

Bellman

Contents and Structure

Key/FK relationships

- important
- poorly documented
- poorly maintained
- heterogeneous: different join paths for different tuples

DB Profiling

Summaries of tables and fields

Visualized for interaction

Enable very quick questions about the data

- Find all joins between this table and others. Direction? Size?
- Find composite fields similar to this one
- Find if table has heterogenous join paths

Food for thought

Q-gram vector distance (from sketch) vs. Q-gram resemblance?

Bellman + Potter's Wheel?

Can use types as virtual fields?

Bellman

schema browser

#tables, #rows, #distinct, #nulls

convert everything to a string before you start

keys, multiset signatures, q-gram sigs, q-gram sketches

Bellman's summaries

Set signatures

For set resemblance (Jaccard index)
Provides size of intersection if you also know cardinality (#distinct values) per set
 $|A \cap B| = r/(r+1)(|A|+|B|)$
Trick: $Pr[\minHash(A) = \minHash(B)] = \text{resemblance}$
But, high variance. So do this with N hash functions, divide hit-rate by N
This set signature can also estimate $r_{\{A \setminus B\}} = |A \setminus B|/|A \cup B|$
To "sum" 2 signatures and form $S(A \cup B)$, just take for all hash functions s_i the minhash position of the 2 that's smaller.
In Bellman: for every field with ≥ 20 values, form a set signature in a metadata table. Self-join of the metadata table gives likely matching columns.

multiset signature

minHash Count: how often does the minhash value occur, for lots of choices of hash function? This gives a sample of the frequency distribution
minHash sample with minHash counts
"summable": sample is min position, count is corresponding count (or sum of counts if they match)
Gives nice picture of "tail" of a distribution ("outlier" values in the head can be captured explicitly)
Can tell us frequency distros of joins (1:how-many) and compare "distro dependence": distro of values with distro of join fanout
Join sizes and frequency distribution of join result

Substring resemblance

Q-grams

just minHash-based Jaccard over QGrams
Can get intersection size of QGrams as above

Q-gram signatures

Similar for QGram set difference (containment)

Q-gram sketches

Since set sigs are easily summed, can find fields A,C whose combination is closer to another field B than either A or C is alone
Nice for pruning join paths

Q-Gram vector: normalized count of each Qgram in the multiset

Q-Gram vector distance: $\sqrt{\text{sum-squared dist}}$

AMS sketch: small-d random projections, preserves distance

UNlike sigs, don't help with containment

Keys

See TANE

Finding Structures

Join Paths

Find pairs of similar fields

For each destination table from source T

- Find matches to key columns of source table
- Find matches to key column of dest table
- Can characterize "strength" of join in terms of fanout

Composite fields

given a source field, find target fields with high q-gram resemblance

For each destination table, take all combos of 2 or more dest fields, and sort by rank
prune after 3 fields?

Heterogeneous Tables

For every field in table, find high resemblance matches

For each destination table, accumulate fields whose mutual intersection with source is small, and come from different destination tables