

Web Data Integration Using Approximate String Join

Yingping Huang and Gregory Madey

Computer Science and Engineering

University of Notre Dame

WWW2004, New York, 5/19/2004

Introduction

- Web data integration is an important preprocessing step for web mining and data analysis.
- Approximate string processing is a fundamental step in many existing data cleansing algorithms.
- Approximate string join seeks to identify (almost) all pairs of strings whose distances are less than a certain threshold.
- Typical string distances include edit distance, q-gram distance and vector cosine similarity.

Related Work

- Li (2003) proposed a mapping algorithm where each string is mapped to a point in a high dimensional euclidean space using FastMap. Then a similarity join algorithm proposed by Hjaltason and Snel (1998) is used to identify close points.
- Gravano (2003) presented a sampling approach for performing text join where each string is represented by a sparse vector in a high dimensional space. Then a join is performed on the resulting vector space.

Drawbacks of Previous Approach

- In Li (2003), the similarity join algorithms is computationally sensitive to the dimensionality of the hosting space. When the dimensionality gets large, the similarity join algorithms becomes very inefficient.
- In Gravano (2003), the sampling method uses a lower dimensional subspace for join. The accuracy of this approach depends on the dimensionality of the subspace. Usually, to obtain a high accuracy, the dimensionality of the subspace is close to the dimensionality of the original space.

Our Approach

- We first form the database of strings to be a $(1,2)$ -B metric space and then map the $(1,2)$ -B metric space into a high dimensional grid space.
- Pairs of points with distance 1 are identified in the grid space. Any two points in the grid space have distance 0, 1, or 2.
- A post join process is performed to remove false positives.

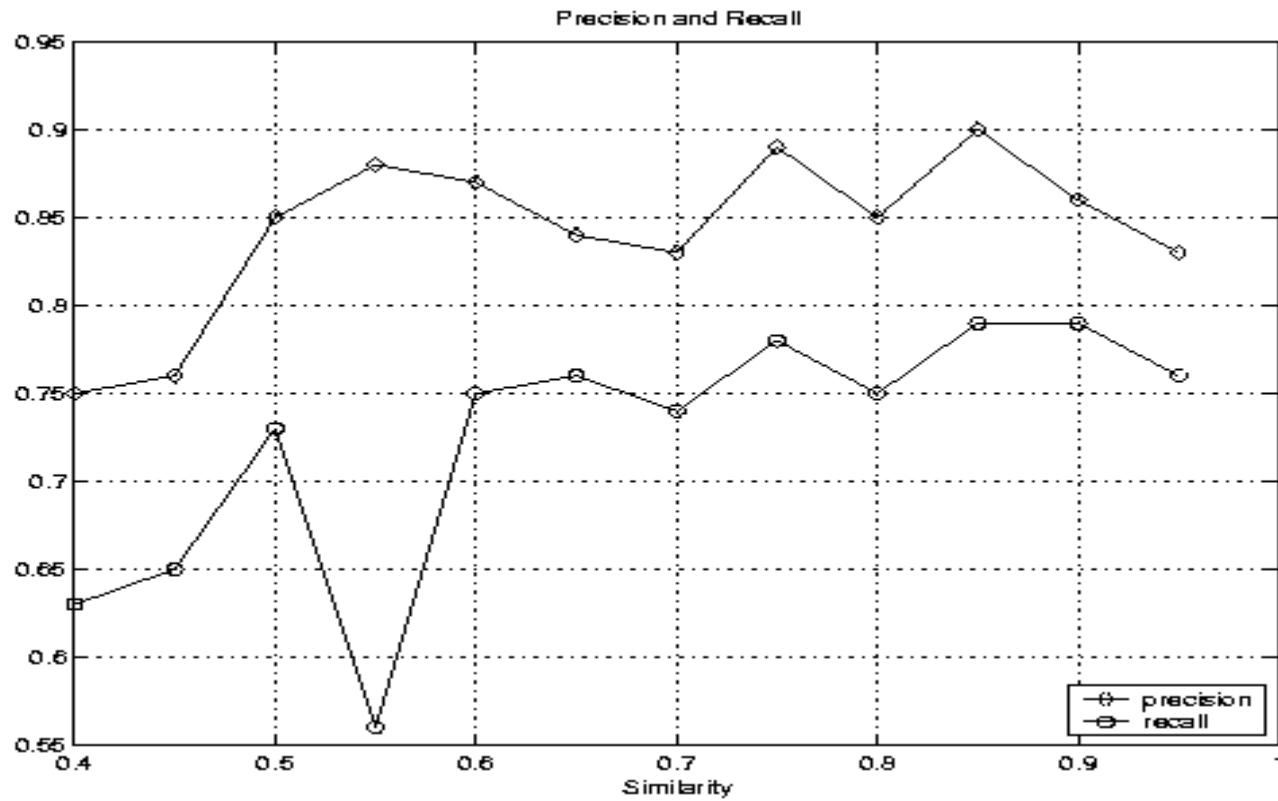
(1,2)-B Metric Space

- A metric space $M=(X,D)$ is called a (1,2)-B metric space, if the distance between any two points is either 0, 1, or 2, and for any point in X , there are no more than B points within distance 1.
- For any two strings s and t , if their string distance is less than k , then we define their new distance to be 1, otherwise, we define their distance to be 2. We also assume that each string has at most B other strings that have string distance less than k .

Lemma

- Guruswami (2003) proved that a $(1,2)$ -B metric space can be isometrically embedded into a high dimensional grid space, with dimensionality $O(B \log N)$ where N is the size of the string database.
- An approximate matrix multiplication method is used to construct the actual mapping.

Results



The precision and recall are both reasonably good.

Summary

- The previous figure shows that our approach achieves good precision and recall.
- It has some potential advantage over the algorithms presented by Li (2003) and Gravano (2003).
- The execution time is almost linear to the dimensionality of the hosting grid space.

References

- Li (2003): L. Jin, C. Li and S. Mehrotra. Efficient record linkage in large datasets. In Proc. 8th international conference on database systems for advanced applications.
- Gravano (2003): L. Gravano and P. Ipeirotis. Text join in an rdbms for web data integration. Proc. 12th international WWW conference.
- Guruswami (2003): V. Guruswami and P. Indyk. Embeddings and non-approximability of geometric problems. In Proc. 14th Annual ACM-SIAM Symposium on Discrete Algorithms.