Virtual Web Data Integration on GRID

Supervised by

Prof. R. K. Ghosh and Prof. S. K. Aggarwal

CS728 – Term Paper

Ravi Kumar Mandala

Janardhan Rao

Overview of our Presentation

- Introduction
- Challenges and Benefits
- Issues and Motivation
- Proposed Solution
- Overview of our Architecture
- Components of our Architecture
- Salient features
- Related work
- Conclusions
- Future Work
- References

Introduction: Virtual Web Data Integration on GRID

- Effective extraction, management and processing of web information on GRID
- Enabling applications/ users to access data as if it is in single database, though it is not.

Challenges and Benefits

Challenges:

- Identifying relevant information from multiple sources
- Querying the sources for results “on-the-fly”
- Transforming and mining data for analysis
- Integrating diverse information within cost & skill constraints
- Information availability, Timeliness and Access performance

Benefits:

- Aids quick business decision making (reduces analysis time and effort)
- Even applications can now access “The Web Data Repository”

Issues

- Unstructured and Semi-structured documents (XML, HTML, eMails, PDF)
- Changing information sources (Data, Schema and/or constraints)
- Pruning irrelevant sources (necessary for scaling-up)
- Insert conditions in query plans to be used in pruning dynamic bindings
- Incremental Web Warehousing
- Autonomy, heterogeneous platforms
- Semantic Interoperability Problem
- Provenance information (Should give audit trail/derivation history of data)

Motivation

- Searching is second most frequent web usage next to e-Mail service.
- Not only users, but also Web applications, PDAs, Mobiles need to interact with search engines
- Need for the flexible, high performance access to all significant data

Some search engines and Meta search engines

- Google
- Yahoo
- Altavista
- Lycos
- Webcrawler
- AskJeeves

- HotBot
- Infoseek
- MyExcite
- MSN Search
- AOL Search
- IceRocket

- MetaCrawler
- DogPile
- GenieKnows
- FastSeeks
- iBoogie
- metaEUREKA

Objectives

- Content integration, Data format conversion
- Opening up access to islands of information
- Mine cached web (in case source is DOWN/ expensive)
- Looking for trends, data correlations
- Generic architecture instead of domain-specific one

Example: Travel Information Service

Provides information regarding…

Air Fares, Airlines Info, Maps, Hotel availability, Place Info etc.,

Heterogeneities to be dealt…

- Technical
  - OS, Hardware
- Interfaces
  - Access, Language
- Semantic
  - Different names for same concept
- Structural
  - Attributes are grouped into different tables

Proposed Solution: Virtual Integration Architecture

- Analyze sources and generate schema-guided wrappers
- Leave the data where they are
- When query comes in...
  - Determine relevant sources
  - Breakdown the query and send sub-queries to sources
  - Get answers from sources and integrate them

Our meta-search engine based information integration system with deep web crawling aims at providing users a single comprehensive view of seamless information that cannot be easily accomplished by a single search engine.

