Data Integration

Hanna Zhong
hzhong@illinois.edu

Department of Computer Science
University of Illinois, Urbana-Champaign

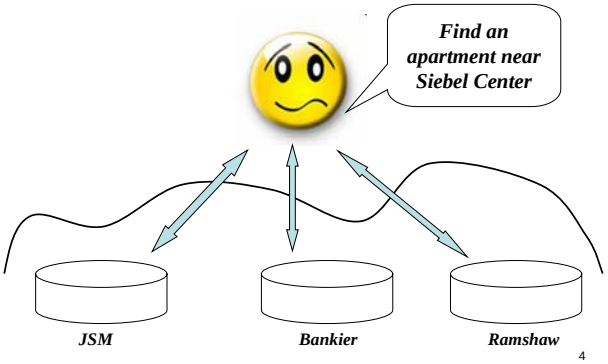
11/12/2009

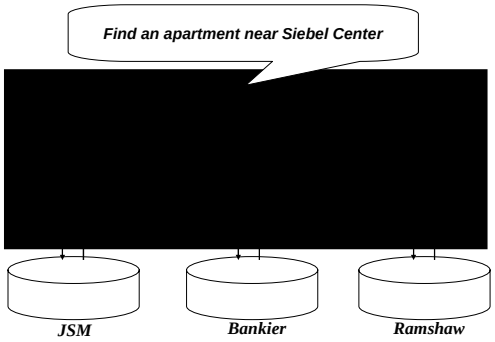
Overview

- Motivation
- Problem Definition
- Data Integration Approaches
  - Virtual integration
  - Data warehouse
- Issues
- Discussion

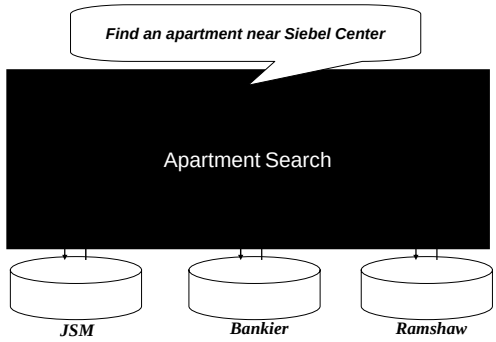
Why Data Integration?

Example: Apartment Search





5



6

More Examples

- People Search
  - Build a yellowpage application on db people
    - Many people doing database stuff in the US
    - How can we find information about a database person, such as classes taught, publications, collaborators, etc?
      - Homepages
  - http://dblife.cs.wisc.edu/

7

Example Systems

- Apartment Search
- DB People Search
- Etc…

8

Data Integration

- Arises in numerous contexts
  - on the Web, at enterprises, military, scientific cooperation, bio-informatics domains, e-commerce, etc.
- Currently very hot
  - in both database research and industry
- Current state of affairs
  - Mostly ad-hoc solution
    - create a special solution for every case; pay consultants a lot of money.

9

Overview

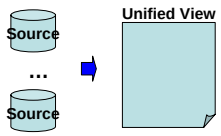
- Motivation
- Problem Definition
- Data Integration Approaches
  - Virtual integration
  - Data warehouse
- Issues
- Discussion

10

What is Data Integration?

The process of

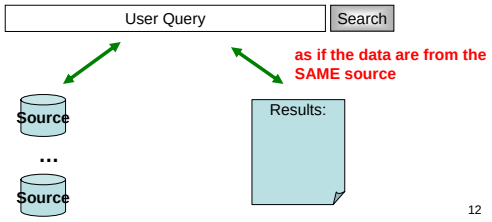
1. Combining data from different data sources
   - Data sources:
     - Databases, websites, documents, blogs, discussion forums, emails, etc
2. Presenting a unified view of these data



11

What is Data Integration? (2)

- The process of
  1. Combining data from different data sources
  2. Presenting a unified view of these data



12

Problem Definition

How can we access a set of **heterogeneous**, **distributed**, **autonomous** databases as if accessing a single database?

