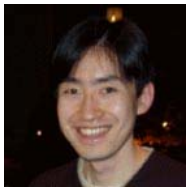
Division of Computational Informatics
Computational Intelligence Group

Hiroyuki Kitagawa
Center for Computational Sciences
Graduate School of Systems and
Information Engineering

Members



- Faculty
  - Hiroyuki Kitagawa (Professor)
  - Toshiyuki Amagasa (Assistant Professor)
  - Hideyuki Kawashima (Assistant Professor)
    (Joined CCS in February 2007)
- Postdoctoral Researchers
  - Yousuke Watanabe (JST/CREST)
  - MoonBae Song (JSPS Postdoctoral Fellow)
- Students
  - Doctoral Program: 5
  - Master Program: 13
  - Undergraduate: 5
  - Research Student: 1

Overview

- Management and utilization of databases and large datasets
- R&D in data engineering and databases
- Main Research Areas
  - Meteorological Databases
    - GPV/JMA data archive
  - Information Integration
    - Integrated use of different types of information sources: Databases, Web, Files, Sensors, …
  - Data Mining and Knowledge Discovery
    - Extraction of useful information from databases and web
  - XML
    - XML: Standard format for data interoperability
    - XML data management and databases

Collaboration

Division of Global Environment and Biological Sciences
Global Environmental Science Group

Division of Computational Informatics
Computational Media Group

- Development and Maintenance of GPV/JMA Data Archive

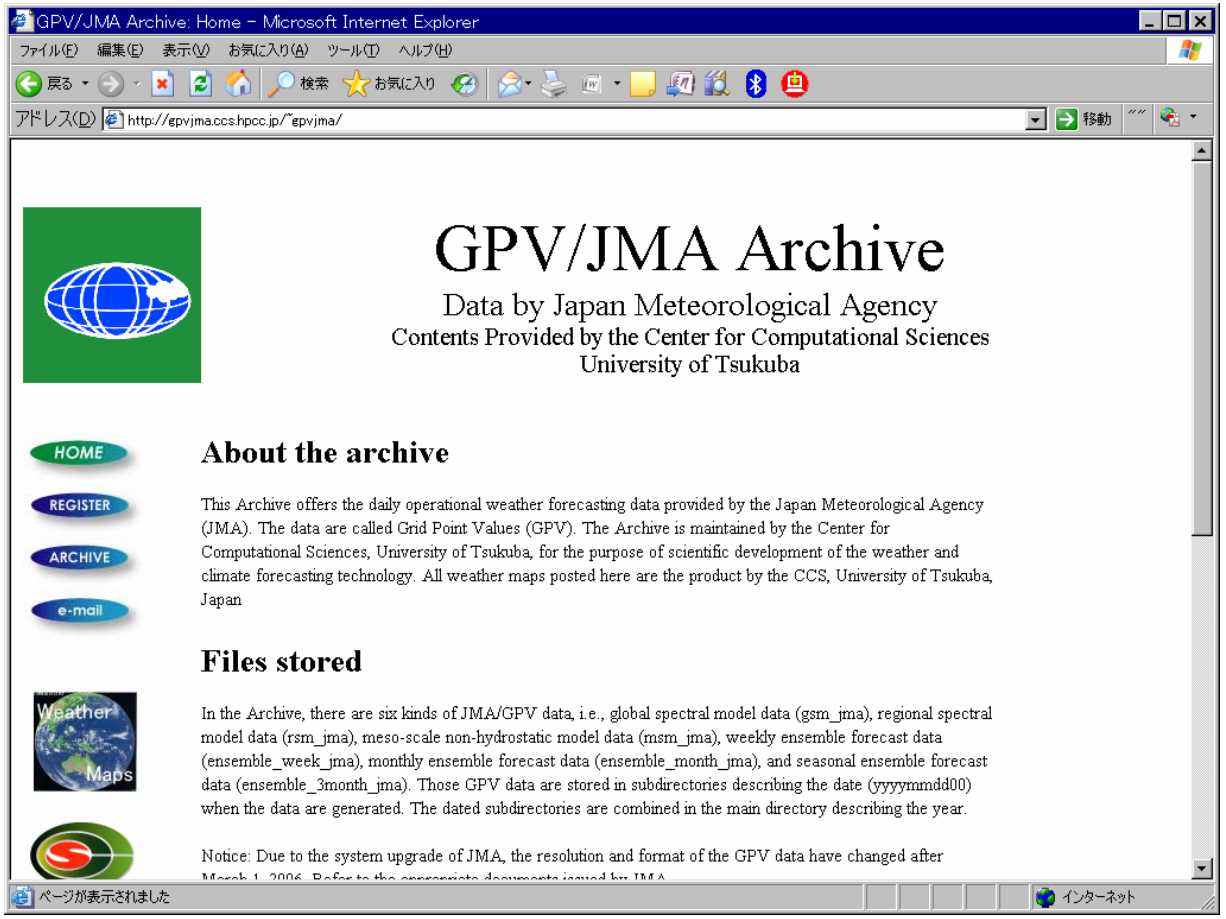
- Monthly Division Meeting (Computational Informatics Seminar)
- Real-world Sensing Data Management

Division of Computational Informatics
Computational Intelligence Group

Research Areas

- ✓ Meteorological Databases
  - GPV/JMA data archive
- Information Integration
  - Integrated use of different types of information sources: Databases, Web, Files, Sensors, …
- Data Mining and Knowledge Discovery
  - Extraction of useful information from databases and web
- XML
  - XML: Standard format for data interoperability
  - XML Data management and databases

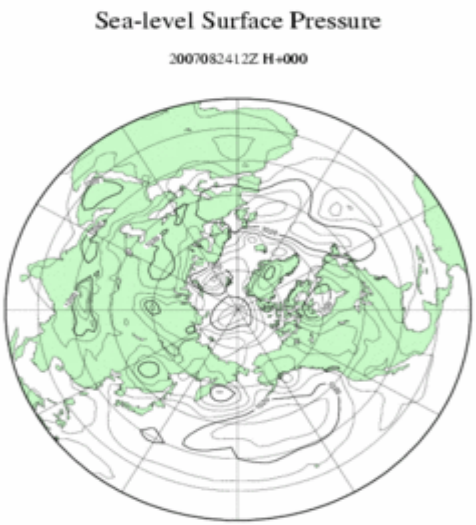
GPV/JMA Data Archive
http://gpvjma.ccs.hpcc.jp



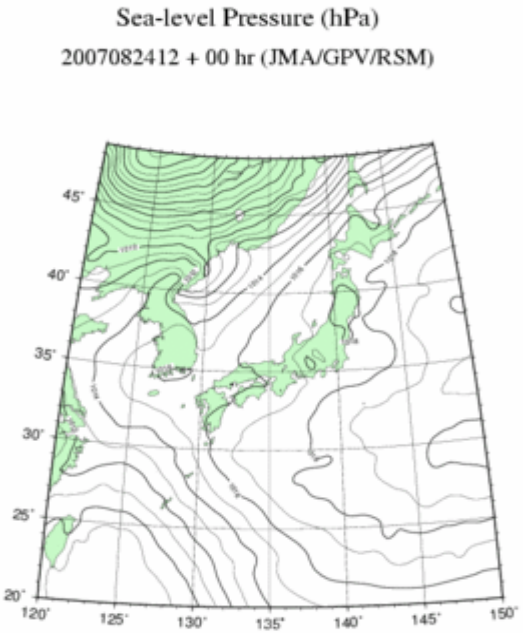
Developed and maintained in collaboration with Global Environmental Science Group since Jan. 2005

GPV/JMA Data Archive

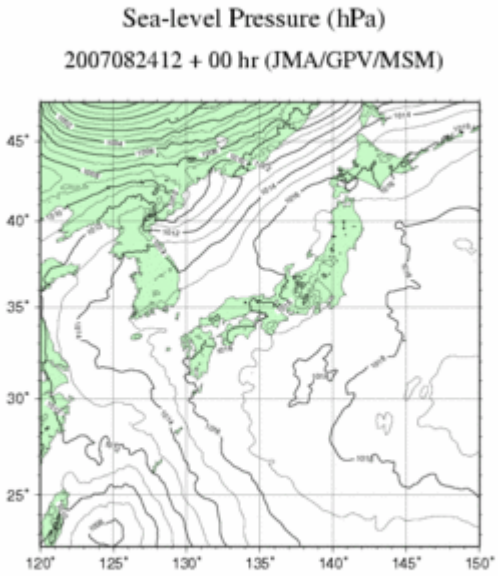
- Daily operational weather forecasting data (Grid Points Values Data (GPV)) provided by the Japan Meteorological Agency (JMA)
- For scientific development of weather and climate forecasting technology



