

**THEORETIC BACKGROUNDS**

The problem of similarity queries on vector space models is a general problem compared to commonly used functions such as Jaccard and cosine. It is a difficult problem on large-scale date sets due to three aspects.

1. Arbitrary token weights integrated in pair-wise similarity.
2. Variant degrees of each token in different records.
3. Given a threshold, exactly answer a query from millions of records.

**CONTRIBUTIONS**

We formally define the problem of similarity queries on collections of records using a vector space model, and solve it efficiently based on a revised inverted index. Our main contributions are

1. Pruning methods with various bounds
2. Solid proofs of their correctness
3. An efficient elastic scheme to iteratively tighten the bounds
4. A family of algorithms based on an inverted index for large data sets
5. Evaluating intensive experiments on dozens million of records.

**FORMAL DEFINITION**

Given a collection of records \( S \), a p-norm selection includes a record \( Q \) and a threshold \( \tau \). Suppose a token has a weight \( w(t) \) and a degree \( f(t, R) \) (in record \( R \)), it is to find all the records \( R \in S \) such that

\[
\left( \sum_{t \in R \cap Q} f(t, R)w(t) \right)^p \left( \sum_{t \in R \setminus Q} f(t, R)w(t) \right)^p \geq \tau
\]

**USABILITY**

Based on the ground truth about similar pairs of records on “AskUbuntu”, the two figures compare the precision and recall of this search using different functions. The results given above say that:

- The tf-idf 2-norm function gave the best accuracy results.
- The two 2-norm weighting functions were better than Jaccard and Cosine.

It was not to show that the p-norm function is the best for all applications. Instead, but a better similarity in certain applications.

**RESULTS**

**ELASTIC BOUNDS**

Significant improvement can be achieved as long as we iteratively tighten the bounds based on decreasing maximal degrees and reversely shrink the length bounds:

- After each step, the minimal and maximal degrees will be recomputed by the decreasing length bounded by the candidate set.
- The updated length bounds are used to shrink the range of the minimal and maximal degrees of unseen tokens.
- The maximal token degrees are associated in the index to iteratively tighten length bounds and maximal degrees until no more changes are observed.

![Scanning record entries using Elastic Bounds](Image)

**A Future Direction**

Can we associate part of the “remaining tokens” in the inverted index, allowing us to do “in-place” pruning. How to decide the associated length needs in-deep researches.

**Source Code**

Please contact by the following email or from Weechart by Email: lwh@whu.edu.cn