More about our Virtual Integration Architecture

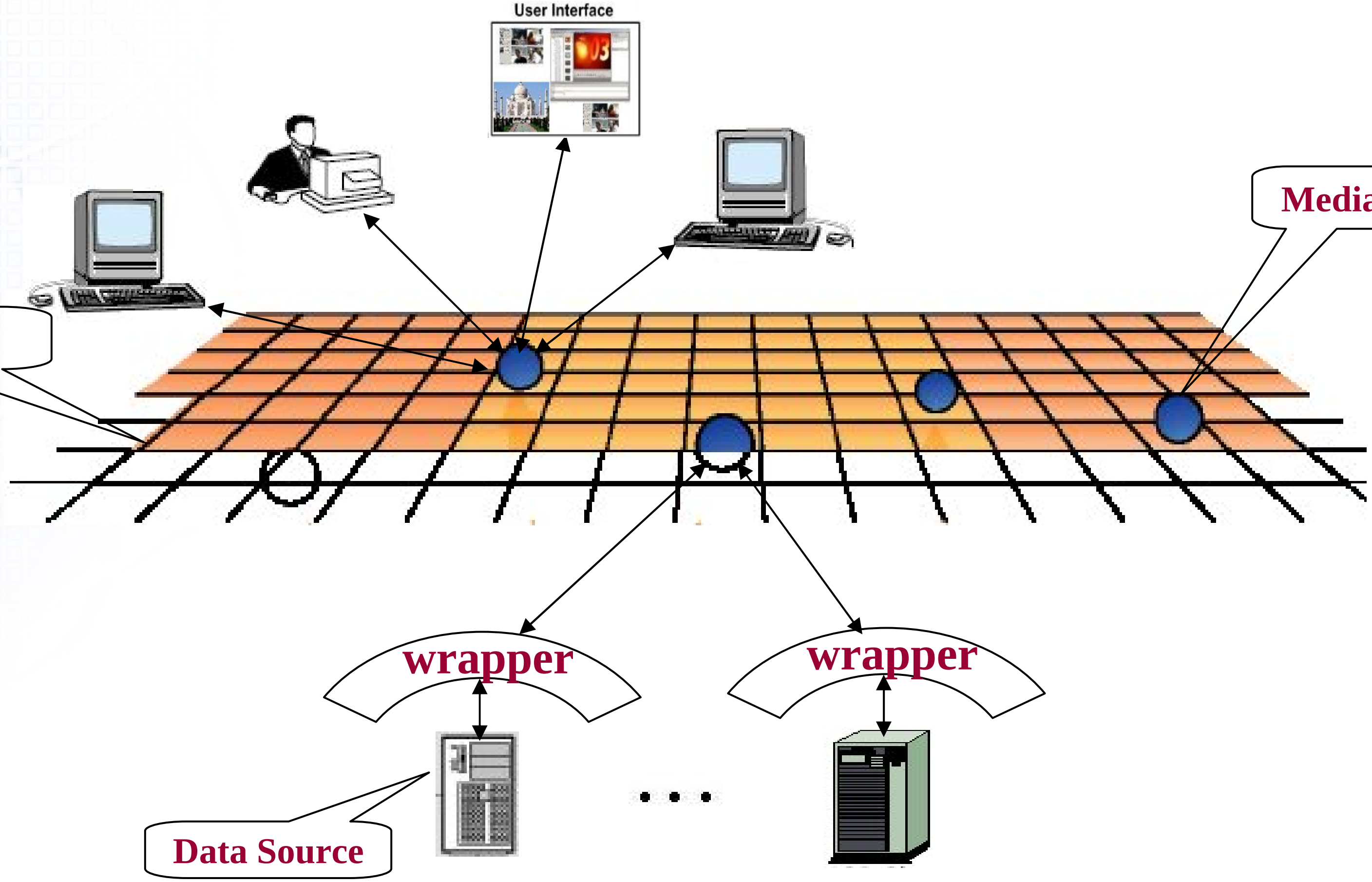
- Set-up time: Initial set-up
  - Analyze sources and generate default wrappers
  - Extract, transform, clean information. Leave the data where they are
- Query processing time: When query comes in...
  - Get user-defined schema and restructure query
  - Breakdown the query and send execution plan to wrappers
  - Get answers from sources and combine them
- Operation time: Handle changes in ([non] cooperating) sources
  - Modifications of the sources are filtered for relevance
  - Propagated changes to upgrade other sources using registry

