Decision Support Systems 62 (2014) 144–157

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A characterization of hierarchical computable distance functions for data warehouse systems

Matteo Golfarelli *, Elisa Turricchia

DISI, University of Bologna, Italy

ARTICLE INFO

Article history:
Received 22 August 2013
Received in revised form 10 March 2014
Accepted 30 March 2014
Available online 12 April 2014

Keywords:
Testing
Similarity measures
Hierarchical data
Categorical data

ABSTRACT

A data warehouse is a huge multidimensional repository used for statistical analysis of historical data. In a data warehouse events are modeled as multidimensional cubes where cells store numerical indicators while dimensions describe the events from different points of view. Dimensions are typically described at different levels of details through hierarchies of concepts. Computing the distance/similarity between two cells has several applications in this domain. In this context distance is typically based on the least common ancestor between attribute values, but the effectiveness of such distance functions varies according to the structure and to the number of the involved hierarchies. In this paper we propose a characterization of hierarchy types based on their structure and expressiveness, we provide a characterization of the different types of distance functions and we verify their effectiveness on different types of hierarchies in terms of their intrinsic discriminant capacity.

1. Introduction

Effectively measuring the similarity, or symmetrically the distance, between objects is a generic research issue whose solution changes depending on the characteristics of the involved data. In this paper we focus on the distance functions for categorical and hierarchical attributes. On the one hand categorical data are intrinsically unordered and this limits the possibility of defining an effective distance measure [1]. On the other hand, the presence of hierarchies of concepts enriches the description of the objects and provides a tool for partially restoring an ordering between them.

Categorical and hierarchical attributes are particularly relevant since they are one of the building bricks for multidimensional data spaces, where a single data object is described by several of this attributes. We will refer to such data spaces as *Hierarchical Non-Ordered Discrete Data Spaces — HNODDSs* that extend the acronym NODDS coined in [2]. Similarity search for HNODDSs is becoming increasingly important for several application domains such as multimedia information retrieval, statistical data analysis, scientific databases and data mining [3]. In particular, HNODDSs are at the core of data warehouses that are huge multidimensional repositories used for statistical analysis of historical data [4]. In a data warehouse events are modeled as multidimensional cubes where cells store numerical indicators while dimensions describe the events from different points of view. For example, a SALE cube would store the quantity sold and the corresponding sale amount; each sale would be defined by a CITY, a PRODUCT and a DATE. All the attributes are categorical, except DATE. Fig. 1 shows an example for the CITY hierarchy. Identifying the distance between a couple of cube cells/events has several applications. For example a user would benefit in automatically retrieving events that are similar to those she is currently browsing. On the other hand, the identification of events that are very dissimilar from all of the others (i.e. outliers) would be very useful to both end-users (e.g. detection of anomalous behaviors) and system administrators (e.g. detection of erroneous data during the data warehouse ETL).

Distance functions for hierarchical data have been proposed in many different contexts. [5] defines a set of measures for assessing the distance between words exploiting a taxonomy of concepts, [6] shows that these methods work well on a large set of a taxonomies of medical terms. [7] analyzes different similarity criteria and tests them in the area of data warehousing on user labeled data to understand which is the one that matches the human perception of similarity at best. All the previous papers state that, when categorical and hierarchical attributes are involved the least common ancestor — LCA between values at the lowest level of the hierarchy plays a crucial role in defining a user-meaningful distance function. [7] uses the term hierarchical computable for such type of distances. Please note that distances included in this class do not keep information coming from a corpus into account (e.g. the frequency a particular city has in the data set). All the previous papers investigate the effectiveness of hierarchical computable distances that is measured a posteriori typically through a manual tagging of the results. The authors also debate on the weakness of their measures, but limit their discussion to empirical considerations, failing to provide a well-founded answer due to the lack of an analytic model.

Effectiveness of hierarchical computable distances depends on both the LCA's features used to define them (e.g. the level of the LCA or the

* Corresponding author at: DISI - Via Sacchi, 3 - 47521, Cesena (FC), Italy.
E-mail addresses: matteo.golfarelli@unibo.it (M. Golfarelli), elisa.turricchia2@unibo.it (E. Turricchia).

http://dx.doi.org/10.1016/j.dss.2014.03.011
0167-9236/© 2014 Elsevier B.V. All rights reserved.

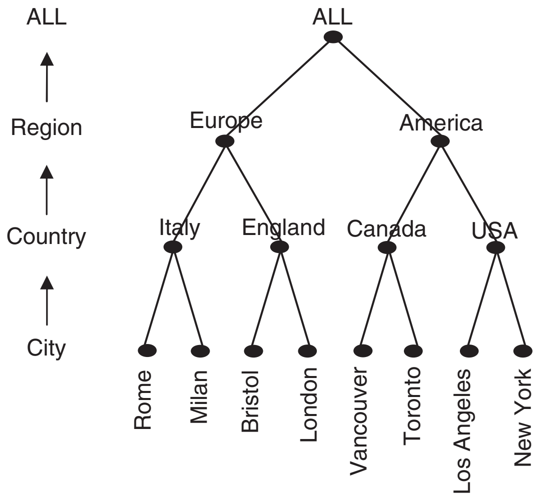

Fig. 1. A hierarchy and its instances.

distance from it or the number of data objects subsumed by the LCA) and on the hierarchy structure. In particular the quantity of information coded in a hierarchy, and consequently the level of precision of a hierarchical computable distance, changes with its depth and size.