GSM:
Global Spectral Model



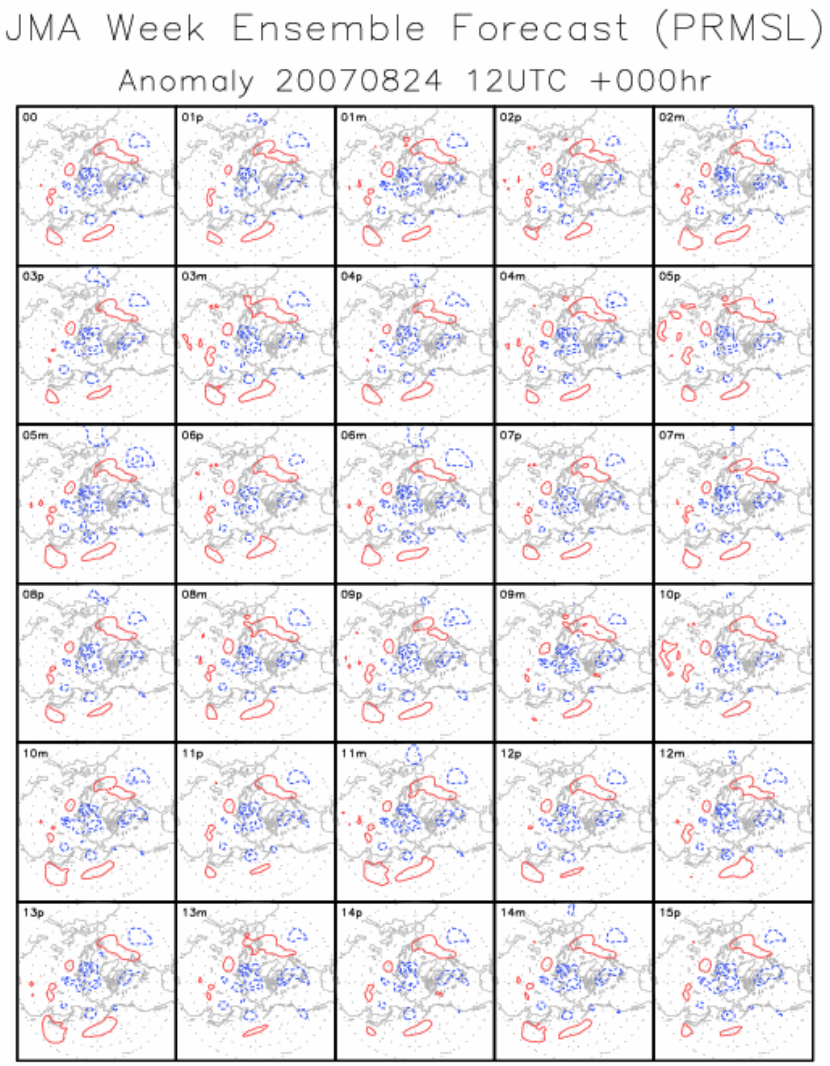
RSM:
Regional Spectral Model



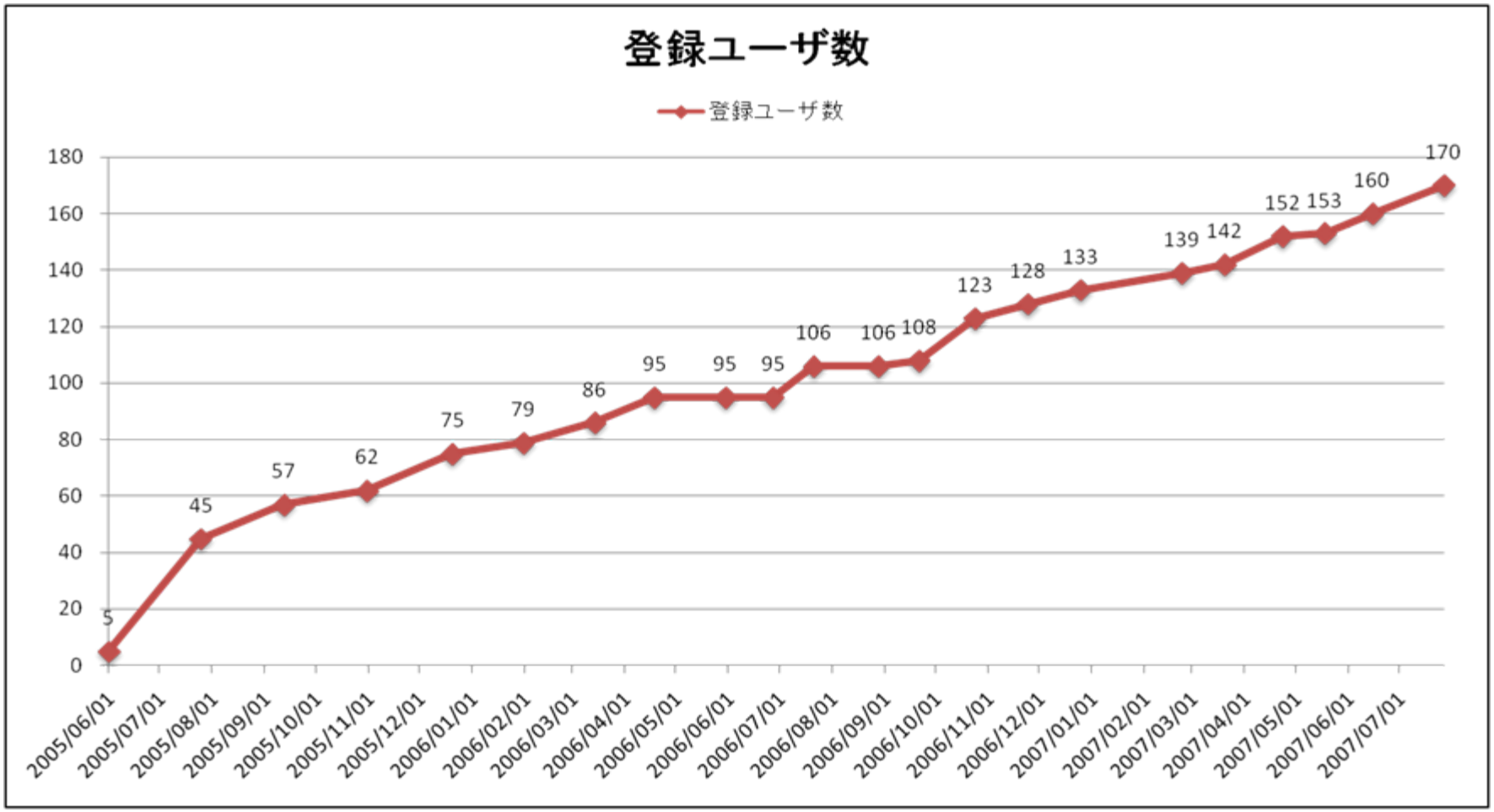
MSM:
Meso-Scale non-hydrostatic Model

GPV/JMA Data Archive

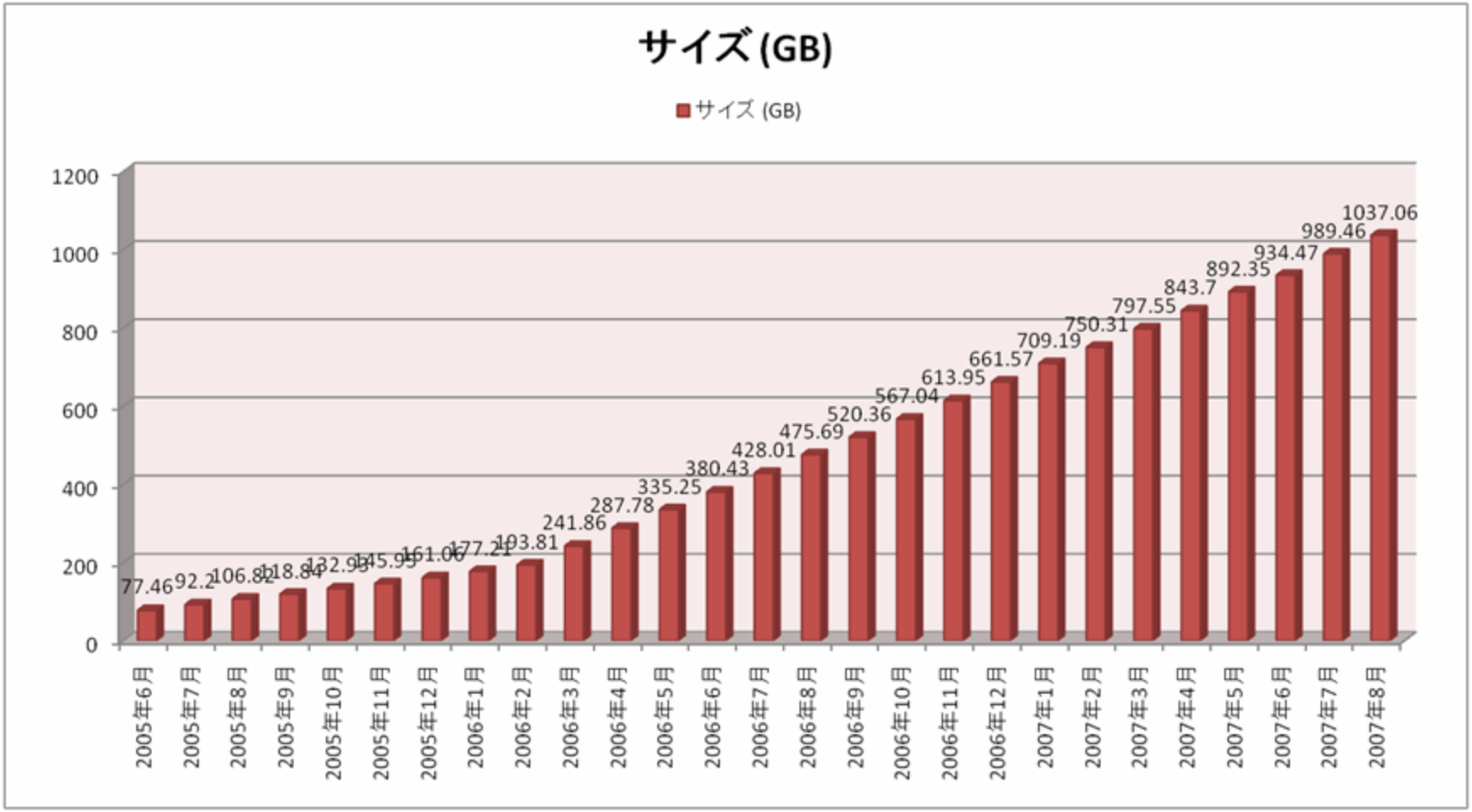
- Weekly ensemble
- Monthly ensemble
- Seasonal ensemble



GPV/JMA Data Archive
Number of Registered Users

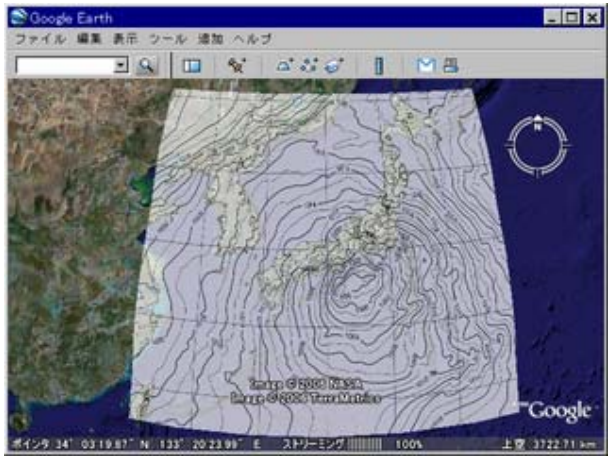


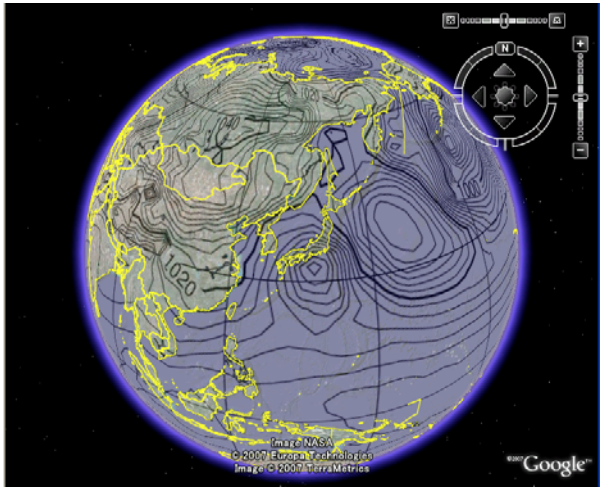
GPV/JMA Data Archive
Archived Data Size



