Linköping Electronic Articles in
Computer and Information Science
Vol. 5(2000): nr 4

Using the Real Dimension of the Data

Christian Zirkelbach

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/2000/004/

Posted on March 8,
formally published on December 5, 2000
by Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in
Computer and Information Science, Vol. 5(2000): nr 4.
http://www.ep.liu.se/ea/cis/2000/004/. *December 5, 2000.*

This URL will also contain a link to the author's home page.

Abstract

This paper presents a method for extracting the real dimension of a large data set in a high-dimensional data cube and indicates its use for visual data mining. A similarity measure structures a data set in a general, but weak sense. If the elements are part of a high-dimensional host space (primary space), for instance a data warehouse cube, the resulting structure doesn't necessarily reflect the real dimension of the embedded (secondary) space. We show that a metric-structured set has, in general, a fractal dimension. This means that the data set is a finite subset of a fractal secondary space of lower dimension.

Mapping the set into the secondary space of lower dimension will not result in loss of information with regard to the semantics defined by the measure. However, it helps to reduce storage and computing efforts. Additionally, the secondary space itself reveals much about the set's structure and can facilitate data mining.

The main problem with the secondary space is that it is unknown, and if it is not a linear sub-space of $\Re^d$, then there is no algorithm to determine it. We make a proposal for adding the property of a dimension to a metric and show that this is compatible to our customized understanding of a a dimension. We present an algorithm which computes, in optimal time, an index on the data set. The index can be regarded as a materialized view for representing the structure of the elements with respect to the given metric. The algorithm works independently of the dimension of the primary space, i.e. pumping up the dimension of the primary space wil not affect the result and the computational effort. Thus, besides supporting navigation and answering neighborhood queries, the algorihm also determines the real dimension of the underlying data set, even if the set is not linearly structured. We prove that the algorithm is robust with respect to skewed data or a distortion of the distance measure.

Author's address

University Gh Kassel - Fachbereich 17
D-34109 Kassel
Germany
E-mail: czi@db.informatik.uni-kassel.de

For technical reasons, this article is stored as two separate computer files, labelled "Part I" and "Part II", respectively. Please replace the present page by a printout of the "Part II" file.