Towards Standardized Vectorial Resource Descriptors on the Web

Wolfgang Orthuber[1]

[1]Department of Orthodontics
University Clinic Schleswig-Holstein
Arnold Heller Str. .3, House 26
24105 Kiel / Germany
orthuber@kfo-zmk.uni-kiel.de

Stefan Dietze[2]

[2]Knowledge Media Institute
The Open University
MK7 6AA
Milton Keynes / United Kingdom
s.dietze@open.ac.uk

Abstract: Resources with quantitative properties, e.g. measurable resources or sources for feature extraction (e.g. fingerprints), play an important role, particularly in scientific areas such as Life Sciences, the medical domain and nature sciences. In this paper we propose similarity-based representation of resources using so called Vectorial Resource Descriptors (VRDs) on the Web. The VRDs are standardized data structures which build the basis of Vectorial Web Search. Every VRD contains a feature vector and a Vector Space Identifier (VSI), and further data. In contrast to conventional keyword search, which requires matching of free text, Vectorial Web Search is well defined similarity search of numeric data. Users provide a VRD, or only the searched numeric data (i.e. the feature vector, as sequence of numbers) together with the VSI. The VSI is a HTTP URI which identifies the vector space of the feature vector, and which points to a standardized Vector Space Descriptor (VSD). So the valid distance function and the meaning of every dimension (number) of the feature vector is known by the system. For quantification of similarity the (in the VSD specified) distance function of the chosen vector space is used. The smaller the distance, the greater is the similarity of a VRD, and the higher is its rank in the search result.

1 Introduction

Due to the enormous amount of data on the Web there is an increasing need for information integration. This lead to the Semantic Web [Be01] and the Linked Data approach [Be06][Bi09] which aim at meshing together meaningful machine readable data on the Web. The mesh defines neighborhood: data which are directly linked together are neighboring in the mesh. While that approach requires to establish links explicitly, there is a well known mathematical concept for definition of neighborhood which allows implicit inference of neighborhood: similarity in a vector space.

While similarity computation usually exploits linguistic or structural similarities, we propose vectorial representation of resources to facilitate computation of similarities by means of distance functions (metrics) in shared vector spaces according to individual requirements. A vector space is much finer than any mesh of links. In our view, vectorial search complements but does not substitute traditional search approaches. In applications

in which it is usable, e.g. representation of measurements or other Quantifiable Resources (QR, see 2.1), it is a very efficient extension. A compact data structure called "Vectorial Resource Descriptor" (VRD) can serve as connector between the mesh of linked information and vector spaces, it is introduced below (chapter 3).

The usage of vector spaces for data integration is already topic of research, e.g. the Conceptual Space approach. Conceptual Spaces (CS) [Ga00][Pr97] follow a theory of describing entities at the conceptual level in terms of their natural characteristics similar to natural human cognition in order to avoid the symbol grounding issue [Di09a]. CS 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 semantic 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, speech and various types of fingerprints 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 [Gu06]. Vectors, embedded in VRDs, can describe all quantitative (and with this also all measurable) properties. This is the basis of *Vectorial Web Search* which means web based similarity search for quantitative data in standardized vectorial representation, using the well known metric space approach [Ze05]. In contrast to conventional keyword search [Br98], which requires matching of words, Vectorial Web Search is similarity search which exploits the ordered nature of quantitative (numeric) data. This is even necessary: quantitative descriptions are often very precise, having numeric representations with many digits and/or many dimensions, that not even one 1:1 match can be found worldwide, but many similar matches. These can be sorted in well defined way. In this paper we describe the application of vector spaces for representation of quantitative properties and data integration in general and propose a data structure and framework for efficient implementation of vectorial similarity search in the current Web infrastructure.

2 Approach

Precondition for vectorial (numeric) representation is reproducible quantification. To be suitable for a vectorial representation, a web resource must have one or several quantitative properties, i.e. one or several attributes which each have an inherent directed order (from "little" to "great"). We will call such a resource "Quantifiable Resource" (QR). One can say that “all possible” resources are quantifiable, because also quantification of the property “non existing” or “existing” can be done by associating both possibilities with numbers (e.g. with “0” or “1”). In many applications, however, fine distinction is required. When defining a vectorial (numeric) representation, aim is to find most important (i.e. decision relevant) ordered properties of a resource, so that most important variants of the resource can be represented by as few as possible numbers and small changes of the resource are mapped to small changes of the numbers.

