www.elsevier.com

The Merkurion approach for similarity searching optimization in Database Management Systems

Marcos V.N. Bedo, Daniel S. Kaster, Agma J.M. Traina, Caetano Traina

PII: S0169-023X(16)30224-5
DOI: http://dx.doi.org/10.1016/j.datak.2017.09.003
Reference: DATAK1614

To appear in: *Data & Knowledge Engineering*

Received date: 29 September 2016
Revised date: 3 August 2017
Accepted date: 18 September 2017

Cite this article as: Marcos V.N. Bedo, Daniel S. Kaster, Agma J.M. Traina and Caetano Traina, The Merkurion approach for similarity searching optimization in Database Management Systems, *Data & Knowledge Engineering,* http://dx.doi.org/10.1016/j.datak.2017.09.003

The Merkurion approach for similarity searching optimization in Database Management Systems

Marcos V. N. Bedo[a,b,*], Daniel S. Kaster[c], Agma J. M. Traina[a], Caetano Traina Jr.[a]

[a] *Institute of Mathematics and Computer Sciences, USP, São Carlos, Brazil*
[b] *Fluminense Northwest Institute, UFF, St. A. Pádua, Brazil*
[c] *Computer Science Department, UEL, Londrina, Brazil*

Abstract

Modern Database Management Systems (DBMSs) retrieve songs that resemble those in a music dataset, identify plagiarism in a set of documents, or provide past cases to physicians by taking into account the characteristics of a query exam. All such tasks require the comparison of data by similarity, which can be expressed in terms of distance-based queries in metric spaces. Traditional query processing relies mostly on histograms for describing the data distribution space and choosing a data retrieval path that quickly leads to the answer, discarding comparisons of most unwanted data. However, DBMSs still lack adequate support for selectivity estimation of query operators for data types embedded in metric spaces. This article addresses a novel strategy that extends the query optimizer of any DBMS, so that it can also perform both logical and physical query plan optimizations in searches that include similarity predicates. The proposal, named Merkurion, updates the concept of Data Distribution Space and captures data distributions according to the distances between the elements within a dataset. Moreover, it employs concise representations of such distributions, called synopses, for the definition of rules that enable similarity searching optimization. An extensive evaluation of Merkurion in real-world datasets has proven its effectiveness and broad applicability to many data domains.

Keywords: Similarity Searching, Query Optimization, Selectivity Estimation, Design and Implementation Techniques

1. Introduction

Similarity searching is an important paradigm for modern computer applications and has been widely employed for the retrieval, clustering and classification of data [1, 2, 3]. Accordingly, many studies have designed extensions for the support of similarity searching in current DBMSs [4, 5]. A similarity-extended DBMS should handle dynamic queries whose predicates include relational conditions, such as identity (=, ≠)

[*] Corresponding author. São Carlos/SP – Brazil. P. O. Box 668. Zip Code: 13566-590.
Email address: bedo@icmc.usp.br (Marcos V. N. Bedo)