Efficient Implementation of Large-Scale Multi-Structural Databases

Ravi Kumar
Yahoo! Research
ravikumar@yahoo-inc.com
Joint work with

- Ronald Fagin, IBM Almaden
- Phokion Kolaitis, IBM Almaden
- Jasmine Novak, Yahoo! Research
- D. Sivakumar, Google
- Andrew Tomkins, Yahoo! Research

Work done at IBM Almaden Research Center
Outline of the talk

• Multi-structural databases (MSDB)
• Queries
• A new algorithm
• Conclusions
Given a database of news articles, categorized by media type, topic, company, geography, publication date, etc, ask:

There seem to be a lot of documents talking about politics – do they come predominantly from any particular geography and/or time?

Ans: Eg, documents about politics are much more likely to come from Europe, published in 2005, or from Korea, published in June of 2004
A motivating example

Given a database of news articles, categorized by media type, topic, company, geography, publication date, etc, ask:

Among documents that mention finance what are the topics that are strongly correlated with a particular geography?

Ans: Eg, globalization is strongly correlated with California, India, and China; and currency is strongly correlated with Europe
Given a database of news articles, categorized by media type, topic, company, geography, publication date, etc, ask:

What are the three combinations of geography and media type that have grown most significantly over the last year?

Ans: Eg, Japanese press releases, Asian newspapers, and Iranian blogs
Some observations

Japanese press releases, Asian newspapers, and Iranian blogs

• Consist of combinations of geography and media type
  – Express “clusters” in easy-to-understand terms

• Select different levels of the geographic hierarchy in the same result set
  – May select one or more appropriate granularities for the answer

• Japan is part of Asia, but fortunately newspapers and press releases are disjoint
  – Will not return overlapping regions of the multi-dimensional space
A framework …

Multi-Structural Databases (MSDB)
• **PODS 2005**: “Multi-Structural Databases”
  – Described basic framework
  – Gave three example query types
  – Small study

• **This talk**: “Efficient Implementation of Large-Scale Multi-Structural Databases”
  – Broader family of queries based on the framework
  – New algorithms for certain cases
  – Larger study: real-time queries on 4B web pages
We, the people, in order to form a more perfect union, do hereby...
**Dimensions**

- **Hierarchical dimensions**
  - Eg, a tree of geographic locations, topics
  - Restrictions are nodes of the tree
    - Eg, restrict to documents from Europe

- **Numerical dimensions**
  - Eg, timestamp, price, temperature
  - Restrictions are intervals (ranges)
    - Eg, documents from 1720 to 1860

- **Most generally, lattice dimensions**
  - Dimension is a bounded lattice
  - Restrictions are lattice elements

We consider only hierarchical and numerical dimensions
### An example MSDB

<table>
<thead>
<tr>
<th>Doc ID</th>
<th>Name</th>
<th>Content type</th>
<th>Geography</th>
<th>Topic</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Declaration of Independence</td>
<td>Historical document</td>
<td>Pennsylvania</td>
<td>Politics</td>
<td>1776</td>
</tr>
<tr>
<td>2</td>
<td>The Dilbert Principle</td>
<td>Book</td>
<td>US</td>
<td>Humor</td>
<td>1995</td>
</tr>
<tr>
<td>3</td>
<td>Yahoo! announces earnings</td>
<td>News article</td>
<td>Washington, DC</td>
<td>Financial</td>
<td>2004</td>
</tr>
</tbody>
</table>
Combining dimensions

**Geography**

Europe

- France
- Spain

**Content type**

Media

- TV
- Radio

Multi-Dimensions: Easy to name and communicate elements and is based on user’s view of the world

Eg, ({Time, Topic}): 2003 Politics, 1990s Music
Pairwise-Disjoint Collections (PDCs)

- Elements of a PDC represent disjoint parts of the concept space
- No document overlap unless documents appear at multiple locations
Inputs to an MSDB query

Multi-Dimension

European Media

- French Media
- Spanish Media
- European TV
- European Radio

Document subset

- French Radio
- Spanish TV
- French TV
- Spanish Radio

PDC size

3
Outputs of an MSDB query

Multi-Dimension

Document subset

PDC size

Output is PDC of size at most 3 that is maximal under some measure
Particular query is determined by the measure
What is the measure/query?

Queries correspond to solving an optimization problem.

