

**Search in data mining,
semi-structured data,
and data at various levels of abstraction (data versus meta-data)**

Dirk Van Gucht
Sept 20, 2006

`www.cs.indiana.edu/~vgucht`

Overview

- Search of data at various levels of abstraction; **data** versus **meta-data**
- Search in **semi-structured** data; search in **XML** documents using **XPath**
- Search for **frequent patterns** in data;
 - A measure-theoretic framework; bounds and constraints
 - Analysis of frequent pattern mining algorithms

Search of data at various levels of abstraction

The 3-level architecture for databases

View Level

Conceptual Level

Physical Level

- **Physical Data Independence:** data access specification (queries, searches) should be independent from the physical representation (files, indexes, disk) of the data.
- **Logical Data Independence:** views, applications etc should be minimally affected by changes at the conceptual level.

All access to a database should be specified in a **declarative** language (SQL, Google Access Language), rather than a **procedural** language.

View Level

View_CSDepartment

View_Bursar

View_Scheduling

Conceptual Level

University Database

Students, Course, Enrollment ...

Constraints: student must enroll ...

Physical Level

Spreadsheets, tables, XML docs

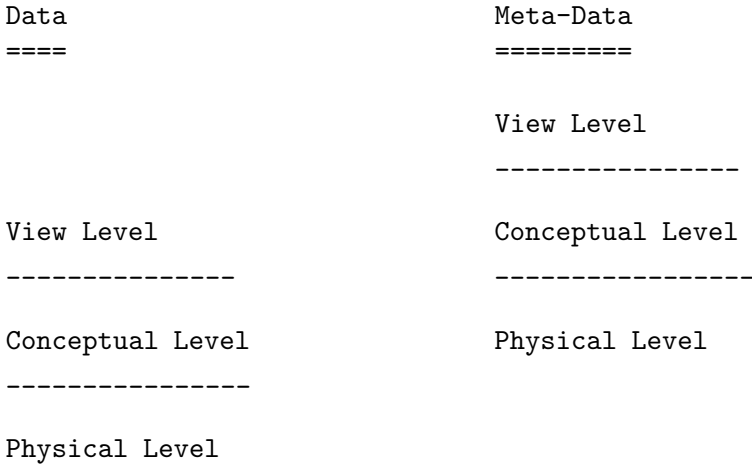
Indexes on properties data;

Partitioning of data;

Algorithms for key operations;

Layout of records on disk.

Data Versus Meta-data



Integrated (Uniform) Model of Data and Meta-data

Data & Meta-Data Combined

==== =====

View Level

Standard User

Database Administrators

Applications developers (e-commerce)

Conceptual Level

Integration of (multiple) databases
and catalogs ...

Constraints ...

Workloads ...

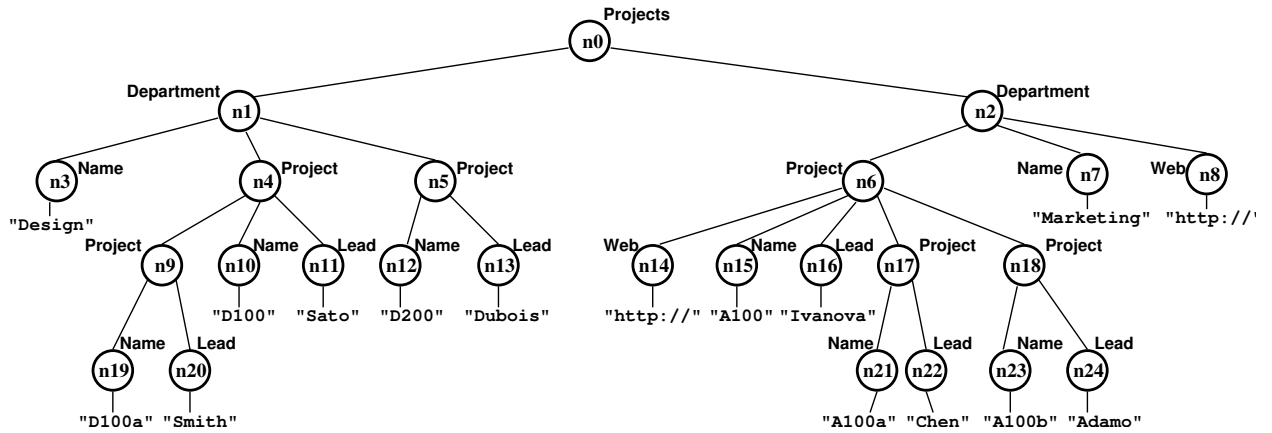
Physical Level

Tables, XML documents, indexes etc...