Weather Maps on GoogleEarth

- GoogleEarth
  - Virtual globe program by Google
  - Can show several kinds of images overlaid on the surface of the earth
- KML (Keyhole Markup Language)
  - Tag-based file format used to display geographic data in an earth browser





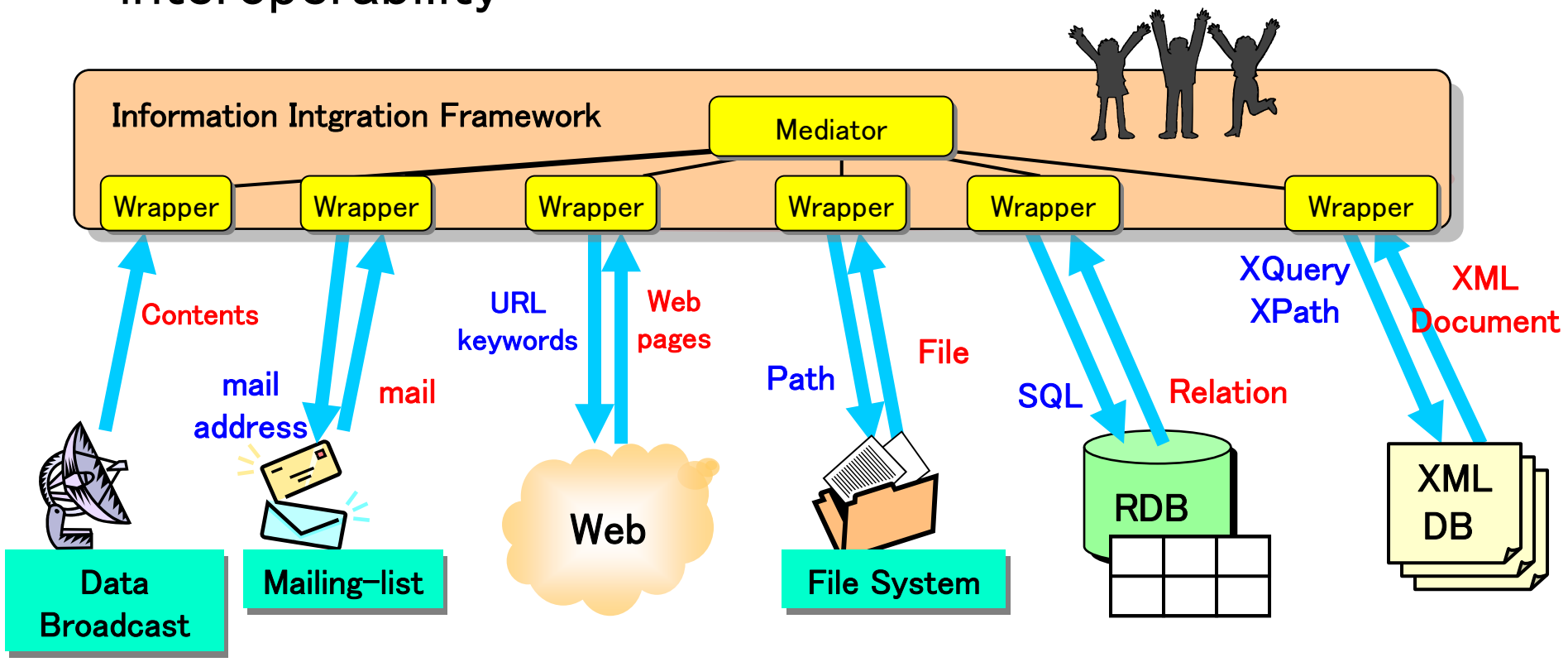
T. Amagasa, H. Kitagawa and T. Komano, “Constructiong a Web Service System for Large-scale Meteorological Grid Data,” IEEE Conf. on e-Science and Grid Computing, Dec. 2007.