Solution with existing Technologies

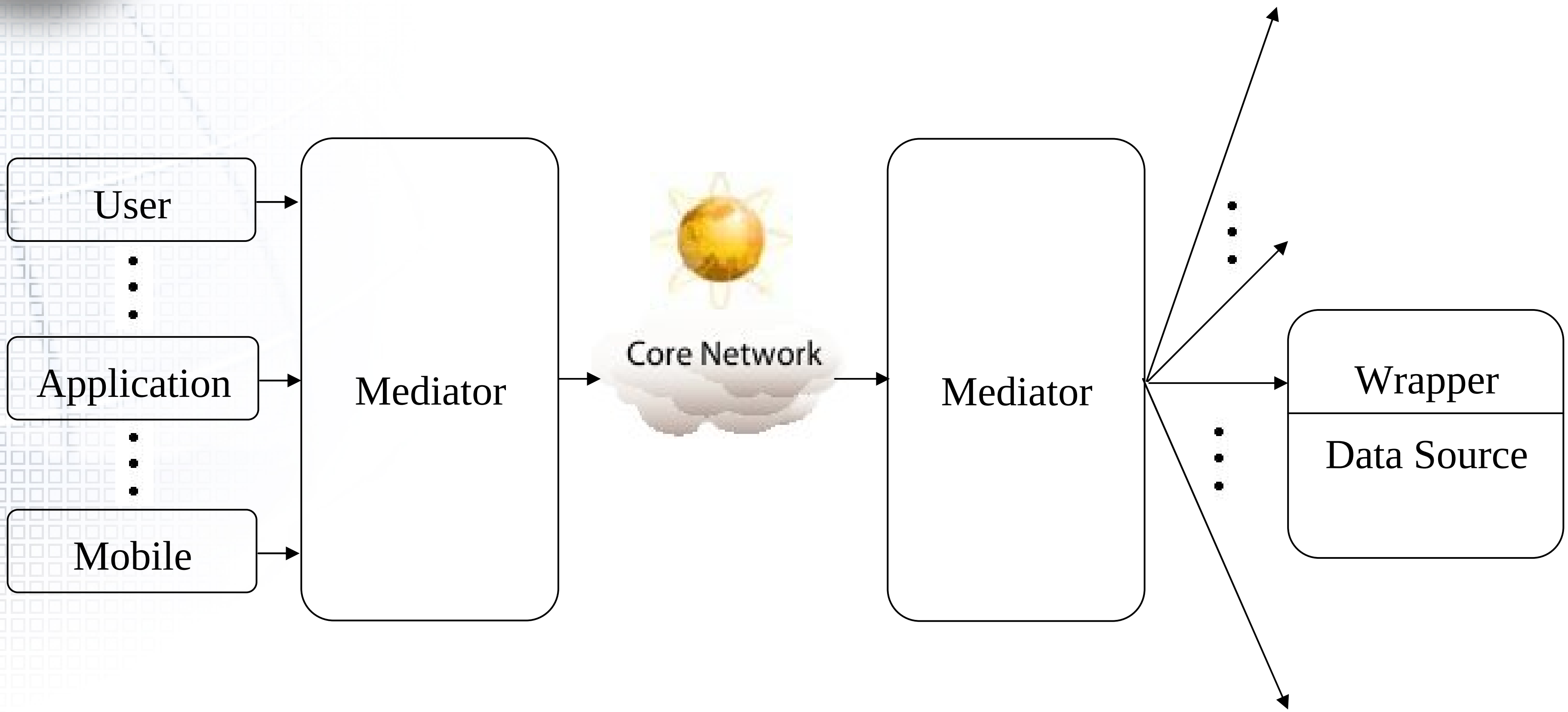
Wrappers	-	Heterogeneity of platforms and meta-data maintenance
XML	-	Heterogeneity of source documents
SOAP	-	Distribution
WSDL	-	Service interoperability between sources (Use IDL)
UDDI	-	Publication & Discovery of Information
RDF	-	Semantics interoperability of Sources
OWL	-	Ontology Language for semantics of documents on web

- ✓ Find a new source of information using UDDI
- ✓ Understand its syntax using WSDL
- ✓ Understand its semantics using RDF
- ✓ Get/ access it using SOAP
- ✓ The information in XML, can be restructured and integrated automatically