13

Data Integration is Hard ☹

- Data sources are heterogeneous, distributed, and autonomous
  - Sources Type
    - Relational database, text , xml, etc
  - Query-Language
    - SQL queries, keyword queries, XQuery
  - Schema
    - Databases have different schemas
  - Data type & value
    - The same data are represented differently in different sources
      - Type (e.g. *time* represented as varchar or timestamp)
      - Value (e.g. *8pm* represented as 8:00pm or 20:00:00)
  - Semantic
    - Words have different meanings at different sources (e.g. *title*)
  - Communication
    - Some sources are accessed via HTTP, others FTP

14

Overview

- Motivation
- Problem Definition
- Data Integration Approaches
  - Virtual integration
  - Data warehouse
- Issues
- Discussion

15

Event Search

- Provide a comprehensive search on Champaign-Urbana events in one place
  - Search events by its title, description, location proximity, dates, venues, and/or data sources

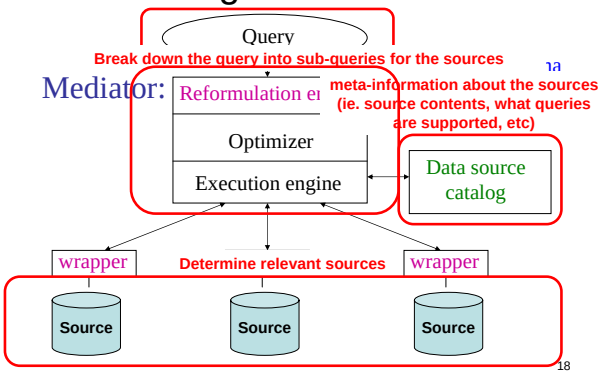
16

Virtual Integration Approach

- Leave the data in the sources
- When a query comes in:
  - Determine the relevant sources to the query
  - Break down the query into sub-queries for the sources
  - Get the answers from the sources, and combine them appropriately
- Data is fresh

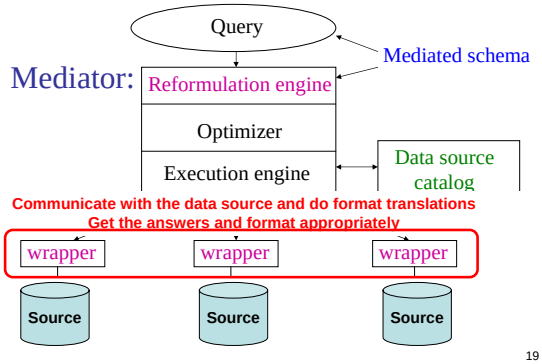
17

Virtual Integration Architecture



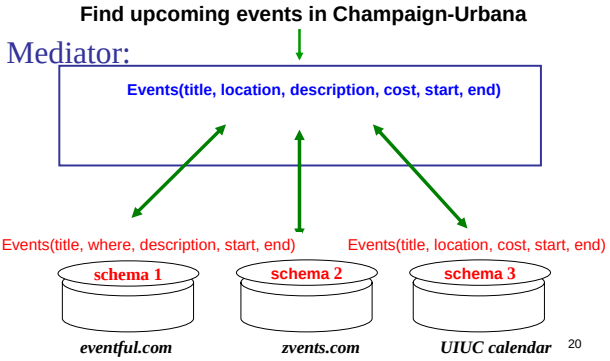
18

Virtual Integration Architecture



19

Virtual Integration Example



20

Virtual Integration Example

Find upcoming events in Champaign-Urbana

Mediator:

Events(title, location, description, cost, start, end)

title	where	description	start	end
Music	Canopy	Music	9pm	11pm

wrapper

Events(title, where, description, start, end)

schema 1 | schema 2 | schema 3

eventful.com | *zvents.com* | *UIUC calendar*

21

Virtual Integration Example

Find upcoming events in Champaign-Urbana

Mediator:

Events(title, location, description, cost, start, end)

title	location	cost	start	end
Music	Canopy Club	$5	8pm	11pm

wrapper | wrapper | wrapper

Events(title, location, cost, start, end)

schema 1 | schema 2 | schema 3

eventful.com | *zvents.com* | *UIUC calendar*

22

Challenges

Data Integration: the process of combining data from different data sources and presenting a unified view of these data

title	location	description	cost	start	end

Events(title, location, description, cost, start, end)

Schema Matching

title	where	description	start	end
Music	Canopy	Music	9pm	11pm

title	location	Cost	start	end
Music	Canopy Club	$5	8pm	11pm

23

Challenges

Data Integration: the process of combining data from different data sources and presenting a unified view of these data

title	where	description	cost	start	end

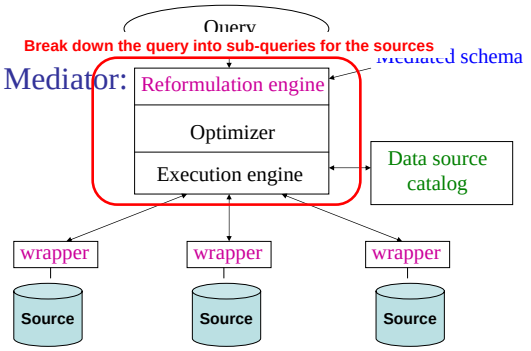
Events(title, location, description, cost, start, end)

title	where	description	start	end
Music	Canopy	Music	9pm	11pm

title	location	Cost	start	end
Music	Canopy Club	$5	8pm	11pm

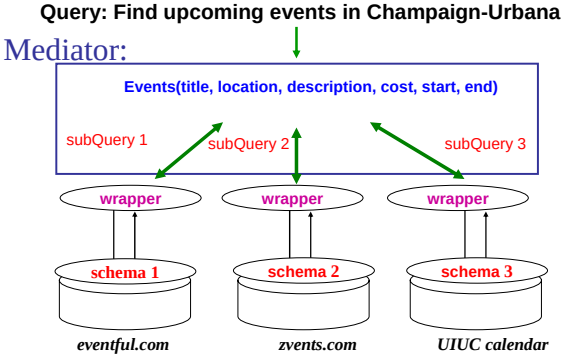
24

Virtual Integration Architecture

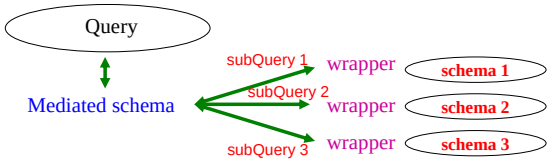


Virtual Integration

Query: Find upcoming events in Champaign-Urbana



Mediators



- Global-as-view
  - express the mediated schema relations as a set of views over the data source relations
- Local-as-view
  - express the source relations as views over the mediated schema

Global-as-View GAV

- Express the mediated schema relations as a set of views over the data source relations
  - The mediated schema is modeled as a set of views over the source schemas
- Design the mediated schema around the source schemas
- Mediated schema:
  Events(title, location, description, cost, start, end)
- Source schema:
  - S1: Events(title, where, description, start, end)
  - S2: Events(title, location, description, cost, start, end, performer)
  - S3: Events(title, location, cost, start, end)
- GAV:

```
Create View Events AS
select title, where AS location, description, NULL, start, end from S1
UNION
select title, location, description, cost, start, end from S2
UNION
select title, location, NULL, cost, start, end from S3
```

Global-as-View GAV (2)

- Adding sources is hard
  - The core work is on how to retrieve elements from the source databases
  - Need to consider all other sources that are available

29

Global-as-View GAV (3)

- Mediated schema:
  Events(title, location, description, cost, start, end)
  Venues(location, city, state)
- Source schema:
  - S4: Events(title, description, city, state)

GAV:

```
Create View Events AS
select title, NULL, description, NULL, NULL, NULL from S4

Create View Venues AS
select NULL, city, state from S4
```

What if we want to find events that are in Champaign?