Research Areas

- ✓ Meteorological Databases
  - GPV/JMA data archive
- ✓ Information Integration
  - Integrated use of different types of information sources: Databases, Web, Files, Sensors, …
- ■ Data Mining and Knowledge Discovery
  - Extraction of useful information from databases and web
- ■ XML
  - XML: Standard format for data interoperability
  - XML Data management and databases

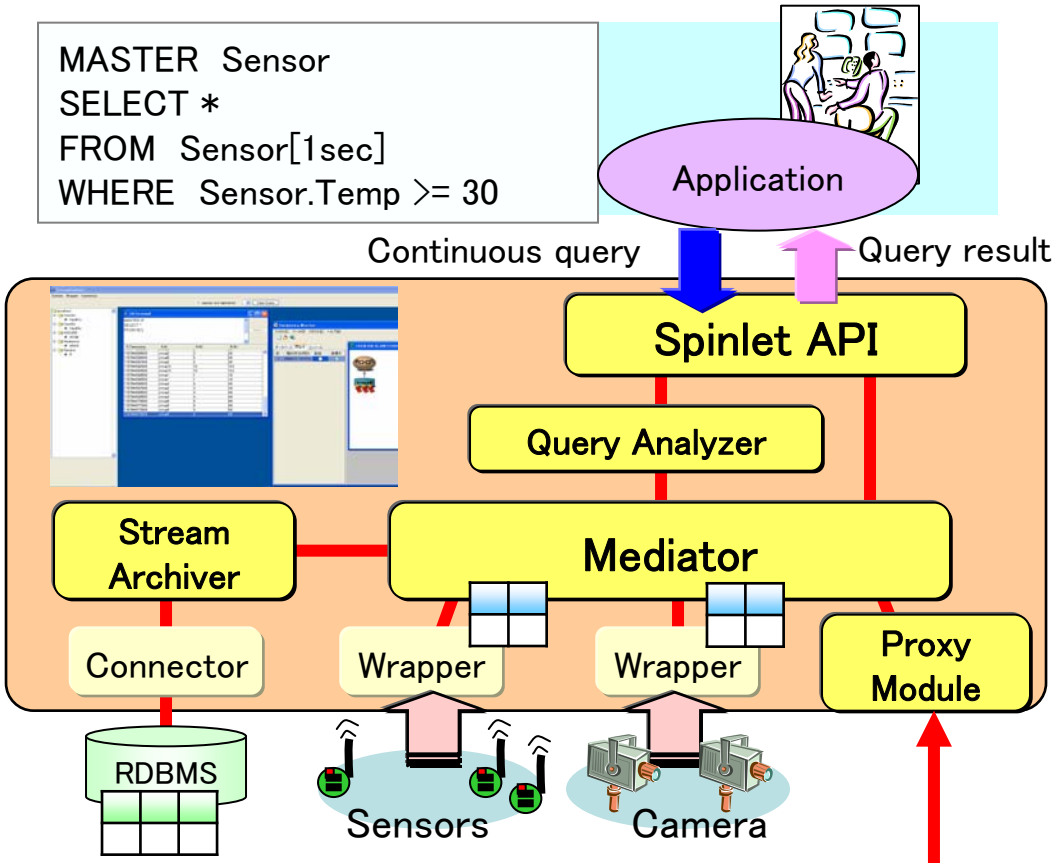
Information Integration

- A huge number of online information sources
  - Different data formats, access methods, query languages, …
- Information integration framework for data interoperability



Information Integration

- StreamSpinner
  - Integration framework for heterogeneous information sources
  - Can cope with streaming data sources such as sensors, location data, streaming media, etc.
  - SQL-like continuous query language
  - Distributed stream processing and data integration



Example 1: Simple Filtering

Notify when the temperature value (ttxd11) exceeds 25℃

MASTER Sensor

SELECT timestamp, fsr
FROM Sensor[1]
WHERE Sensor.ttxd11 > 25

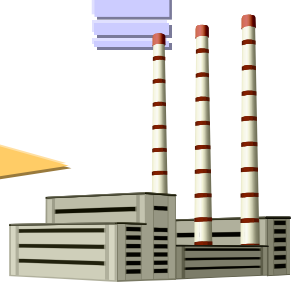
(ttxd11 expresses a value of the temperature sensor)

timestamp	ttxd11	ttxd12	tnhrpm	fsr
9867664…	23.3	23.9	1234	87.2
9867664…	23.4	24.1	1234	87.3
9867664…	23.6	24.1	1235	87.5
9867664…	24.1	24.4	1244	87.8

Turbine stream

Temperature
RPM
…

Sensors monitor behavior of a plant



Sensor streams

Example 2: Storing Stream Data into Database

```
MASTER  Sensor
INSERT INTO SensorDB VALUES (
    SELECT *
    FROM  Sensor [now]
)
```

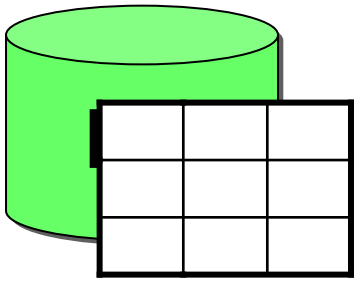
System can meet various insertion requirements

- Store data directly
- Store data after filtering
- Store the average value

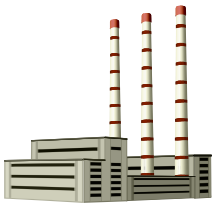
Store sensor data into a database when a new data item arrives