Overview of our Architecture



CnC view of the Architecture

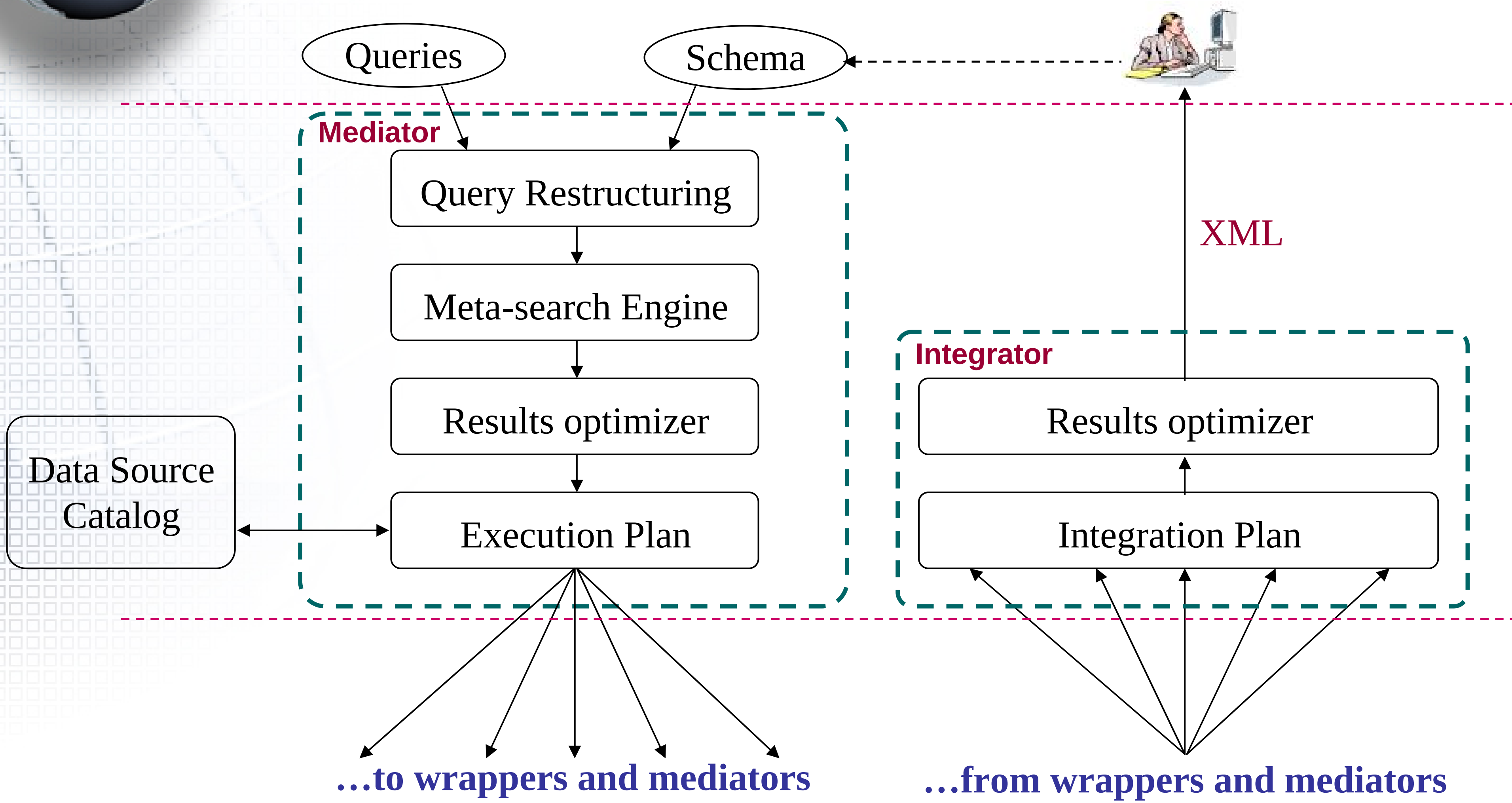


Mediation Approaches

- Query-based Mediation: Partition queries into steps executed at sources
  - Simpler to implement
- View-based Mediation: Find ways to construct given query from global views
  - Determine relevant sources
  - Breakdown the query and send sub-queries to sources
  - Get answers from sources and integrate them

Our meta-search engine based information integration system with deep web crawling aims at providing users a single comprehensive view of seamless information that cannot be easily accomplished by a single search engine.

Mediator and Integrator



Wrappers

To transform unstructured data to semi-structured data
(or semi-structured data to structured data)
The result (XML) can be queried (XQuery) and analyzed (XSLT)

Issues:
(1) Identify semantics of data in an HTML document
(2) Establish the mappings between its structure and semantics.

Need for wrappers:
- ✓ Enhanced way to get relevant info
- ✓ Integrated access to heterogeneous source types

