Hierarchies Measuring Qualitative Variables

Serguei Levachkine and Adolfo Guzmán-Arenas

Centre for Computing Research (CIC) - National Polytechnic Institute (IPN)
UPALMZ, CIC Building, 07738, Mexico City, MEXICO
palych@cic.ipn.mx, a.guzman@acm.org

Abstract. Qualitative variables take symbolic values, such as *hot, shoe, Europe or France*. Sometimes, the values may be arranged in layers or levels of detail. For instance, the variable *place_of_origin* takes as level-1 values *European, African*... as level-2 values *French, German*... as level-3 values *Californian, Texan*... The paper describes a hierarchy, a mathematical construct among these variables. The confusion resulting when using a value instead of another is defined, as well as the closeness to which object *o* fulfills predicate *P*. Other operations among and properties of hierarchical values are derived. Hierarchies are compared with ontologies. Hierarchies find use in measuring linguistic relatedness or similarity. Hierarchical variables abound and are commonly used, often with suggestive string values, without fully realizing or exploiting its properties. We deal with arbitrary hierarchies. Examples are given.

1 Introduction

A datum is a relational entity. Nothing is a datum itself; i.e. a context[1] is required. This thesis is especially true for qualitative data. Notice that many works on qualitative data processing usually omit the problem under consideration context. In contrast, we use the hierarchies to measure similarity and dissimilarity between qualitative values, attempting to keep the context. To some extent, the notion of hierarchy provides an adequate tool for qualitative data analysis, processing and classification, because the hierarchies encapsulate the (sometimes ordered) relations between partitions of the dataset and therefore easily maintain the problem context.

What wearing apparel do we wear for rainy days? *Raincoat* is a correct answer; *umbrella* is a close miss; *belt* a fair error, and *typewriter* a gross error. What is closer to

[1] The notion of context depends on particular environment (subject domain, representation space...) into which the data are embedded. In turn the relatedness between data elements depends on the context. For example, the *pale* and *beige* could be much closed (to indistinguishable) in one context while in another they should be far distanced. Subsequently this paper concerns not only with the problem to appropriately define the closeness of data elements but also to take into consideration the properties of the representation space. This can be observed as a context-oriented approach to qualitative data processing (see also §1.3).