Inserion request

StreamSpinner

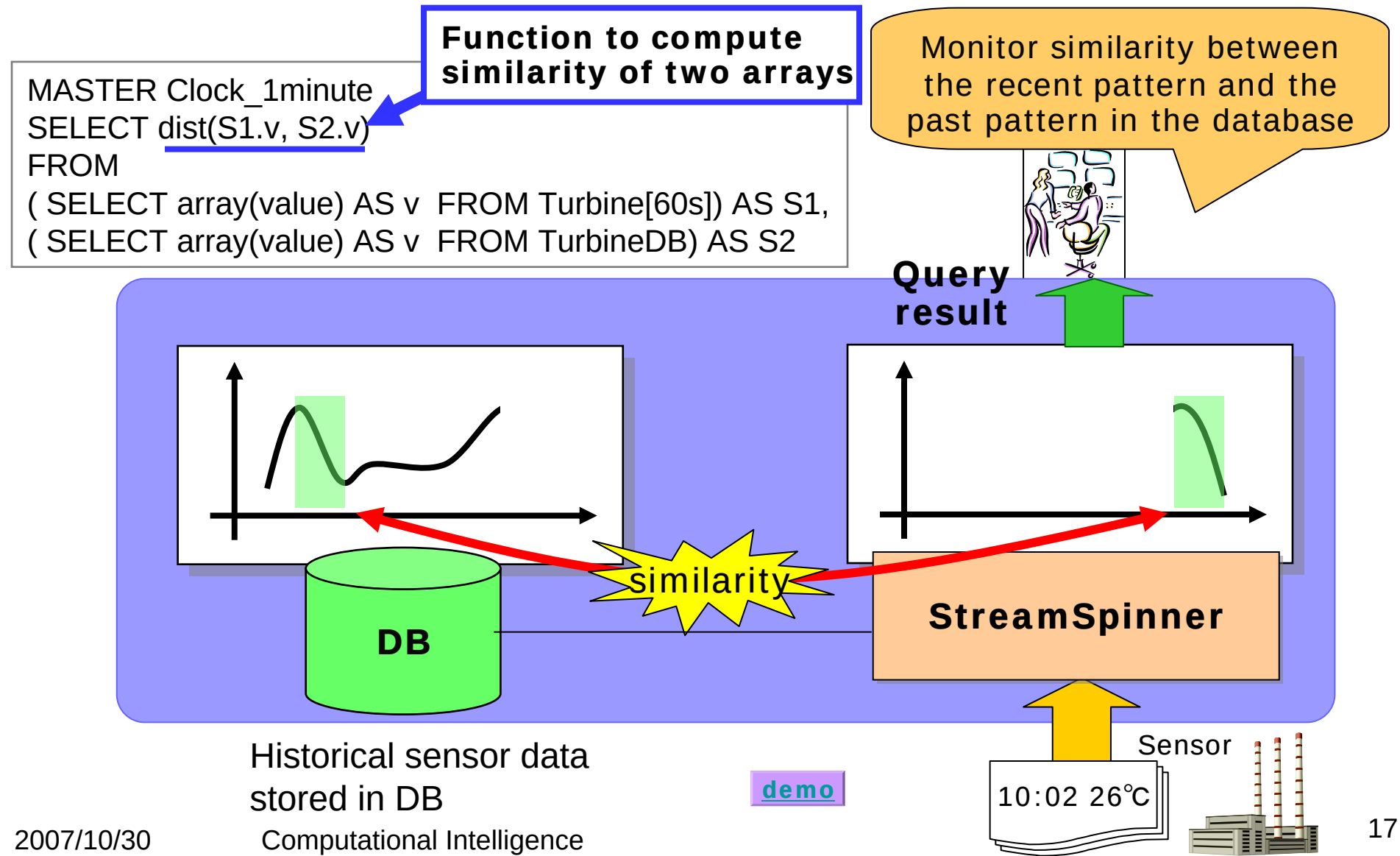


SensorDB



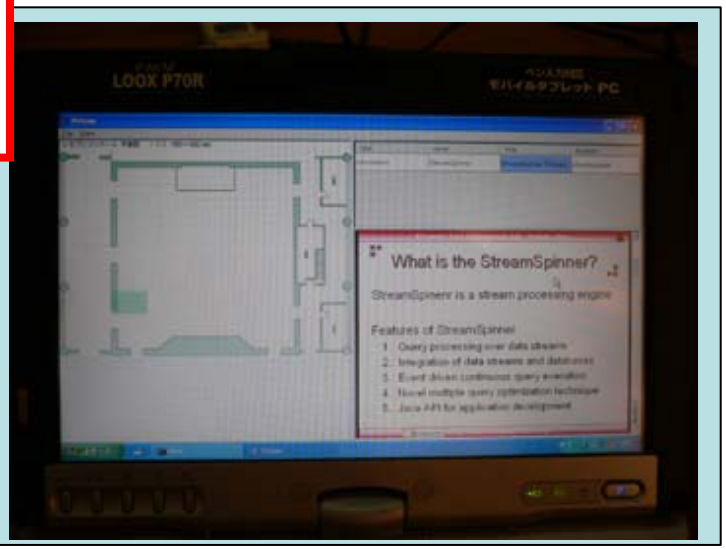
Sensor streams

Example 3: Integration of Streams and Database

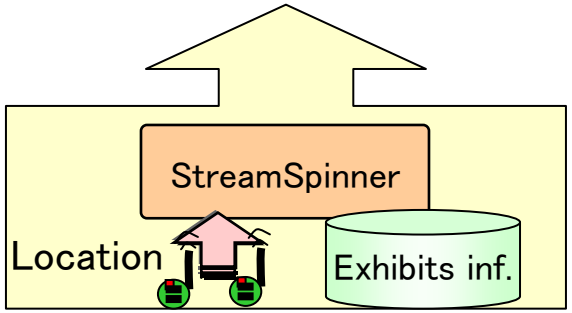


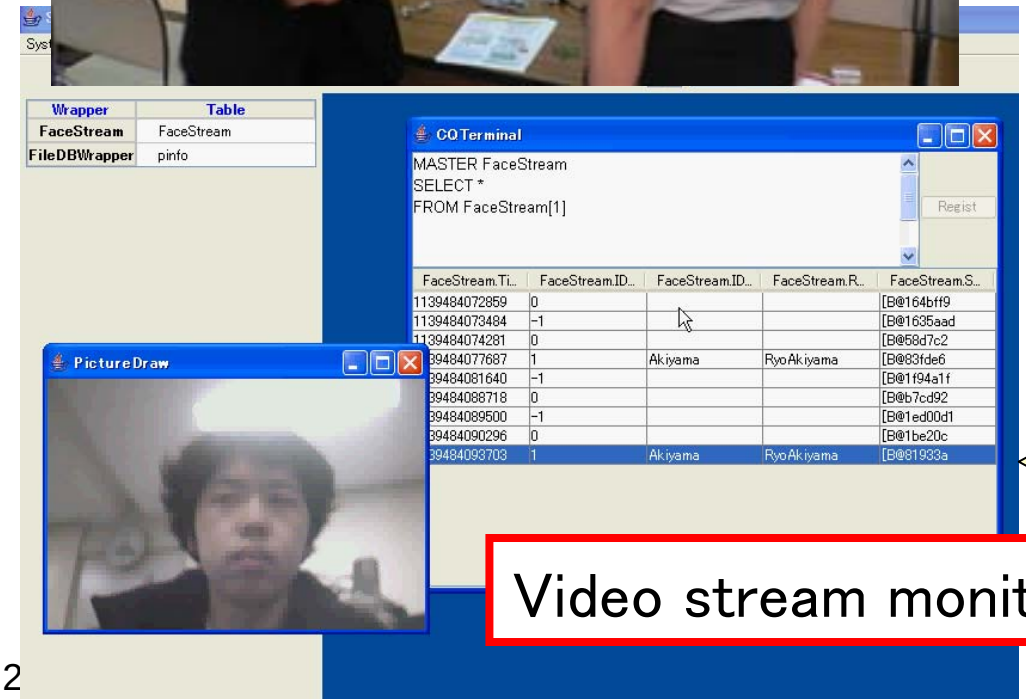
Integration of Streams and Database

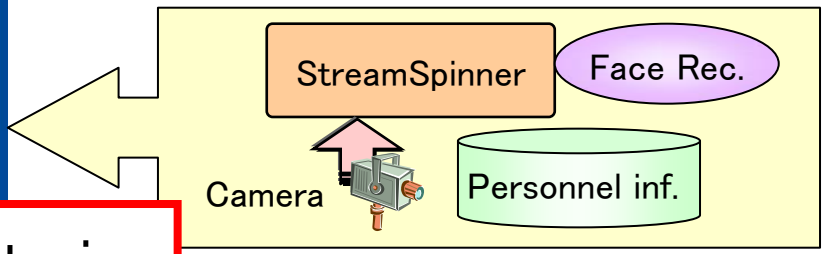
Location-based information delivery



demo







Video stream monitoring

Research Areas

- ✓ Meteorological Databases
  - GPV/JMA data archive
- ✓ Information Integration
  - Integrated use of different types of information sources: Databases, Web, Files, Sensors, …
- ✓ Data Mining and Knowledge Discovery
  - Extraction of useful information from databases and web
- ■ XML
  - XML: Standard format for data interoperability
  - XML Data management and databases