Examples for QR: Human diagnostic parameters and measurements are attractive QR, because their standardized quantification allows world wide grouping of medical records of patients with similar parameters, to find the most efficient treatment for this group [Or06][Or10]. Some further examples for QRs are: Descriptors of personal profiles, digital representations of various items of daily life and their direct recognition (e.g. melodies, faces etc.), measurement and classification schemes of products and services and with this more individual and reproducible adaptation of products and services to the customers needs, semantic web services, concepts or situations, defined as elements (vectors) in conceptual spaces [Di09], numeric quantitative data, e.g. GPS coordinates, higher level sensor data, e.g. measurements of environmental or climate parameters, results of feature extraction, measurements and classifications in all areas of daily life.

3 Vectorial Resource Descriptors (VRDs)

Our proposed standardized data structure for representation of a QR on the Web is called *Vectorial Resource Descriptor* (VRD). It contains:

1. The *Identifier of the QR* (QRI). It is a HTTP URI which points to the resource.
2. The *Vector Space Identifier* (VSI). It is a HTTP URI which points to the *Vector Space Descriptor* (VSD).
3. The *feature vector* (usually a sequence of numbers). It represents the quantitative properties of the resource.

Additionally it can contain:

4. Auxiliary data, e.g. date, keywords.

The VSI is a HTTP URI [Bi09] which uniquely identifies the vector space and with this the meaning of every dimension (number) of the feature vector. The feature vectors of all VRDs with the same VSI are elements of the same vector space and with this comparable. Similarity search is done within this space. Being a HTTP URI, the VSI not only identifies the content of the feature vector, it simultaneously points to the Vector Space Descriptor (VSD).

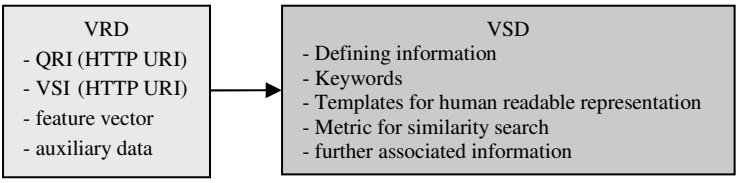

Figure 1. Contents of the VRD and VSD. The VRD's *Vector Space Identifier* (VSI) points to the VSD which provides important information about all VRDs with this VSI.

The VSD provides all necessary information about the vector space, particularly about the metric (distance function) for comparison, definitions of dimensions, templates for human readable representation of instances and links to further related Web content.

VRDs and VSDs are machine readable and can be embedded into the semantic Web as Linked Data [Bi09]. They extend the mesh of linked data by vector spaces, are uniformly

comparable and searchable, and it is possible to share the work for their definition and generation among all domain name owners.

The Resource Description Framework (RDF) [W3C04] can be used for VRD and VSD representation [Or10a]. If wished, a VRD can be also embedded in HTML files via RDFa. The feature vector is usually a sequence of numbers. This sequence can generally represent mathematical objects (e.g. coefficients, matrices of linear operators for conversion of vectors, tensors) whose definition can be done in the VSD. An even more expanded definition is possible: More generally the feature vector is a data structure which is defined in the VSD, usually designed for efficient comparison, so that the result of the comparison is a number. This convention makes the VRDs adaptable to the programmers' needs.

Example: Lets assume that a manufacturer of *round tables* owns the web address http://a.com and wants that his products are searchable according to their main geometric features height and diameter. Therefore he places a VSD on his website on the Web address http://a.com/round-table.rdf. The VSD holds the defining information of the vector space: It is two dimensional and contains feature vectors $V = (v_1, v_2)$ in which v_1 = height of the round table in cm and v_2 = diameter of the round table in cm. Keywords are: "Round Table". Template for human readable representation is: "round table with (D1) cm height and (D2) cm diameter" in which (D1) is placeholder for dimension 1 and (D2) is placeholder for dimension 2 of the feature vector. The metric of similarity search is the (by reciprocals of the standard deviations) weighted Manhattan distance *dws*: Let $V = (v_1, v_2)$ and $U = (u_1, u_2)$ denote two feature vectors, then $dws(V,U) = \left(\left| \frac{v_1 - u_1}{sd_1} \right| + \left| \frac{v_2 - u_2}{sd_2} \right| \right)$, in which sd_1 is the standard deviation of dimension 1 (of the height in cm) and sd_2 is the standard deviation of dimension 2 (of the diameter in cm). Reciprocals of standard deviations can be used as default weighting factors if further information is missing. This compensates inter alia the influence of the chosen unit.

After this the manufacturer (and other people) can add on the Web to every product description of a round table a VRD with QRI = Web address of the product description, VSI = Web the address of the VSD = "http://a.com/round-table.rdf", feature vector $V = (v_1, v_2)$ in which v_1 = height of the round table in cm and v_2 = diameter of the round table in cm. This makes the product descriptions of round tables searchable by height and diameter using vectorial web search (chapter 4).