30

Local-as-View LAV

- Express the source relations as views over the mediated schema
- The mediated schema is already designed
  - Create views on the source schemas
- Mediated schema:
  Events(title, location, description, cost, start, end)
- Source schema:
  - S1: Events(title, where, description, start, end)
  - S2: Events(title, location, description, cost, start, end, performer)
  - S3: Events(title, location, cost, start, end)
- LAV:

```
Create View S1
select title, location AS where, description, start, end from Events

Create View S2
select title, location, description, start, end, NULL from Events

Create View S3
select title, location, cost, start, end from Events
```

31

Local-as-View LAV (2)

- Mediated schema:
  Events(title, location, description, cost, start, end)
  Venues(location, city, state)
- Source schema:
  - S4: Events(title, description, city, state)

What if we want to find events that are in Champaign?

LAV:

```
Create View S4
select title, description, city, state
from Events e, Venues v
where e.location=v.location AND city= "champaign"
```

32

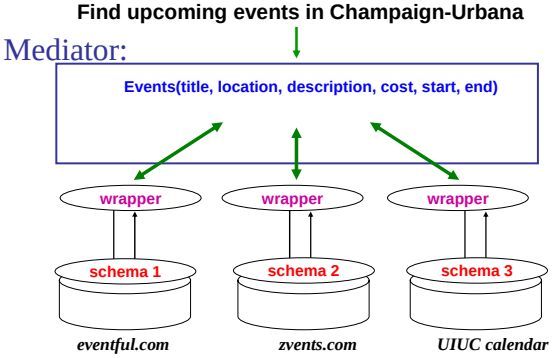
Local-as-View LAV (3)

- Very flexible.
  - You have the power of the entire query language to define the contents of the source.
  - Hence, can easily distinguish between contents of closely related sources
- Adding sources is easy
  - They're independent of each other

33

Virtual Integration Example

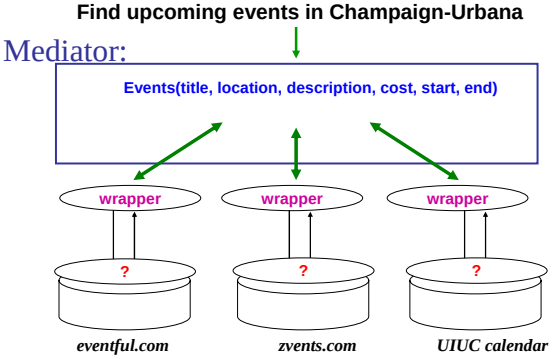
Find upcoming events in Champaign-Urbana



34

What if Schema is Unknown?

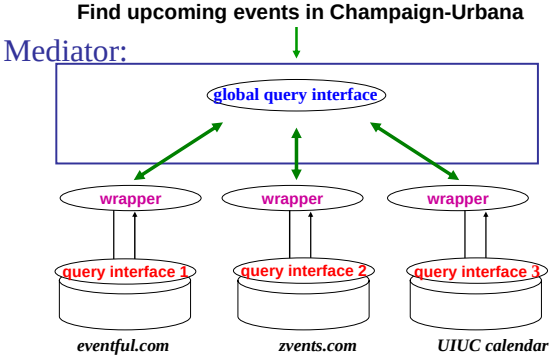
Find upcoming events in Champaign-Urbana



