[16] P. Indyk, R. Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In Proc. of the 30th Symposium on the Theory of Computing (1998), 604-613.

[17] M.W. Kattan, D.H.Y. Leung, M.F. Brennan. Postoperative Nomogram for 12-Year Sarcoma-Specific Death. Journal of Clinical Oncology, **20**(3) (2002), 791-796.

[18] L. Fang, D.C. Gossard. Multidimensional curve fitting to unorganized data points by nonlinear minimization. Computer-Aided Design **27** (1) (1995), 48–58.

[19] LOINC®. Logical Observation Identifiers Names and Codes. http://loinc.org/, viewed 2010-02-10