4 Vectorial Web Search

The VRDs are the fundament of Web-scale vectorial similarity search (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 *V* 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*)
- 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 1) the search engine calculates distances between the feature vector *V* and the feature vectors of the collected VRDs and sorts them according to distance using e.g. introsort [MU97] or another algorithm which allows fast parallel processing.
- 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 *V*, among all VRDs those with this *VSI* are selected (an index can accelerate this) and after optional further confinement used for comparison, i.e. the distances between their feature vectors and the searched feature vector V are calculated. In the search result the rank of VRDs and (via QRI, as described in chapter 3) associated resources is the higher, the less the distance is, i.e. most similar resources are listed first, using the proved and tested metric space approach [Ze05]. Concerning technical limitations of the approach we want to mention the curse of dimensionality. Low dimensional vector spaces are preferable because they can be handled much more efficiently than high dimensional spaces. Therefore it can become necessary that the search engine introduces an upper limit of dimensionality. Practical limitations can result from inefficient or redundant definitions of vector spaces. Therefore the VSD contains keywords, so that is possible to search for existing definitions to a topic before making a new definition.

5 Conclusions

Conventional language-based Web search does not facilitate similarity search of resources with certain quantitative properties, e.g. measurements. With this article we introduced a framework for standardized Vectorial Resource Descriptors (VRDs) on the Web, so that Vectorial Web Search, and with this similarity search of resources with quantitative properties can be realized in efficient way.

If there is enough support, further advanced evaluation of the approach will be part of our future work. The next important step is an online prototype. The concrete RDF format of VRDs and VSDs is currently topic of discussion [Or10a], comments are welcome. Advanced implementations could support flexible queries using e.g. SPARQL query language and/or realize a first practical medical or commercial application.

Bibliography

[Be01] Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. Scientific American 284(5) 2001, pp. 34-43.

[Be06] Berners-Lee, T. Linked Data. W3C Design Issues, 2006.

[Bi09] Bizer, C., Cyganiak, R., Heath, T. How to Publish Linked Data on the Web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/, viewed 2009-05-27.

[Br98] Brin, S., Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proc. of 7th WWW Conference 1998, pp. 107-117.

[Di09] Dietze, S., Gugliotta, A., and Domingue, J.: Exploiting Metrics for Similarity-based Semantic Web Service Discovery, IEEE 7th International Conference on Web Services (ICWS), 2009, Los Angeles, CA, USA.

[Di09a] Dietze, S., Orthuber, W., and Domingue, J. Blending the Physical and the Digital through Conceptual Spaces, Workshop: OneSpace 2009 at Future Internet Symposium (FIS) 2009, Berlin, Germany.

[Fo02] Fodor I.K. A survey of dimension reduction techniques, US DOE Office of Scientific and Technical Information, 2002.

[Ga00] Gärdenfors, P. Conceptual Spaces - The Geometry of Thought. MIT Press, 2000.

[Gu06] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., ed. Feature Extraction, Foundations and Applications. Studies in Fuzziness and Soft Computing, Springer, Berlin, Heidelberg, 2006.

[MU97] Musser, D., Introspective Sorting and Selection Algorithms. Software: Practice and Experience (Wiley) 27(8), 1997, pp. 983-993.

[Or06] Orthuber, W. Vectorial Web Search. http://www.orthuber.com/wpa.htm, since 2006-06-02.

[Or08] Orthuber, W., Fiedler, G., Kattan, M., Sommer, T., Fischer-Brandies, H.: Design of a global medical database which is searchable by human diagnostic patterns. The Open Medical Informatics Journal 2, 2008, pp. 21-32.

[Or10] Orthuber, W., Papavramidis E., Standardized Vectorial Representation of Medical Data in Patient Records, Medical and Care Compunetics 6, pp. 153-166.

[Or10a] Orthuber, W. Proposal for VRD and VSD representation. http://www.orthuber.com/wvstandard.pdf, since 2010-06-29.

[Pr97] Pratt I., Lemon O.: Logical and Diagrammatic Reasoning: the Complexity of Conceptual Space. Proceedings of the Nineteenth Annual Conference of the Cognitive Science 1997, Stanford, pp. 430-435.

[W3C04] W3C: RDF/XML Syntax Specification (Revised); Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/, viewed 2008-08-12.

[W3C07] W3C: XQuery 1.0: An XML Query Language; W3C Recommendation 23 January 2007. http://www.w3.org/TR/xquery/, viewed 2009-06-03.

[W3C08] W3C: SPARQL Query Language for RDF; W3C Recommendation 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/, viewed 2009-06-03.

[Ze05] Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search. The Metric Space Approach. Series: Advances in Database Systems, Vol. 32., Springer, Berlin, Heidelberg, 2005.