On Multi-Column Foreign Key Discovery

Meihui Zhang   Marios Hadjieleftheriou   Beng Chin Ooi
Cecilia M. Procopiuc   Divesh Srivastava

VLDB: September 2010

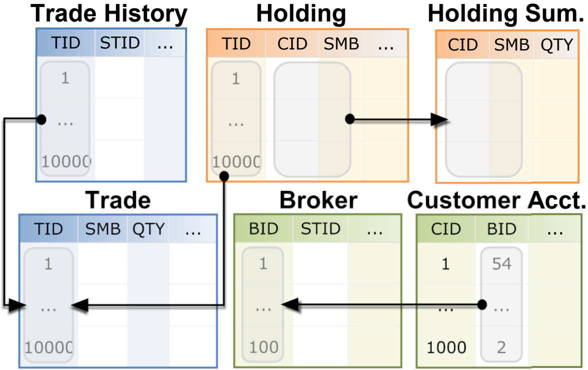
Daisy - AAU

Presented by: Benjamin Krogh

November 11, 2010

Introduction

- Important constraint on data
- Designers frequently fail to specify foreign keys
- Most previous work focuses on inclusion dependencies
- Inclusion dependencies yields many false positives
- Multi-column foreign keys have not been considered yet



Trade History
TID
STID
...
1
...
10000
Holding
TID
CID
SMB
...
1
...
10000
Holding Sum.
CID
SMB
QTY
Trade
TID
SMB
QTY
...
1
...
10000
Broker
BID
STID
...
1
...
100
Customer Acct.
CID
BID
...
1
...
1000
54
...
2

Overview

1. Introduction
2. Traits of Foreign Key
3. Overall Design
4. Inclusion Dependencies
5. Randomness Test
6. Experiments
7. Conclusion

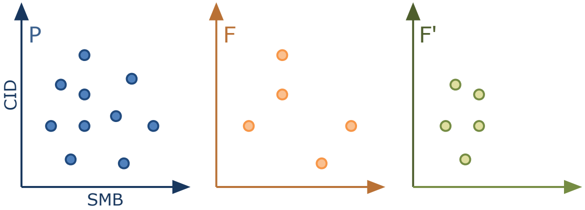
Traits of a Foreign Key

1. Should have significant cardinality
2. Should have good coverage of the primary key
3. Should not be the primary key of many other foreign keys
4. Its values should not be a subset of many primary keys
5. The average length should be similar to that of the primary key
6. The column names of foreign/primary keys should be similar

Traits of a Foreign Key - continued

Randomness

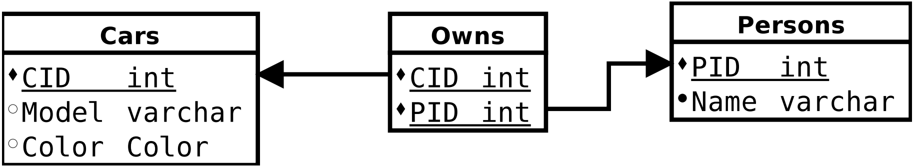
The values of a foreign key will appear to be a random sample of the primary key



Motivation for Randomness

Intuition

The logic that generates the primary key, is disconnected from the logic generating the fk.



Overall Design

1. Find candidate foreign/primary key pairs based on inclusion dependencies
2. Rank the pairs such that pairs with similar distribution scores best
3. Return top X-%

Inclusion Dependencies

Approach

Estimate inclusion dependencies based on a number of samples

- Dirty data due to unenforced constraints.
- Include pairs fulfilling: $\sigma(F, P) = \frac{|F \cap P|}{|F|} \geq \theta$
- Use bottom-k sketches to estimate inclusion dependencies

CID	SMB	...
10	INTC	
2	AAPL	
217	GOOG	

Hash
h(10|INTC) = 10
h(2|AAPL) = 1
h(217|GOOG)=5

Bottom-1
2|AAPL, 1

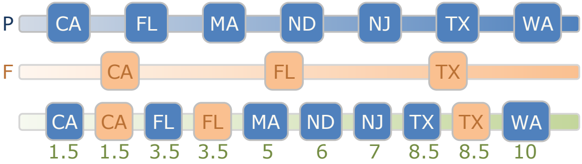
Bottom-1 sketch

Randomness Test - Single-Column

Wilcoxon test

A standard statistical test for randomness

1. Sort values in $F \cup P$
2. Assign ranks
3. Compute rank-sum of duplicate values

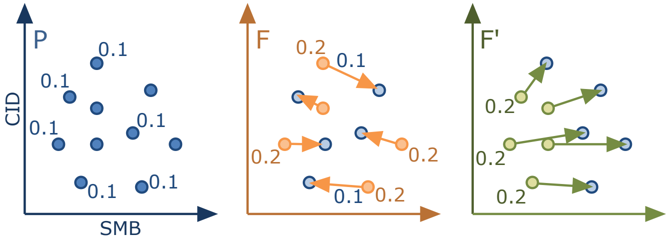


Randomness Test - Multi-Column

Earth Mover's Distance

Assume two piles of dirt A and B, then $EMD(A, B)$ is the amount of work to convert A to B.