**Growth**: Break $X'$ into $k$ pieces of maximum growth.

Eg, In stock price time-series, show 4 intervals of maximum growth.
(f, °) query type

Given \( D' \subseteq D, X' \subseteq X \), parameter \( k \), find PDC \\{ L_1, \ldots, L_k \} such that \\
\[ f(X', L_1) \circ \cdots \circ f(X', L_k) \]

is maximized

Eg, \( f(X', L) = \#\{ x \in X' | x \text{ belongs to } L \}; \circ = + \)

Sum-additive query type

\( A \cap B = \emptyset \Rightarrow f(A \cup B, L) = f(A, L) + f(B, L) \)

\( \circ = + \)
Growth in \((f, \circ)\) language

- Find 4 regions of most rapid growth
- Numerical dimension
- A candidate \(f\)

\[
g(t) = \frac{\#docs@t}{\#docs@(t-1)}
\]

\[
f([a, b]) = \sum_{t \in [a, b]} \log (g(t))
\]

\(\circ = +\)

- Sum-additive type
More query types

- **Divide**: Break X’ into pieces that partition the space and have roughly equal cardinality
- **Differentiate**: Find parts of multi-dimension that occur more often in a foreground set of documents than a background set
- **Discover**: Find parts of X’ defined in terms of multi-dimension that are cohesive with respect to another set of “measurement” dimensions
- **Recency**: Find regions of X’ that have shown significant recent growth
- **Value**: Find parts of X’ that maximize a “value” function
Algorithms for sum-additive queries

Single hierarchical dimension
• $n =$ number of nodes in the tree
• $O(nk^2)$ algorithm

Single numerical dimension
• $n =$ number of time units
• $O(n^2 k)$ algorithm

Dynamic programming!
Given a plot of a stock price, and constrained to hold the stock during only k distinct regions, find the regions which maximize your total profit

A simple abstraction: Given

Maximal subinterval problem: Find k sub-intervals of this sequence whose sum is maximal

1  -3  12  -8  7  -10  15  -5  4
Folklore algorithm for \( k=1 \)

\[
\begin{align*}
1 & \quad -3 & \quad 12 & \quad -8 & \quad 7 & \quad -10 & \quad 15 & \quad -5 & \quad 4 \\
\text{Total: 1} & \quad \text{Total: -2} & \quad \text{Total: 12} & \quad \text{Total: 4} & \quad \text{Total: 11} & \quad \text{Total: 1} & \quad \text{Total: 16} & \quad \text{Total: 11} & \quad \text{Total: 15} \\
\end{align*}
\]

Answer = 16
An \( O(n^2k) \) algorithm

\[
P([1, i], k) = \text{Best solution of } x_1, ..., x_i \text{ with } k \text{ intervals}
\]

\[
P([j, i], 1) = \text{Best solution of } x_j \ldots x_i \text{ with } 1 \text{ interval}
\]

Solvable in \( O(|i-j|) \) time

\[
P([1, i], k) = \max\{j < i\} \ P([1, j-1], k-1) + P([j, i], 1)
\]

Prohibitive even when \( n \) is only large (eg, \#days)

Can we do better?
A bad idea

• For k = 2
  – Pick the best interval (k = 1)
  – Remove the interval
  – Pick the best interval in the rest (k = 1)

• Doesn’t work!

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-9</td>
<td>10</td>
<td>-11</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
1. Solve $k=1$

2. Invert subinterval

3. Solve for $k=1$
Five-way decomposition

A

B

C

D

E

Full interval

Most positive

Most negative of most positive
Five-way decomposition tree

Continue for k levels
Theorem: There is an optimal $k$-element solution such that every interval in the solution is a node of the tree.

Proof idea: Structural analysis of the optimal solution.

Theorem: There is an algorithm on the tree to find the best set of $k$ nodes with running time

$$\min(nk^2, \max(nk, k^5))$$

Proof idea: Dynamic programming on the five-way decomposition tree.
Experiments and graphs

See the paper!
Summary

• An abstraction of query types to capture many interesting queries for MSDB

• An almost optimal algorithm for the sum-additive query type for numerical dimension

• Large-scale experiments
Open problems

• New types of queries
• Broadening of the PDC concept
• Many interesting algorithmic open problems
  – Variants of many of these problems have been studied before
Thank you!

ravikumar@yahoo-inc.com