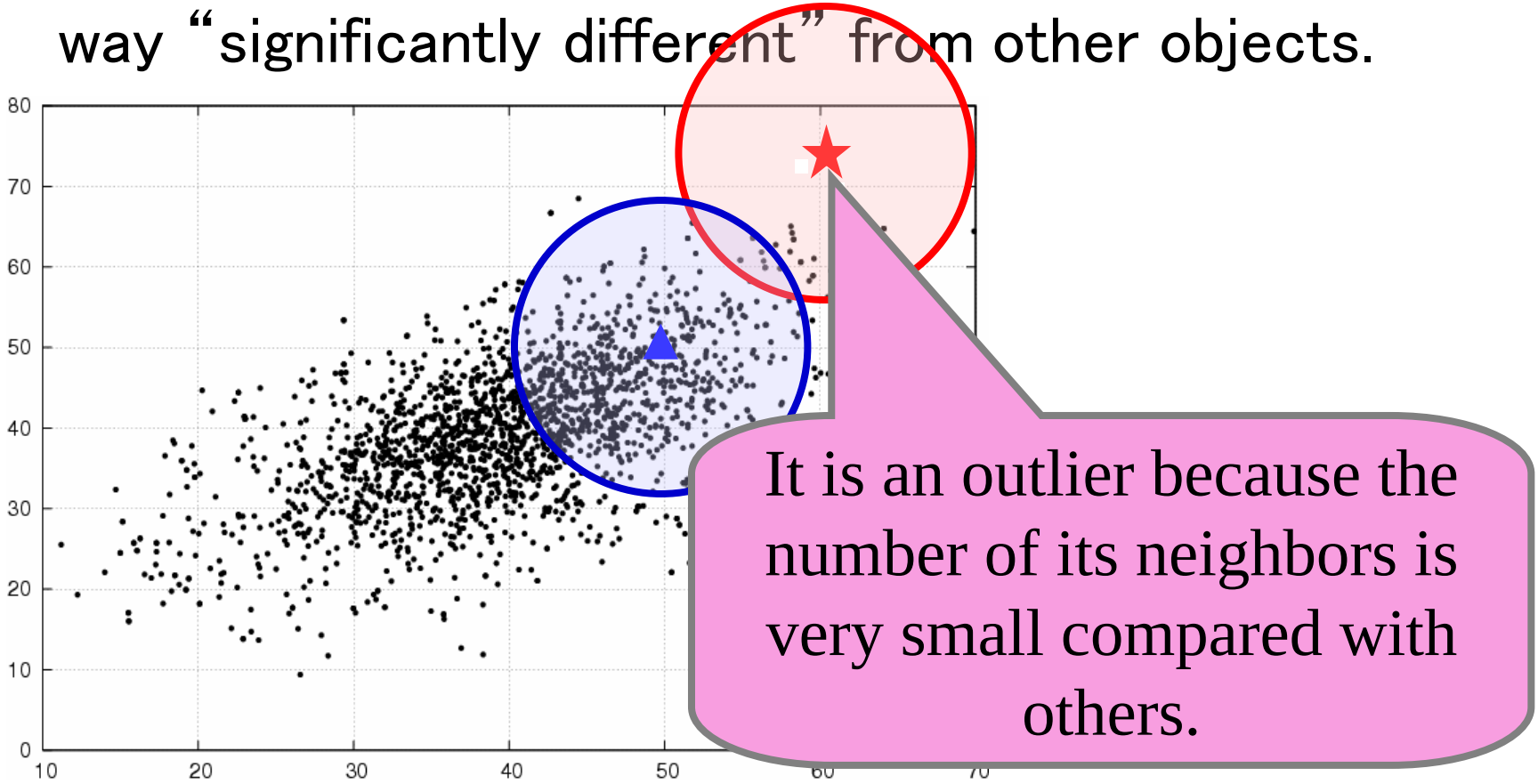
Data Mining and Knowledge Discovery

- Outlier Detection
- Ratio Rule Mining
- Information Extraction from Document Databases
- Novelty-based Document Clustering
- Topic Detection from Documents

…

Outlier Detection

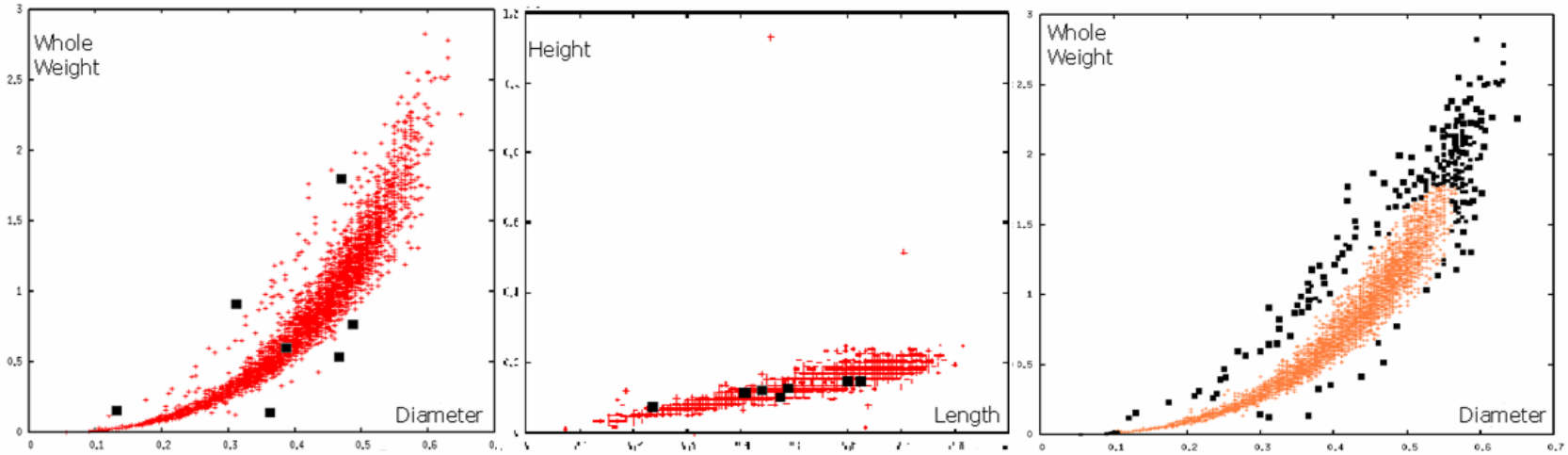
- Detecting outliers is an important problem with many applications such as anomaly and other interesting event detection.
- Intuitively, an object is an “outlier” if it is in some way “significantly different” from other objects.



Example-based Outlier Detection for High Dimensional Data

- Focusing on low dimensional projections to detect outliers in high dimensional datasets.
- Based on user supplied outlier examples.

Data : Abalone (UCI Machine Learning Repository) 4177 items



Outlier Examples in Subspace (A) | Outlier Examples in Subspace (B) | Detected Outliers in The Optimal Subspace

Outlier Detection for Categorical Records

Detects anomaly records in which many attribute values are not observed despite they should occur in association with other attribute values.

Animal Data

Each record presents a habit of an individual animal

ID	Egg	Legs#	Aquatic	…
1	Yes	4	Yes	…
2	Yes	4	Yes	…
3	Yes	4	Yes	…
4	Yes	4	Yes	…
5	Yes	0	No	…

Our method shows enough detection accuracies in an accuracies evaluation compared with the recent related work [KDD07, Das] for network intrusion data.

Association rules with high confidence

{ (Egg, Yes) }⟶{ (Legs#, 4) }
{ (Egg, Yes) }⟶{ (Aquatic, Yes) }
{ (Egg, Yes) }⟶{(Legs#, 4), (Aquatic, Yes) }

Support ≧ 40%, Confidence ≧ 75%

Rule's right hand itemset must have a strong association with the left hand itemset.

Outlier degree of record *t*

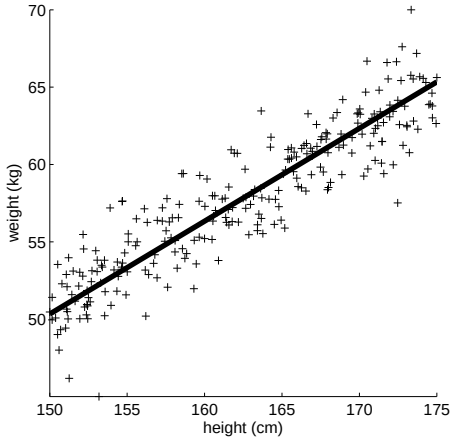
Consider an ideal form t^+ of t including all items which should observed in t.