When multidimensional objects are involved a clear understanding of the characteristics that influence the effectiveness of distance functions is even more crucial since the presence of several hierarchies, possibly with different characteristics, could make correctly capturing similarity more complex or even impossible. As shown in the paper, HNODDSs are subject to the so called curse of dimensionality, thus it is important to know to which extent similarity queries (e.g. range and nearest neighbor queries) make sense.

To the best of our knowledge no paper in the literature proposes a precise characterization of the effectiveness and limitations of hierarchical computable distances in terms of the structure and size of the hierarchy. In this paper we move a first relevant step in this direction, by providing:

- a characterization of hierarchical and categorical attributes according to the structure of their hierarchies (see Section 4);
- a characterization of hierarchical computable distances based on the type of information they consider (see Section 4);
- a probabilistic model that analytically defines the discriminant capabilities of the distance functions. The model works in the mono-dimensional case (see Section 5.1) as well as in the multi-dimensional one (see Section 5.2). Together with the model some indicators are provided to evaluate the discriminant capacity of hierarchies, distance function and HNODDS: some of them are original, while others are enabled in HNODDSs by the model (e.g. Intrinsic dimensionality).
- a set of experimental results, carried out on both real and synthetic data sets, that provide an empirical evaluation of the effectiveness and efficiency of hierarchical computable distances, both for a single categorical and hierarchical attribute and for HNODDSs (see Section 6).

Our contributions are valuable tools for both practitioners and researchers involved in defining similarity functions or in designing hierarchy structures. Indeed, no techniques are currently available to estimate a priori the capabilities/limits of a given similarity measure when applied on a given categorical hierarchy or a given HNODDS. All the evaluations are carried out a posteriori through subjective user's feedbacks. Conversely, the characterization we propose defines a formal framework for such evaluation, gives some rules of thumb for coupling measures and hierarchies and it finally provides indicators (original and non original) for evaluating, at design time and on an objective basis, the discriminant capacity of distance functions and hierarchies.

2. Related literature

The definition of similarity and distance functions is a wide research area that covers several application domains ranging from information retrieval to multimedia applications. Each domain requires a specific definition of distance that should exploit the characteristics of the involved data and should be meaningful for the application users as well. When data are numeric, distances can be derived starting from the classic Euclidean distance function or the Minkowski one that generalizes it. Several more sophisticated concepts have been devised in the literature [8], for example the Mahalanobis distance improves over Euclidean one since it is scale-invariant and it takes into account the correlations of the data set. Recently, with the increasing complexity of data entities across various domains, an increasing interest has been raised by nonmetric distances. Although this type of distances allows the modeling of complex distance concepts, they cannot exploit the nice topological properties of the metric ones and thus require ad-hoc techniques for efficiently running a similarity search [9].

When data compared are categorical, they are typically modeled as sets of elements (e.g. Overlap Coefficient, Jacard's Coefficient). The distance between two collections is then computed on the basis of their set or bag intersection [8]. Based on this idea several more advanced approaches have been proposed for categorical attributes (see [1] for a comprehensive analysis). Basically they exploit the cardinality of the different elements (e.g. Eskin [10]) or the frequency of the considered attribute values (e.g. Inverse Occurrence Frequency — IOF [11], Goodall [12]) to differentiate the similarity between couple of objects. Such information allows the computation of accurate similarity measures.

Intersection-based measures do not accurately capture similarity when data are sparse or when there are known relationships between items. This is the case for hierarchically organized domains whose similarity has been studied first in [13]. In that paper the authors generalize the *cosine-similarity* to take the hierarchy into account. In the generalized-cosine-similarity — GCSM two unit vectors (i.e. two vectors modeling a single leaf of the hierarchy), are no more perpendicular if such leafs share a common ancestor in the hierarchy. In other words, the closer the common ancestor the more similar the two elements. The idea of computing the path through the hierarchy for exploiting hierarchical information has been adopted in many other approaches, for example in [14] it is used to model the semantic similarity in an ontology. A similarity/distance measure is also necessary in outlier detection applications. To the best of our knowledge the only approach that poses the problem in a HNODDS like context and exploits aggregated information is [15] that models criminal incidents as multidimensional cells whose dimensions describe the incident (e.g. type of incident, weapon-used). The distance function used to compute the extremeness of a cell is based on the frequency of events in the current cell compared to those of its neighborhood (i.e. all the possible aggregations of the current cell on the available dimensions). Unfortunately the approach does not use hierarchies: each dimension is characterized by one attribute and aggregations are obtained by considering or dropping it.

As to effectiveness of distance functions for hierarchical data, an interesting paper is the one by Baiakousi et al. [7]. The paper experimentally assesses the effectiveness of some known similarity measures based on a user study. Interestingly the study confirms that the functions that seem to fit the user needs at best are those choosing as the closest point the one with the shortest path through the hierarchy. Unfortunately the study does not evaluate the discriminant capacity of the proposed measures and it considers only data with a very limited dimensionality.