General approach to similarity search of resources with numeric features on the web

Wolfgang Orthuber, 3 March 2013

Abstract

It will be shown and online demonstrated how it is possible to realize systematically on the web similarity search of resources with numeric (quantitative) features using

- User defined metric Feature Vector Spaces (FVSs)
- Vectorial Resource Descriptors (VRDs)

Every VRD is element of a FVS. It maps a resource into this FVS and makes it accessible to metric similarity search by its numeric features. The VRD contains all necessary numeric data and a URI called "topic" which identifies the FVS. Because missing FVSs can be quickly defined by the users, VRD based description is a general approach to realize similarity search of resources with numeric features. VRD search consists (without extension) of 2 systematic steps:

1. Selection of the FVS (e.g. using word based search of the topic)
2. Similarity search of a VRD by its numeric features within the selected FVS or a part of it.

The presented online implementation shows that VRD based description and search has great potential. To store FVSs and VRDs as open data on the web it is recommendable to start a working group for introduction of a web standard for worldwide valid FVS definitions and VRDs.

Introduction

Suppose we have a resource with certain numeric features and want to store it on the web, that search engines can find it systematically by its numeric features. Though this is a very frequent and important problem, up to now it is not solved in general. There is no general approach to it.

In this presentation we describe and practically demonstrate a general approach and solution of this problem using

- User defined metric Feature Vector Spaces (FVSs)
- Vectorial Resource Descriptors (VRDs)

Every FVS is identified by an URI called "topic". VRDs are elements of the FVSs and accessible to numeric similarity search. So they can make resources searchable by their numeric features. This is already implemented on a local database and online since July 2012.

We concretely demonstrate the approach online using the VRD search engine http://nummel.com/

Approach

First we demonstrate the VRD search procedure on the existing database. The general approach (VRD search on user defined FVSs) is shown by the exemplary topic "box". We have a box (with numeric features e.g. Price, Height, Length, Width, Weight, Volume) and want to make it searchable by its numeric features. For this we first try to find a FVS with the topic "box" or something synonymous like "case", "chest" by conventional text search.

- If we find it, and if the FVS contains all numeric dimensions which are important for us, we can (after login) immediately provide a VRD with these numeric data, which will be demonstrated.
- If we find it, but some dimensions are missing, we can inform the owner of the FVS about the missing dimensions and hope that the owner will append these. The owner should be interested to do this to keep the own FVS attractive.
- If a FVS with synonymous topic and satisfying definition is finally not available, we can consider the definition of a new FVS (e.g. with dimensions Price, Height, Length, Width, Weight, Volume). This will be also demonstrated. Then it is possible to provide an appropriate VRD of the box which we want to make searchable. After this also other users can provide VRDs of this new FVS and so make their boxes accessible to similarity search using the dimensions of this FVS (e.g. Price, Height, Length, Width, Weight, Volume).

So it has been shown how VRD based description can make resources with any numeric features (or dimensions) accessible to numeric similarity search. If wished also other possibilities can be demonstrated, e.g. usage of numeric dimensions for representation of keywords or user-definable lists, partial search (using min-max ranges for certain dimensions), graphical representation of selectable dimensions in the search result, immediate statistical evaluations of FVSs, XML export of FVS definitions and VRDs. Operating instructions for the implementation are available in http://nummel.com/intro.pdf , where also the used distance functions for sorting the search result and details about the used data structures are described.

Some possible questions and answers

1. *Meanwhile there is much progress in conversion of unstructured data into structured form. Do we really need a standard for structured input of numeric data?*

Answer: Even if the techniques for conversion of unstructured to structured Data would be perfect, they can only extract existing numeric data. Often important numeric data are missing, because the writer does not know well the expectations and interests of the reader. The list of interesting numeric data depends on the topic. For example the list of interesting numeric data in case of the topic "traffic-connection" largely deviates from the list of interesting numeric data in case of the topic "disc-herniation". Therefore it is efficient to define in dependence of the topic a (usually nonbinding) table of interesting numeric data and so define a topic specific interface between writer and reader. This is the FVS which is identified by the URI called "topic". By definition of international FVSs the motivation can be given to publish completely new (topics with) numeric data on the web to make them generally accessible to similarity search (and more).

2. *Why has VRD based description and search higher resolution than word based description and search?*

Answer: For a word which is (more than grammatically) different from other words we need an extra definition. But for all (different) VRDs which belong to the same FVS we need only one definition - the definition of the FVS. This definition is usually also more precise than the definition of a word, and can be made for all topics which are of interest for the users. There can be much more different VRDs (even in only one FVS) than there are different words. Therefore VRD based description and search has much higher range and resolution than word based description and search.