t_5^+ = { (Egg, Yes), (Leg#, 0), (Aquatic, No), (Legs#, 4), (Aquatic, Yes) }

$$od(t) = \frac{|t^+ - t|}{|t^+|} = \frac{2}{5} = 0.4$$

Ratio Rule Mining

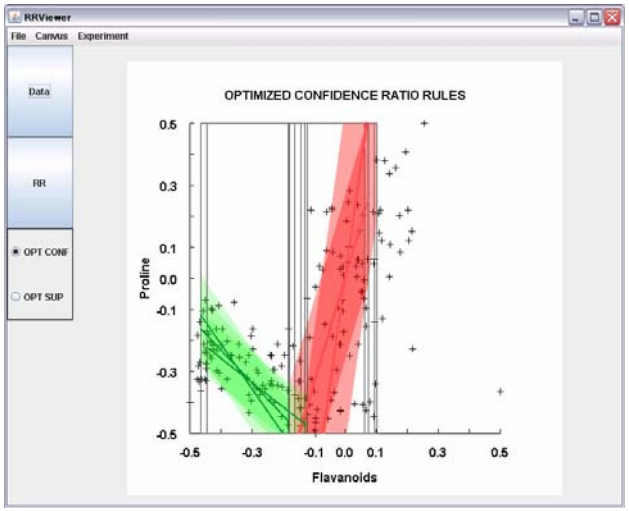
(weight) = 0.6 × (height) - 40

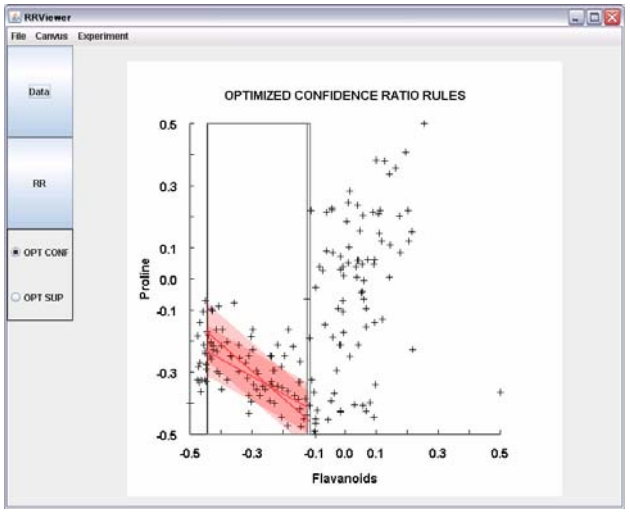


- Extract Ratio Rules (linear relationships) in numeric data
  - Capture linear relationships in the data
- Features
  - Can extract local linear relationships as well as global ones
  - Introduce Support (population size relevant to the rule) and Confidence (confidence degree of the rule) to characterize each ratio rule

Ratio Rule Mining

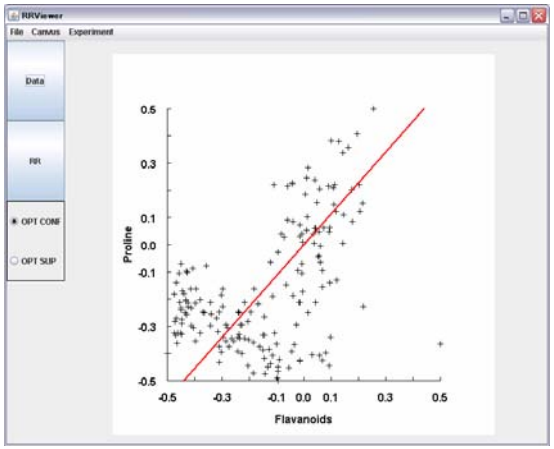
Example: Wine Data





Our Ratio Rules extract both two linear relationships separately

Only strong correlation can be extracted by setting appropriate parameters

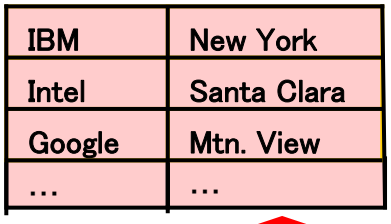


Linear regression: only the global linear relationship

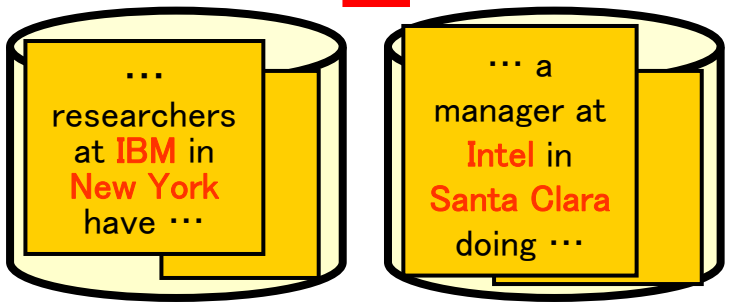
Record Extraction from Document Databases

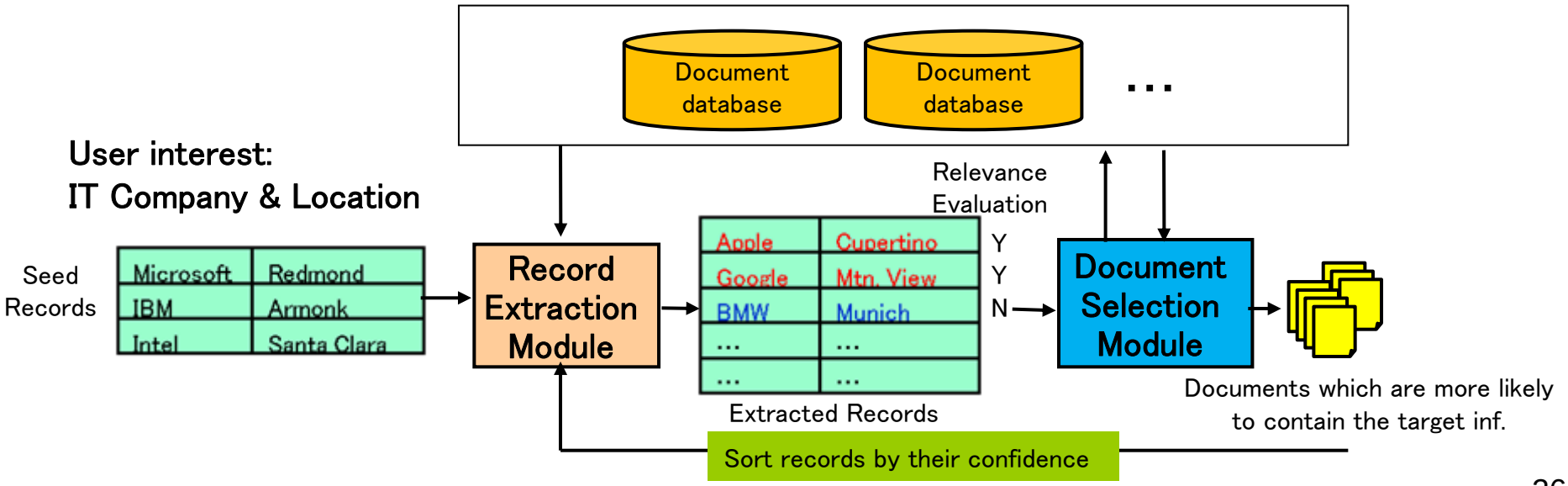
Problem: Extract records fit for user interests with high accuracy and efficiency

Basic idea: Narrow the search to documents which will contain target information



Record extraction from documents





Research Areas

- ✓ Meteorological Databases
  - GPV/JMA data archive
- ✓ Information Integration
  - Integrated use of different types of information sources: Databases, Web, Files, Sensors, …
- ✓ Data Mining and Knowledge Discovery
  - Extraction of useful information from databases and web
- ✓ XML
  - XML: Standard format for data interoperability
  - XML Data management and databases

Processing XML data in PC cluster systems

Background

- XML data processing is not cheap

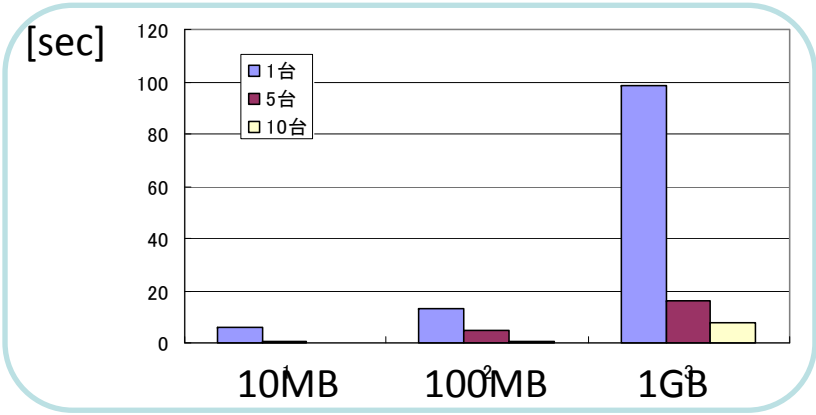
➔**Parallel XML processing to cope with growing XML data volume**

