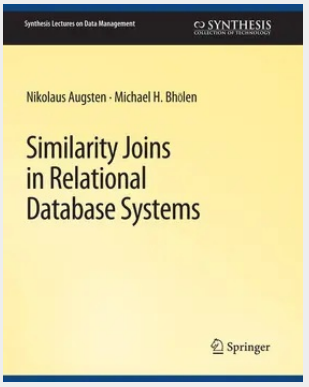
Augsten / Bohlen

Similarity Joins in Relational Database Systems



State-of-the-art database systems manage and process a variety of complex objects, including strings and trees. For such objects equality comparisons are often not meaningful and must be replaced by similarity comparisons. This book describes the concepts and techniques to incorporate similarity into database systems. We start out by discussing the properties of strings and trees, and identify the edit distance as the de facto standard for comparing complex objects. Since the edit distance is computationally expensive, token-based distances have been introduced to speed up edit distance computations. The basic idea is to decompose complex objects into sets of tokens that can be compared efficiently. Token-based distances are used to compute an approximation of the edit distance and prune expensive edit distance calculations. A key observation when computing similarity joins is that many of the object pairs, for which the similarity is computed, are very different from each other. Filters exploit this property to improve the performance of similarity joins. A filter preprocesses the input data sets and produces a set of candidate pairs. The distance function is evaluated on the candidate pairs only. We describe the essential query processing techniques for filters based on lower and upper bounds. For token equality joins we describe prefix, size, positional and partitioning filters, which can be used to avoid the computation of small intersections that are not needed since the similarity would be too low.

State-of-the-art database systems manage and process a variety of complex objects, including strings and trees. For such objects equality comparisons are often not meaningful and must be replaced by similarity comparisons. This book describes the concepts and techniques to incorporate similarity into database systems. We start out by discussing the properties of strings and trees, and identify the edit distance as the de facto standard for comparing complex objects. Since the edit distance is computationally expensive, token-based distances have been introduced to speed up edit distance computations. The basic idea is to decompose complex objects into sets of tokens that can be compared efficiently. Token-based distances are used to compute an approximation of the edit distance and prune expensive edit distance calculations. A key observation when computing similarity joins is that many of the object pairs, for which the similarity is computed, are very different from each other. Filters exploit this property to improve the performance of similarity joins. A filter preprocesses the input data sets and produces a set of candidate pairs. The distance function is evaluated on the candidate pairs only. We describe the essential query processing techniques for filters based on lower and upper bounds. For token equality joins we describe prefix, size, positional and partitioning filters, which can be used to avoid the computation of small intersections that are not needed since the similarity would be too low.

35,30 €
32,99 € (zzgl. MwSt.)

Lieferfrist: bis zu 10 Tage

Artikelnummer: 9783031007231
Medium: Buch
ISBN: 978-3-031-00723-1
Verlag: Springer International Publishing
Erscheinungstermin: 19.11.2013
Sprache(n): Englisch
Auflage: Erscheinungsjahr 2013
Serie: Synthesis Lectures on Data Management
Produktform: Kartoniert
Gewicht: 255 g
Seiten: 106
Format (B x H): 191 x 235 mm

Kundenservice Verlag Dr. Otto Schmidt KG
Neumannstraße 10, 40235 Düsseldorf | kundenservice@otto-schmidt.de | 0221-93738-997 (Inland)