Still, ALL access is through declarative languages:

- SQL + dynamic generation of SQL code + EVAL operation
- SQL with uniform access to data and meta-data

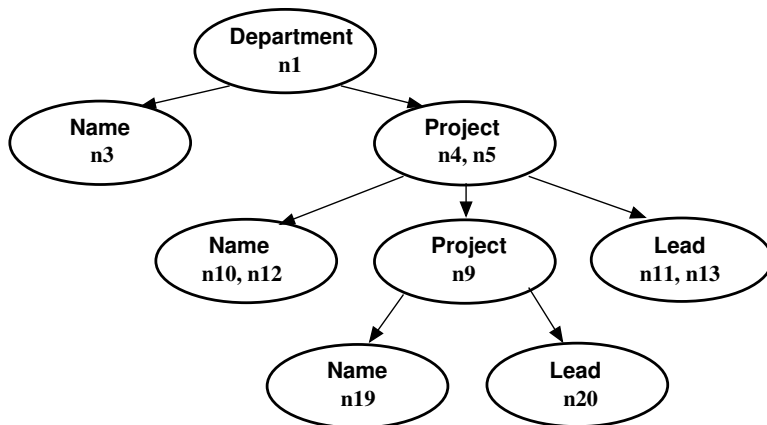
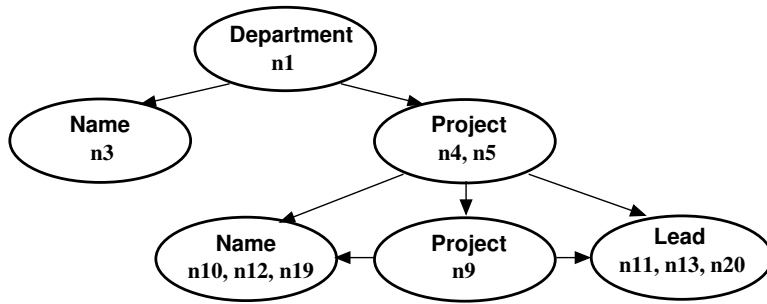
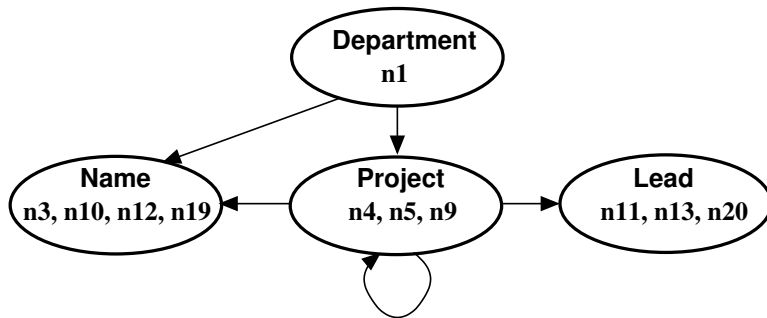
Search in semi-structured data



- Search language: (fragments) of XPath (navigational language)
 - XSLT
 - Xquery
- Index structures on XML documents for efficient access
 - Value-based indexes
 - Structural indexes

Structural Indexes

Partitions the nodes by similarity of structural information that surrounds them in the document.



XPath Fragments

- XPath 2.0
- Xpath 1.0
- Xpath with only child navigation
- Xpath with only parent navigation
- XPath without predication
- ...

Research: couple each XPath fragments with its optimal structural index.

- Theory: based on logic, set theory, meta-analysis
- System: How can these ideal couplings be implemented in such a way that XPath expressions can be evaluated “optimally” or “alternatively”.

Database Search versus Data Mining

Fundamental difference:

- Databases: search is in a polynomial space of data
 - Permits reasoning about **classes of searches** (SQL queries, Google queries)
 - General systems performing these searches.
- Data Mining: search is in an exponential space of data
 - Reasoning about a **single search**.
 - Best algorithms (with heuristics) for search in exponential space.
 - * empirical
 - * analytical
 - Can these algorithms be made generic for classes of related data mining tasks? (Library of polymorphic data mining algorithms.)

Frequent Pattern Mining

- The Market Mining Problem: Given a (large) list of baskets, each containing a set of items sold at the market, find combinations of items (item sets) that are frequently bought together.
- Graph mining: Given a (large) list of graphs, find subgraphs that occur frequently in these graphs (Chemical data mining)

The **frequency function** (**freq**) is a **measure** that computes how many times a given item set occurs in the baskets.

$$\begin{aligned}\text{freq}(I) &\geq \text{freq}(I \cup J) \\ \text{freq}(I \cup J) + \text{freq}(I \cap J) &\geq \text{freq}(I) + \text{freq}(J) \\ &\dots\end{aligned}$$

Many classes of measures are **strongly related** to frequency functions: probability functions, belief functions (reasoning about uncertainty), co-occurrence functions (text mining), diversity function (ecology, economics) etc.. Cross-disciplinary!

A mathematical framework for measures

- Computing bounds of frequencies.
- Deriving frequencies from known frequencies (Constraints)
- How can this theory be utilized in practical data mining algorithms working on real-data. (Proper utilization depends on how data is distributed.)
- How does an algorithm or a heuristic developed in one domain help search in data mining problems in other domains?
- Empirical versus analytical analysis. At least 50 frequent item set algorithms and 1000's of empirical data points to test this algorithms. Find common techniques used in algorithms and theoretically analyze them under various data distribution assumptions (verify and predict).