35

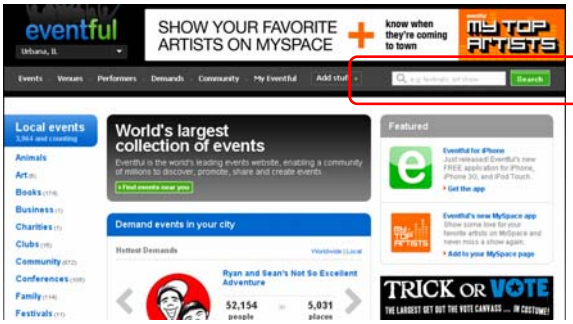
Query Interface

Find upcoming events in Champaign-Urbana



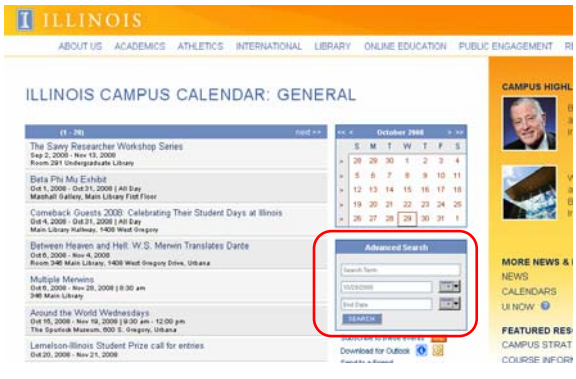
36

Query Interface



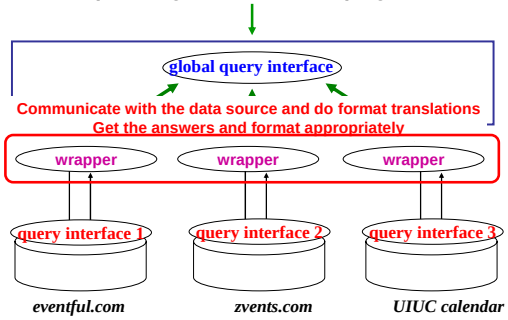
37

Query Interface



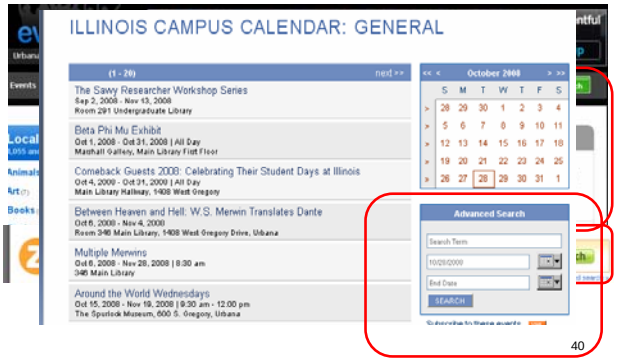
Wrappers

Find upcoming events in Champaign-Urbana



39

Wrappers



40

Wrappers (2)

- Once the query is submitted via the query interface, results are returned
  - Formatting is specific to each source

41

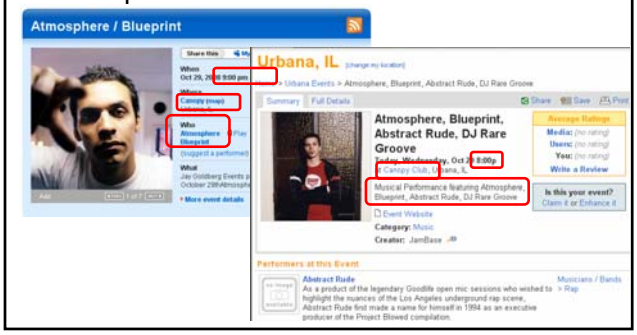
Wrappers (2)

Information Extraction

- What information to extract?

Information Extraction (2)

- Complications…



Wrappers

- Hard to build and maintain (very little science)
- Major approaches
  - Machine Learning
  - Data-intensive, completely-automatic
    - Roadrunner
      http://portal.acm.org/citation.cfm?doid=564691.564778
- Data sources are accessed via query interfaces
  - Query interface to each data source is different
- Scalability
  - One wrapper per source vs one wrapper per domain

45

Overview

- Motivation
- Problem Definition
- Data Integration Approaches
  - Virtual integration
  - Data warehouse
- Issues
- Discussion

