Searching with Numbers

IBM Thinkpad (750 MHz, 196 MB DRAM, ...)
Dell Computer (700 MHz Celeron, 256 MB SDRAM, ...)

Catalog Database

800 200
800 200 3 lb

Storage & Querying of eCommerce Data

eCommerce Applications
SELECT name, output FROM H

Data stored in conventional way
name monitor recharge output scan ...
PANL75 7 inch Built-in - - ...
KLH 221 - - S-Video - ...

1. Problem with Conventional Schema
- Large number of Columns
- Sparsity
- Constant schema evolution
- Performance

Vertical Table (V)

name key val
0 name PANL75
0 monitor 7 inch
0 Recharge Built-in
0 Output Digital
...
1 name KLH 221
1 Output S-Video

Horizontal View (HV)

Query Mapping Layer

Optimized Operator Implementation

3. Solution: Query Mapping Layer
- Hides complexity of vertical representation
- Fast performance

Recommendations for Database Vendors:
- Artial Indices
- Enhanced Table Functions (TF)
- First Class treatment of TF

Other Applications:
- Stores for XML, RDF, LDAP and Data Mining

R. Agrawal, A. Somani and Y. Xu, "Searching and Querying of E-Commerce Data", VLDB 2001

Catalog Integration

- ICS
- DSP
- Mem
- Logic
- Master Catalog
- New Catalog
- After integration:
- ICS
- DSP
- Mem
- Logic

Goal
- Use affinity information in new catalog.
- Accuracy boost depends on match between two categorizations.

Problem Statement
- Given
  - master categorization M
    - categories C_1, C_2, ..., C_n
    - set of documents in each category
  - new categorization N:
    - categories S_1, S_2, ..., S_m
    - set of documents in each category
- Standard Alg: Compute Pr(C_i | d)
- Enhanced Alg: Compute Pr(C_i | d, S)

Enhanced Naïve Bayes classifier
Pr(C_i | d, S) = Pr(C_i | d) Pr(d | C_i)
Pr(d | S) = \sum_{C_i} Pr(C_i | d) Pr(d | C_i)

Pr(C_i | S) = \sum_{C_i} Pr(C_i | d) Pr(d | C_i)

- Use tuning set to determine w.
  - Defaults to standard Naïve Bayes if w = 0.
  - Only affects classification of borderline documents.

Empirical Results


Database Technologies For Electronic Commerce
Rakesh Agrawal, Ramakrishnan Srikant, Yirong Xu