1. Assign a probability *mass* to each point, such that the sum of probabilities is 1.
2. For each point a in A, find the distance to the nearest point in $B \setminus A$ and multiply by a's *mass*.



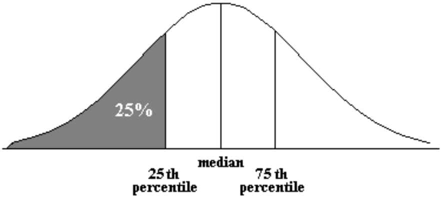
Randomness Test - Multi-Column continued

Problem

Exact algorithm for EMD has cubic complexity, i.e. unfeasible for large tables.

Solution

Use quantiles to summarize values.
The motivation for *n*-quantiles is to divide ordered data into *n* essentially equal-sized data subsets.



An example 4-quantile, dividing a collection into 4 equal sized parts.

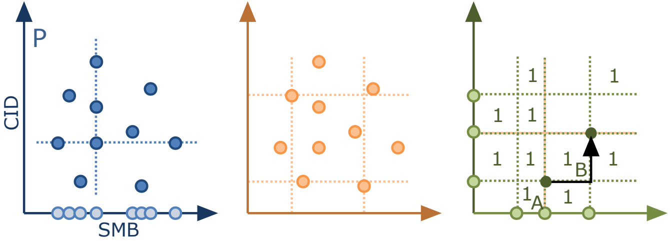
Randomness Test - Multi-Column continued

Problem

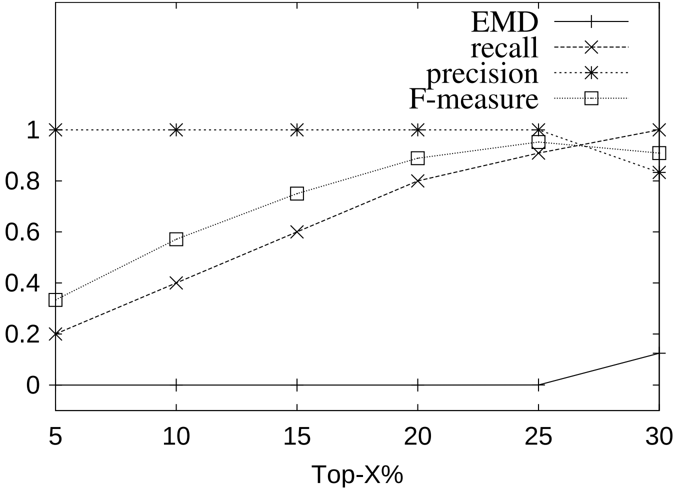
Exact algorithm for EMD has cubic complexity, i.e. unfeasible for large tables.

Solution

Use quantiles to summarize values.

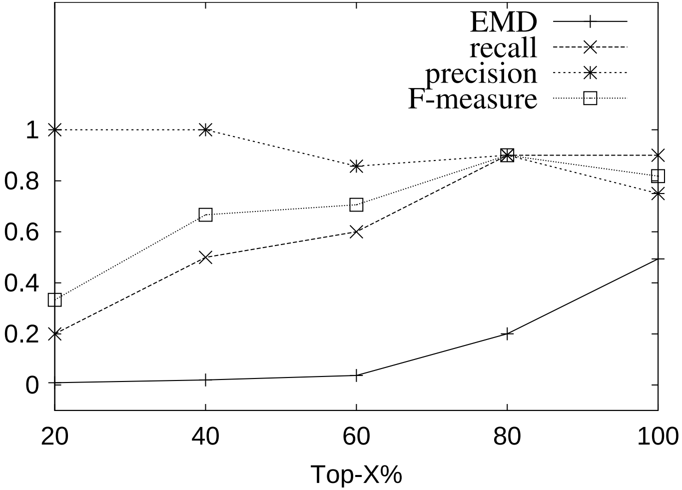


Experiments - 1



TPC-H

Experiments - 2



Wikipedia

Conclusion

- Linear approach to finding foreign key
- Ranking based on *Randomness* property
- Distance measure quantifying randomness
- Fast approximate algorithms for evaluating randomness over a large set of columns
- Comprehensive experimental validation using both synthetic and real datasets.

Good

- A better approach to fk discovery, requiring no domain knowledge
- Novel idea of *Randomness*
- Evaluation on real world schemas and datasets
- They propose a solution to a very real and very important problem
- The first to consider multi-column fks.

Bad

1. "Highly unlikely that a database design incurs a bias" on p. 2 is an unsupported claim.
2. Algorithm in appendix is very dense and unexplained (A more little handholding please).
3. It is unclear why Wilcoxon test is relevant, nearly half a page is spent.
4. Approximate quantiles on p. 6 are not defined.
5. Example on fig. 1 is not too helpful.
6. Paper is fragmented into many topics, difficult to read.
7. From the description of EMD it is unclear how the probability mass affects the result.
8. Not explained why a dataset is dirty.

Relevance to Project

- Foreign keys are an integral part of good database design
- Given foreign keys many additional checks can be implemented
- The majority of open source systems we have examined does not have fks

Questions

?