

Applications of LSH

Entity Resolution
Fingerprints
Similar News Articles

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University



Entity Resolution

- The *entity-resolution* problem is to examine a collection of records and determine which refer to the same entity.
 - *Entities* could be people, events, etc.
- Typically, we want to merge records if their values in corresponding fields are similar.

Matching Customer Records

- I once took a consulting job solving the following problem:
 - Company A agreed to solicit customers for Company B, for a fee.
 - They then argued over how many customers.
 - Neither recorded exactly which customers were involved.

Customer Records – (2)

- Each company had about 1 million records describing customers that might have been sent from A to B.
- Records had name, address, and phone, but for various reasons, they could be different for the same person.

Customer Records – (3)

- **Step 1:** Design a measure (“*score*”) of how similar records are:
 - E.g., deduct points for small misspellings (“Jeffrey” vs. “Jeffery”) or same phone with different area code.
- **Step 2:** Score all pairs of records that the LSH scheme identified as candidates; report high scores as matches.

Customer Records – (4)

- **Problem:** $(1 \text{ million})^2$ is too many pairs of records to score.
- **Solution:** A simple LSH.
 - Three hash functions: exact values of name, address, phone.
 - Compare iff records are identical in at least one.
 - Misses similar records with a small differences in all three fields.

Aside: Hashing Names, Etc.

- How do we hash strings such as names so there is one bucket for each string?
- **Answer:** Sort the strings instead.
- Another option was to use a few million buckets, and deal with buckets that contain several different strings.

Aside: Validation of Results

- We were able to tell what values of the scoring function were reliable in an interesting way.
- Identical records had a creation date difference of 10 days.
- We only looked for records created within 90 days of each other, so bogus matches had a 45-day average.

Validation – (2)

- By looking at the pool of matches with a fixed score, we could compute the average time-difference, say x , and deduce that fraction $(45-x)/35$ of them were valid matches.
- Alas, the lawyers didn't think the jury would understand.

Validation – Generalized

- Any field not used in the LSH could have been used to validate, provided corresponding values were closer for true matches than false.
- **Example:** if records had a **height** field, we would expect true matches to be close, false matches to have the average difference for random people.