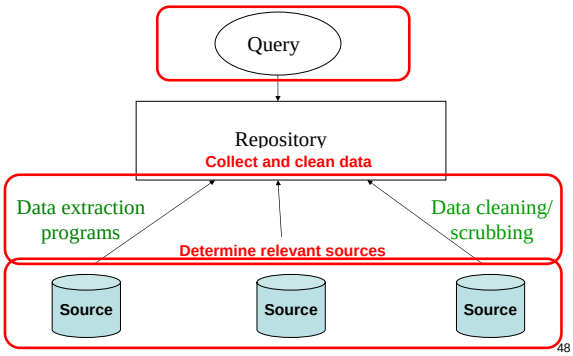
46

Data Warehousing

- Load all the data periodically into a central database (warehouse)
  - Performance is good
  - Data may not be fresh
  - Need to clean, scrub you data

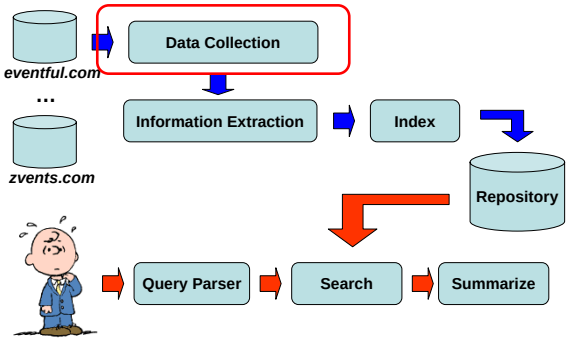
47

Data Warehousing Architecture



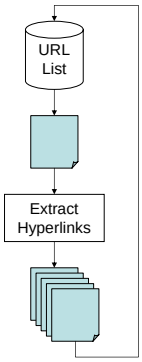
48

Data Warehousing Example

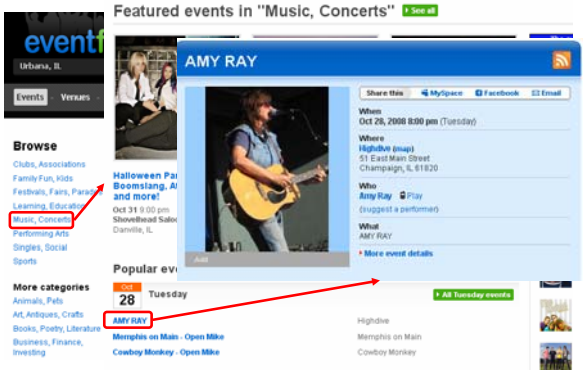


Data Collection

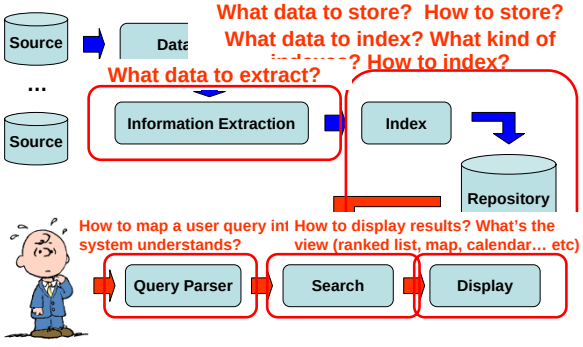
- RSS feed
- Crawlers
  - Programs that browse the Web in a methodical, automated manner
    - Link following



Crawlers (2)



Data Warehousing Example



Data Integration Approaches

- Virtual integration
  - No data are collected offline
  - On a search, data are collected and processed from various sources at runtime
- Data warehouse
  - Data are collected offline and stored in a central repository
  - Search is performed on the repository
- When should we take the virtual integration approach?
- When should we take the warehousing approach?

53

Overview

- Motivation
- Problem Definition
- Data Integration Approaches
  - Virtual integration
  - Data warehouse
- Issues
- Discussion

54

Data Integration Issues

- Data collection
  - Wrappers, crawlers, RSS
  - Duplications, spams
  - Freshness, completeness, etc
- Information extraction
  - What information to extract?
- Schema matching
- Query optimization
- Query reformulation
- Scalability
  - When there are many sources out there, does the solution still work?

55

Discussion

56

References

- Some slides taken from Professor Anhai Doan, from FALL2005 CS511, UIUC

57