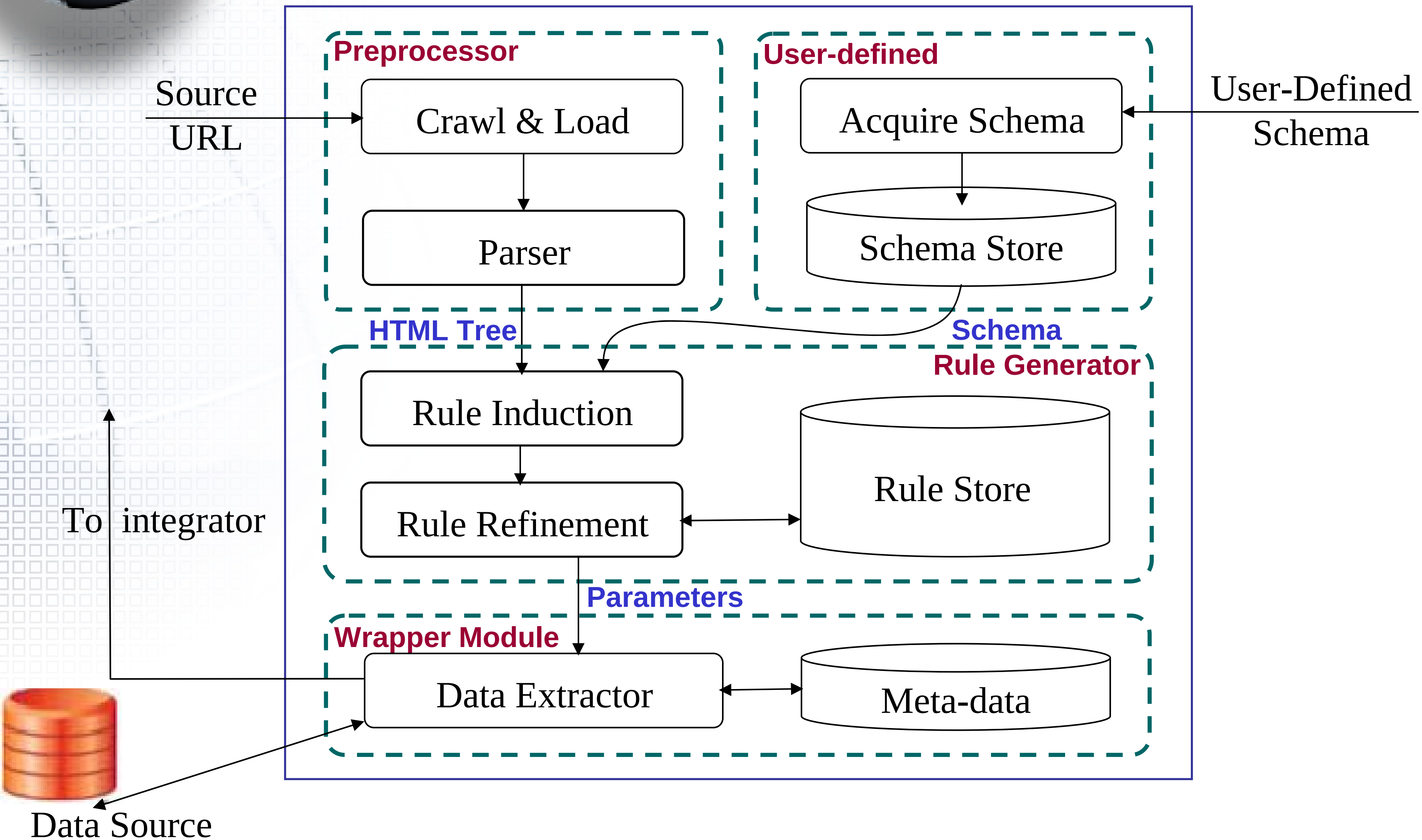
Wrapper Generation methods

- Manual generation (M)
- Semi-automated generation (SA)
  - Supervised methods (Schema-guided)
  - Machine learning methods
- Fully automated generation (A)

Our Wrapper Generation (S/A)

- If the source is previously unexplored, generate new wrapper.
- Generating extraction rules with the guidance of user defined schema
- The wrapper generated based on the rules could be more accurate and better reflect the users requirements.
- Schemas provide query-specific parameters/ rules for the wrapper

Wrapper Architecture



Data Source Wrapper

- Preprocessor
  - Document Loader
  - Parser
- Schema guide
  - Schema Collector
  - Schema Store
- Rule Generator
  - Rule Inducer
  - Rule Refiner
- Data Extraction
  - Data & semantic extraction
  - Meta-data extraction

Preprocessor

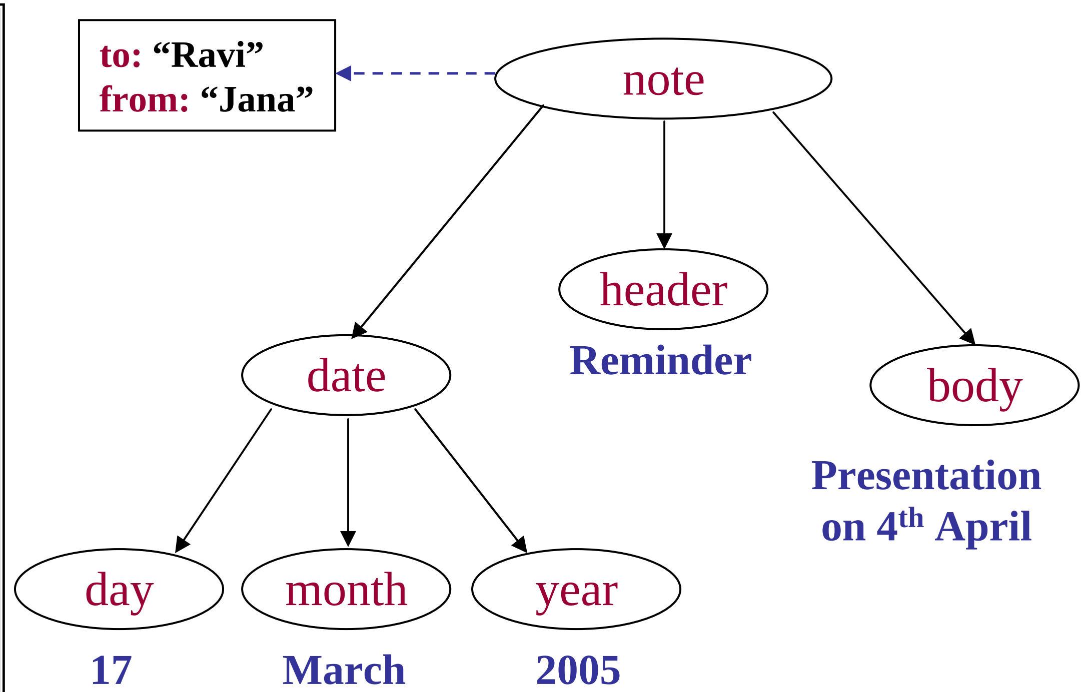
- Sets up the environment for the system
- Fetches Web page using user-given URL
- Parses the document (Handles HTML tag errors)
- Generates HTML/XML trees, Semantic trees, meta-data

