

**String Similarity Search: A Hash-Based Approach**

**Abstract:**

String similarity search is a fundamental query that has been widely used for DNA sequencing, error-tolerant query auto-completion, and data cleaning needed in database, data warehouse and data mining. In this paper, we study string similarity search based on edit distance that is supported by many database management systems such as Oracle and PostgreSQL. Given the edit distance, ed(s, t), between two strings, s and t, the string similarity search is to find every string t in a string database D which is similar to a query string s such that ed(s, t) \_ \_ for a given threshold \_. In the literature, most existing work take a filter-and-verify approach, where the filter step is introduced to reduce the high verification cost of two strings by utilizing an index built offline for D. The two up-to-date approaches are prefix filtering and local filtering. In this paper, we study string similarity search where strings can be either short or long. Our approach can support long strings, which are not well supported by the existing approaches due to the size of the index built and the time to build such index. We propose two new hash-based labeling techniques, named OX label and XX label, for string similarity search. We assign a hash-label, Hs, to a string s, and prune the dissimilar strings by comparing two hash-labels, Hs and Ht, for two strings s and t in the filter step. The key idea behind is to take the dissimilar bit-patterns between two hash-labels. We discuss our hash-based approaches, address their pruning power, and give the algorithms. Our hash-based approaches achieve high efficiency, and keep its index size and index construction time one order of magnitude smaller than the existing approaches in our experiment at the same time.

**Existing System:**

The existing approaches need to build a large index when the strings are longer in a dataset. The larger the index is, the more time it requires to process queries. Accordingly, the performance decreases. In addition, such performance will be affected by a large number of strings in a dataset. In this paper, we focus on new hash-based approaches to deal with large short/long string datasets.

**Proposed System:**

In this paper, we study string similarity search, when the query string s and the average string t in D can be long. The up-to-date approaches cannot efficiently process long string similarity search for the following main reasons. For the prefix filtering approaches, the main idea is to use a small number of q-grams for filtering. When strings become long, the pruning power of such a small number of q-grams will reduce significantly. In addition, the prefix filtering approaches need to know \_ before the index construction. However, when the average strings become long, users want to use different \_ for string similar search: a small \_ for short strings and a large \_ for long strings.

We propose new hash-based labeling for string similar search.