A Near-Optimal Domain Independent Approach for Detecting Approximate Duplicates

Authors : Abdelaziz Fellah, Allaoua Maamir

Abstract : We propose a domain-independent merging-cluster filter approach complemented with a set of algorithms for identifying approximate duplicate entities efficiently and accurately within a single and across multiple data sources. The near-optimal merging-cluster filter (MCF) approach is based on the Monge-Elkan well-tuned algorithm and extended with an affine variant of the Smith-Waterman similarity measure. Then we present constant, variable, and function threshold algorithms that work conceptually in a divide-merge filtering fashion for detecting near duplicates as hierarchical clusters along with their corresponding representatives. The algorithms take recursive refinement approaches in the spirit of filtering, merging, and updating, cluster representatives to detect approximate duplicates at each level of the cluster tree. Experiments show a high effectiveness and accuracy of the MCF approach in detecting approximate duplicates by outperforming the seminal Monge-Elkan's algorithm on several real-world benchmarks and generated datasets.

Keywords : data mining, data cleaning, approximate duplicates, near-duplicates detection, data mining applications and discovery

Conference Title : ICDMMIP 2017 : International Conference on Data Mining, Multimedia and Image Processing

Conference Location : Paris, France

Conference Dates : February 23-24, 2017

Open Science Index, Computer and Information Engineering Vol:11, No:02, 2017 publications.waset.org/abstracts/64998.pdf