HTML page is represented as a tree using the document object model (DOM)

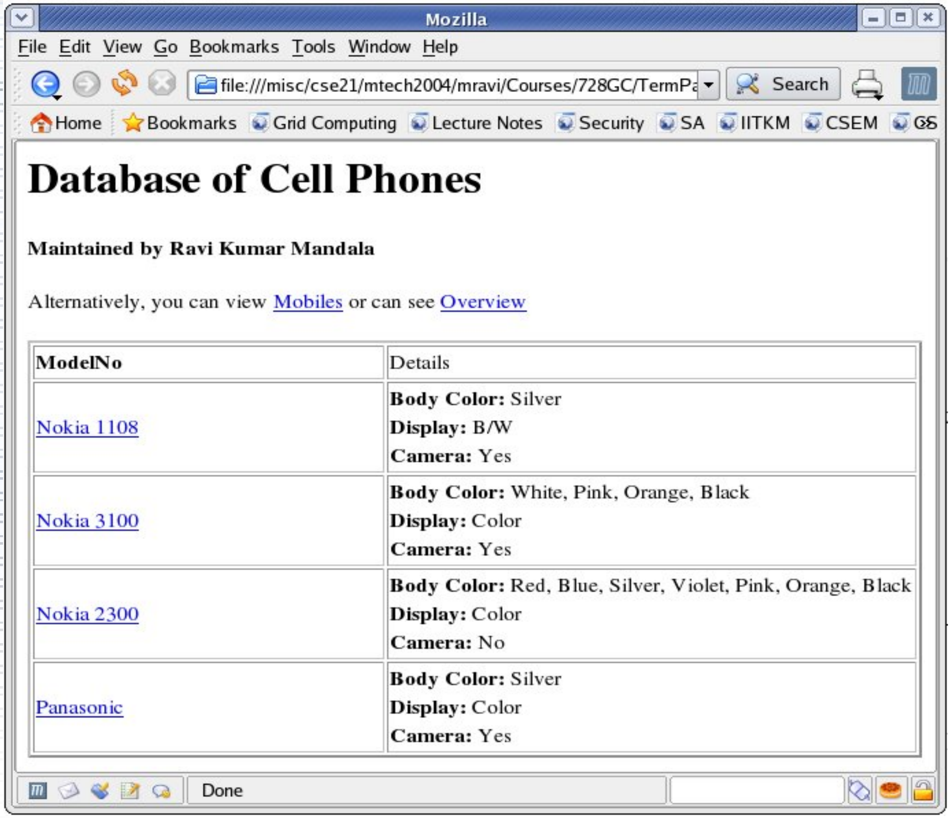
Data item is leaf node and its path gives its location

XML Tree Generation: Example

```
<?xml version="1.0" ?>
<note to="Ravi" from="Jana">
<date>
    <day>17</day>
    <month>March</month>
    <year>2005</year>
</date>
<header>Reminder</header>
<body>
    Presentation on 4th April
</body>
</note>
```



