BANKS

Browsing and Keyword Search in Relational Databases

B. Aditya, Gaurav Bhalotia, Soumen Chakrabarti, Arvind Hulgeri, Charuta Nakhe, Parag, S. Sudarshan

IIT Bombay

http://www.cse.iitb.ac.in/banks/
Motivation

Web search engines are very successful
- Simple and intuitive keyword query interface

Database querying using keywords is desirable
- Query languages, e.g., SQL/QBE, are not appropriate for casual users
- Form interfaces cumbersome, give limited views

Examples of keyword queries on databases
- e-store database: “camcorder panasonic”
- Book store: “sudarshan databases”

Differences from IR/Web Search
- Normalization splits related data across multiple tuples
- Answer to a query is a set of (closely) connected tuples that match all given keywords
Basic Model

Database: modeled as a graph

- Nodes = tuples
- Edges = references between tuples
  - foreign key, inclusion dependencies, etc.
  - Edges are directed

BANKS: Keyword search...
MultiQuery Optimization

Charuta  S. Sudarshan  Prasan Roy

paper  writes  author
**Answer Model**

- Rooted, directed tree connecting keyword nodes
  - May include internal nodes that contain no keywords
  - Root node has special significance
    - May be restricted to relations representing entities
    - Avoid relations representing relationships, e.g. “writes”

- An example: “sudarshan roy”

- Multiple answers may exist
  - Ranked by proximity + prestige
Relevance Calculation

§ Proximity
- Forward edges: foreign key → primary key
- Weight of forward edge is based on schema
  - E.g. “cites” link weight greater than “writes” link weight
- May need backward edges to form answer tree
  - Weight of backward edge $u \rightarrow v \propto$ indegree of $u$

§ Node prestige based on indegree

§ Answer tree relevance
- Edge score $E = 1 / \Sigma$ edge-weights
- Node score $N = \Sigma$ root- and leaf-node-weights
  - Ignore weights of internal nodes
- Normalize and combine using weighting factor $\lambda$
  - Additive: $(1 - \lambda) E + \lambda N$; multiplicative: $EN^\lambda$
Answer Trees

- **Anecdotal results**
  - "Mohan": C. Mohan at the top based on prestige (\# of papers)
  - "Transaction": Jim Gray’s classic paper and textbook at the top based on prestige (\# of citations)
  - "Sunita Seltzer": No common papers, but both have papers with Stonebraker; system finds this connection

- **Backward expanding search algorithm**
  - Start at leaf nodes each containing a query keyword
  - Run concurrent single source shortest path algorithm from each such node, traversing edges backwards
  - Confluence of backward paths identify answer tree roots
  - Answer trees may not be generated in relevance order
    - Insert answers to a small buffer (heap) as are generated
    - Output highest ranked answer from buffer when buffer is full
The BANKS System

Available on the web, with (part of) DBLP data
- http://www.cse.iitb.ac.in/banks/

No programming needed for customization
- Minimal preprocessing to create indices and give weights to links

Provides keyword search coupled with extensive browsing features
- Schema browsing + data browsing
- Hyperlinks are automatically added to all displayed results
- Browsing data by grouping and creating crosstabs
- Graphical display of data: bar charts, pie charts, etc