3. *How are VRDs related to linked data?*

Answer: In the current implementation FVS definitions and VRDs are accessible via HTTP URIs and contain data structures called "keycomment" which each contain short optional text and one or more keywords with optional HTTP links. Additional to the possibility for *connection* via HTTP URIs every VRD has by definition a well defined *similarity relation* at least to all other VRDs of the same FVS. A web standard could introduce a clear relation between the topic of a FVS and its HTTP URI (a new domain name end can be useful for this). Such a standard could allow to make FVS definitions not only new but also partially or fully from compositions of subvector definitions which represent other (smaller) already existing FVS definitions using their (HTTP URIs or) topics and so get also a similarity relation between all VRDs with a common (HTTP URI of a) subvector definition. It may make sense to derive elementary low dimensional FVS definitions and dimension definitions (for compositions of higher dimensional FVSs) directly from already existing ontologies of linked data.

4. *The web covers a large range of topics. How can VRDs efficiently represent any user defined data?*

Answer: To cover the large range of topics on the web it is necessary to use the working capacity of all web users. They can make useful FVS definitions about all topics which are of common interest. They can also write software which generates and utilizes VRDs of these FVSs. Every web user can become owner and creator of a FVS definition and so not only define the numeric content of every VRD of this FVS, the owner can define also the meaning of the HTTP links of the VRD. It is efficient to store directly in the VRD only the searchable numeric data (those which are quickly comparable with other VRDs of this FVS to quickly calculate a well defined distance for similarity comparison) and to make further (large amounts of) data accessible via HTTP links. The owner of a FVS can for example define that the first HTTP link of every VRD of this FVS points to a topic specific dataset (e.g. a picture, song, data generated by software of the FVS owner) and the numeric content is a searchable topic specific feature extraction of this dataset. If the feature extraction is appropriate, the VRD makes the dataset available to topic specific similarity search by all VRD search engines.

5. *How can we quickly detect equivalent definitions of FVSs and find the most relevant among them?*

Answer: In case of doubt we can ask search engines using specific text search within the keywords of FVS definitions to get a list of FVSs which cover a certain topic or something similar. The list can be ordered e.g. by the size of the FVSs (the count of contained VRDs). This can be used to quickly find the most frequently used FVS among equivalent FVS definitions. A check of their definitions can help to find the best fitting FVS for our individual application.

6. *How can we guarantee reliability of FVS definitions?*

Answer: To guarantee reliability, FVS definitions must be stored in a reliable database (perhaps at ICANN), which is open for read and which accepts only compatible changes, e.g. expansions of a FVS by additional dimensions, no deletion or change of already defined dimensions. Every dimension should have a short name (which is different from the names of other dimensions of this FVS) and a concise and clear definition using conventional text, not only a hyperlink. If the owner of the FVS definition is no longer available, the FVS could be marked as free for sale, and owners of other FVS definitions which use it could be informed about this. A new owner can make all compatible changes, e.g. modify free parts of the definition, mark dimensions as outdated and add further dimensions. If no one is interested in owning this free FVS, the database should keep the FVS definition stored as long as necessary. It is possible to store different FVS definitions with the same topic, if they have non-overlapping periods of validity.

7. *Can you give examples, how VRD based description and search can be used for decision support?*

Answer: The probably simplest possibility is to enter wished numeric data (e.g. "0" as Price, or GPS coordinates) and make similarity search (usage of the "sim" field in the shown implementation). In the search result those resources whose data are nearest to the own wish are listed first and can be checked. If only resources with certain numeric data are interesting, similarity search can be restricted to a part of the FVS by predefinition of ranges for certain dimensions via usage of the "min" and/or "max" fields in the search window. It is also possible to check immediate statistics of decision relevant dimensions (e.g. average blood pressure) in case of selected ranges of other influencing dimensions (e.g. doses of medication). The smaller the standard deviation of the checked decision relevant dimension, the better influencing dimensions have been covered. Statistics can be also used to detect dependencies. Of course statistical results are the more reliable, the larger the database is. This is one of many arguments for open worldwide FVSs.

Conclusion

It has been shown that VRD based description and search has great potential and can become an important addition to word based description and search. Up to now FVSs definitions and VRDs must be stored in the database of the VRD search engine. To store FVSs and VRDs as open data on the web it is recommendable to start a working group for introduction of a web standard for worldwide valid FVS definitions and VRDs.