Semantic Tree Generation: Example

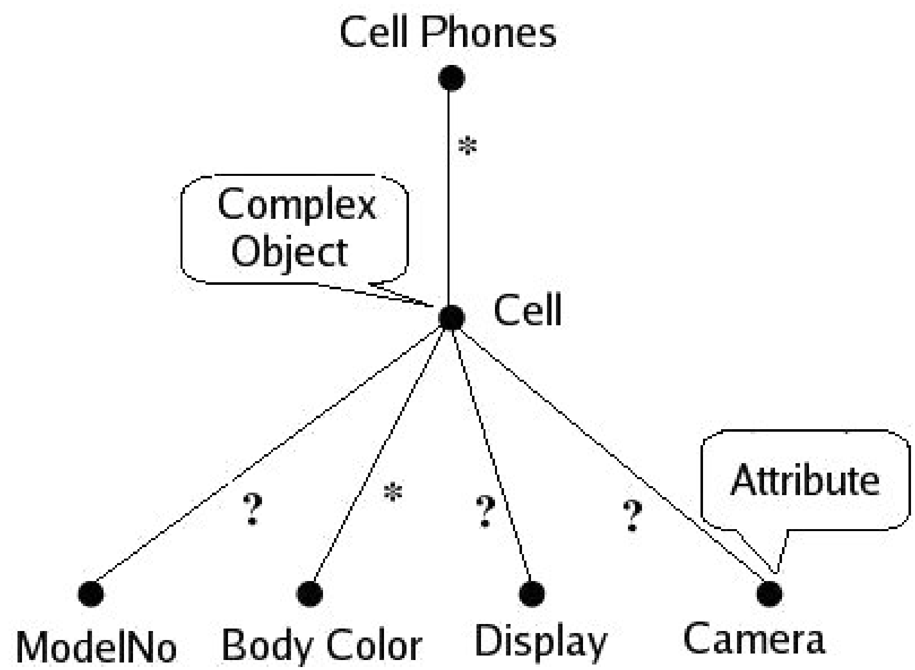


Database of Cell Phones

Maintained by Ravi Kumar Mandala

Alternatively, you can view Mobiles or can see Overview

ModelNo	Details
Nokia 1108	**Body Color:** Silver **Display:** B/W **Camera:** Yes
Nokia 3100	**Body Color:** White, Pink, Orange, Black **Display:** Color **Camera:** Yes
Nokia 2300	**Body Color:** Red, Blue, Silver, Violet, Pink, Orange, Black **Display:** Color **Camera:** No
Panasonic	**Body Color:** Silver **Display:** Color **Camera:** Yes



Adoption of rich ontology languages of Semantic Web with good lexicalizations and links to reference ontology would help in semantic tree generation.[3]

User defined Schema Guidance

- ✓ User profiling is best known approach for knowing user preferences and requirements.
- ✓ Some machine learning techniques are under research
- ✓ Default preferences and schema is used
- ✓ Topic categories and associated common keywords are specified
- ✓ User specifies query parameters like …
  - Domains of interest
  - Types of files to be explored
  - Freshness required
  - Where should my keyword occur in the webpage
  - Languages of interest
  - Filters (like safety, …)
  - Granularity of provenance information required

Rule Generator

- In manual wrapper generation these two stages are repeated several times

Rule Inducer

- Combines preprocessor output with user-schema

Rule Refiner

- Removes inconsistent rules
- Performs Semantic Integration at data level
  - Raw components are collected
  - Combine Semantically equivalent (representing same real-world object) data
    - ✓ Disambiguation or conflict removal
  - Abstraction
  - Augment additional information

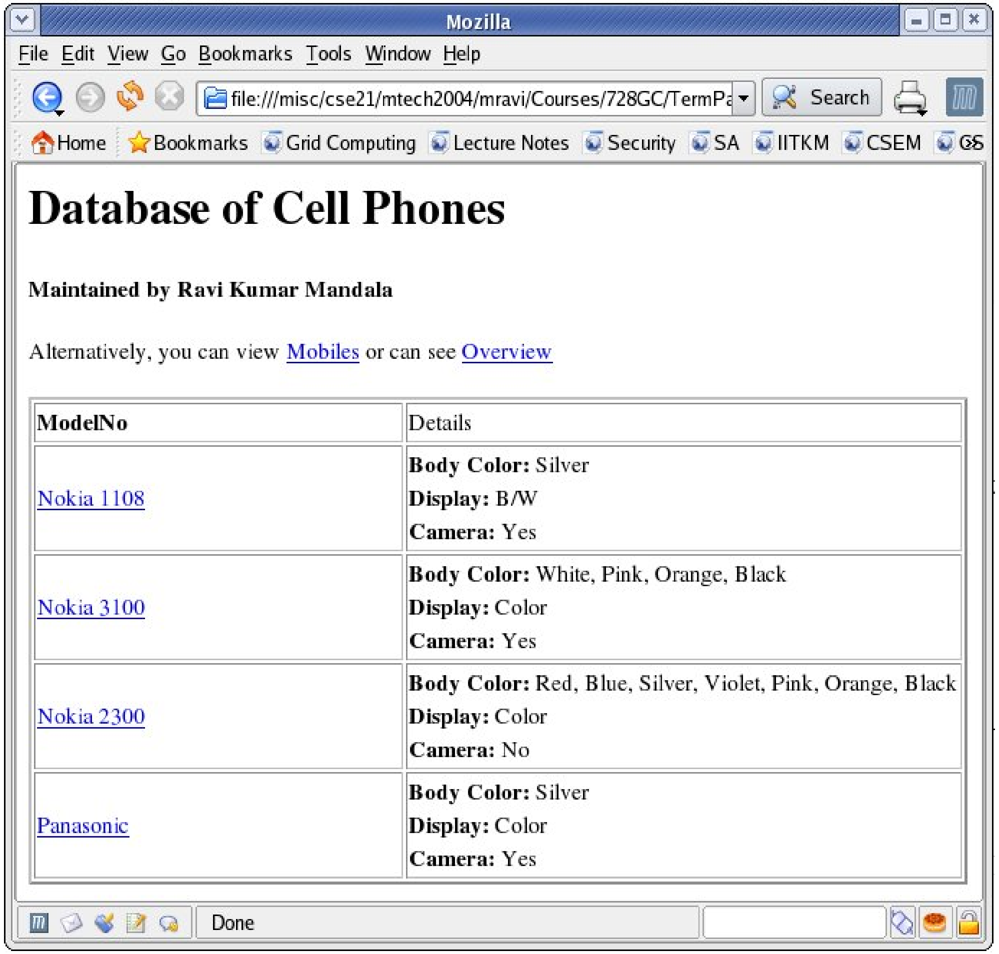
Data rule Extraction: Example

A sample data extraction rule for the example HTML page is as follows:

```
CellPhones:html {
        Cell: body.table[0].tr[]{
                ModelNo:     a[0].txt;
                BodyColor:   table[0].tr[0].td[0].txt;
                Display:     table[0].tr[1].td[0].txt;
                Camera:      table[0].tr[2].td[0].txt;
        }
        …
}
```

Mapping captured from the example:

html.body.table[0].tr[0].a[0] <=> CellPhones.Cell.ModelNo "Nokia 1108“



Database of Cell Phones

Maintained by Ravi Kumar Mandala

Alternatively, you can view Mobiles or can see Overview

ModelNo	Details
Nokia 1108	**Body Color:** Silver **Display:** B/W **Camera:** Yes
Nokia 3100	**Body Color:** White, Pink, Orange, Black **Display:** Color **Camera:** Yes
Nokia 2300	**Body Color:** Red, Blue, Silver, Violet, Pink, Orange, Black **Display:** Color **Camera:** No
Panasonic	**Body Color:** Silver **Display:** Color **Camera:** Yes

Data extraction

Data extraction guided by given set of parameters, is fairly simple task

Can use Topic maps for Data extraction and Knowledge management (Generation, codification, transfer)

- Similar to Semantic Networks
- Indexing, library science and knowledge representation with advanced techniques of linking and addressing
- Indexing: "see", "see also" reference
- Thesaurus: "Narrow term", "Broader term", "user for", "related term"
- Topic name, Topic Type
- Occurrence, Occurrence roles/ role types (Eg. portrait, example, definition)
- Association, Association roles/ types (Eg.written_by, born_in, guide_of)

Resource Meta-data

USE: To process user query and use past results for future processing

Meta-data can be viewed in 3 different perspectives

- Conceptual: Semantic information (using Web Ontology Language OWL)
- Logical: Structure of data in source documents (HTML/ XML Trees)
- Physical: How to locate & access data (SOAP to locate the source & Sitemap)

Salient features

- ✓ Efficient management of data and Meta-data in dynamic-distributed environments
- ✓ On demand creation of databases, federation of existing databases, on demand translation from one data model to another and querying of data in a distributed environment
- ✓ Provides flexible, high performance access to all significant and federated resources
- ✓ Facilitates trend analysis and business decision making
- ✓ Even applications can now access “The Web Data Repository”

Related Work

1. GriPhyN – Grid Physics Network
2. Chiron – Quarknet Education Project
3. GridAnt – Flexible workflow model based on Jakarta Ant

Conclusions

We have investigated and proposed frameworks that best fit in the context rather than reinventing-the-wheel. Semantic Integration has satisfactory results and is active areas of research. Usage of these techniques ought to produce better result which can be realized with proper implementations. In short, Information integration is an excellent resource to be harnessed.

Future Work

To simulate the system in a test environment with real-time difficulties and test performance

References

1. Structured and Semi/Unstructured Information Retrieval, Integration and Management - Mukesh Mohania, IBM India research Lab
2. Anywhere to Anywhere data integration – PowerXML
3. Ontology-Driven Geographic Information Systems – Frederico T. Fonseca and Max J. Egenhofer
4. News Item Extraction for Text Mining in Web Newspapers – Kjetil, Randy
5. Knowledge Management in Heterogeneous Environments - Larry Kerschberg
6. Information integration: A research agenda – Jhingran, Mattos, Pirahesh
7. Complex Queries over Web Repositories: Sriram Raghavan, Hector Garcia-Molina, C.S.E-Stanford University
8. Data warehousing and Data Mining – Michlen and Kamber
9. Building the Data Warehouse (3rd Edition) – W. H. Inmon, W. H. Inmon

Questions?

THANK YOU