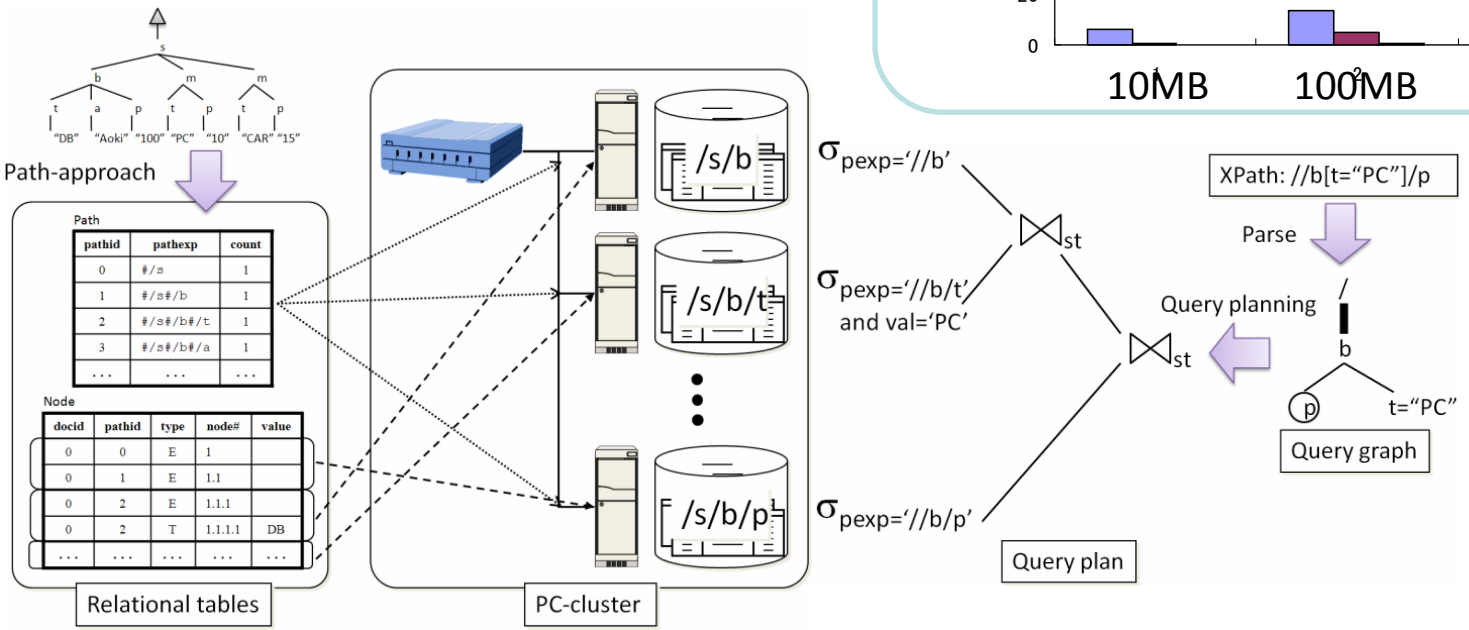
Approach

- Path-based partitioning of XML data
  - Cost estimation for XPath queries
  - Using GA to compute optimal partitions and allocation based on workload information

Results

- Nearly optimal partitioning and allocation of XML fragments
- Good scalability for XML data size





Storage and retrieval of XML in P2P

Motivation

- Overlay network (P2P)
  Infrastructure for sharing information among distant / different organizations

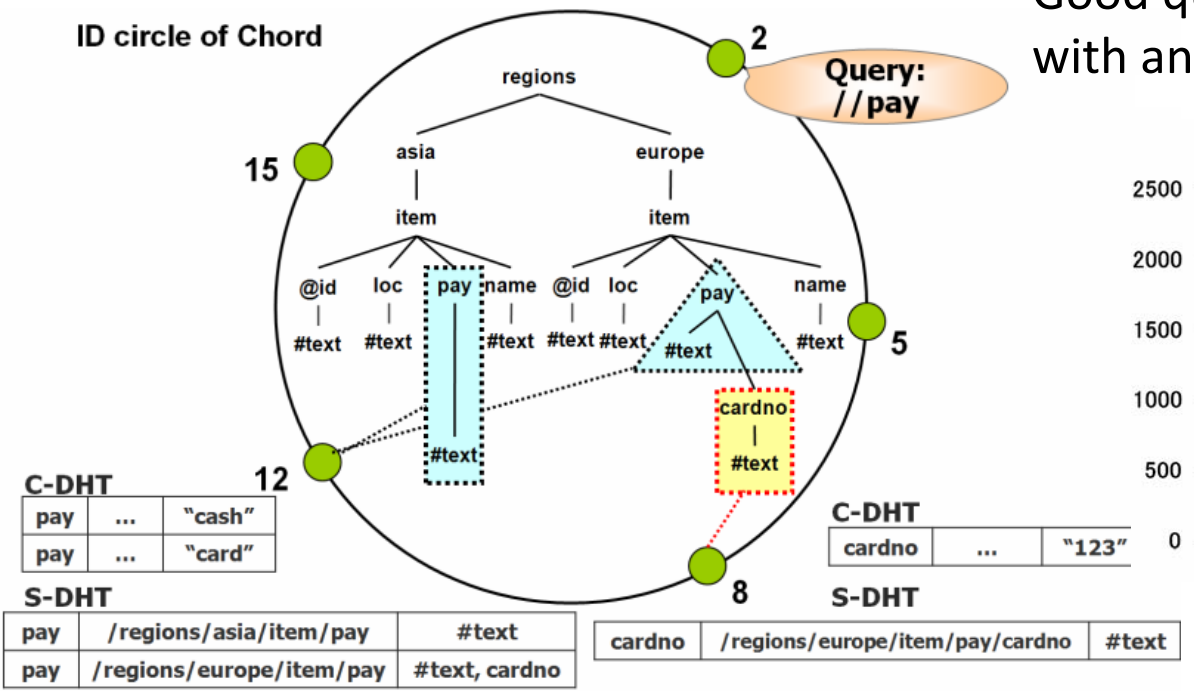
➔ **Efficient XML query processing over P2P network**

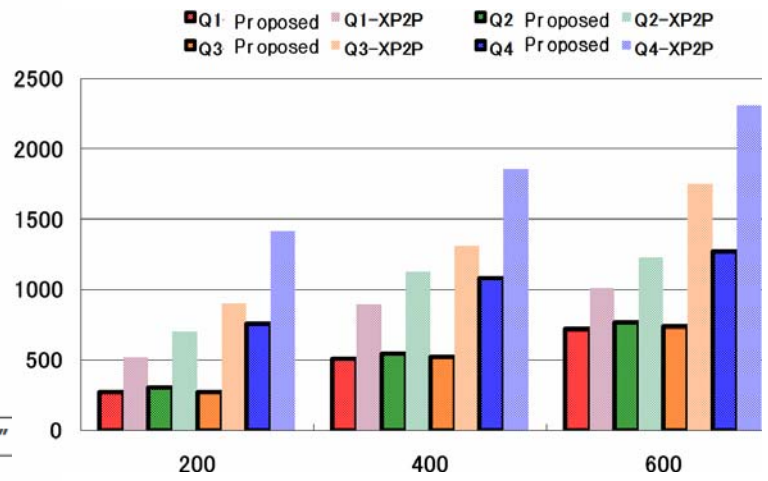
Approach

- Based on DHT
- Separate DHTs for XML contenxts and structures
  Contents (C-DHT)
  Structure (S-DHT)

Results

- Prototype system implementation
- Good query performance compared with an existing method





Analytical processing of XML (XML OLAP)

Motivation

- In-depth analysis on large volume of XML data
  - Decision making
  - Detailed analysis of scientific data

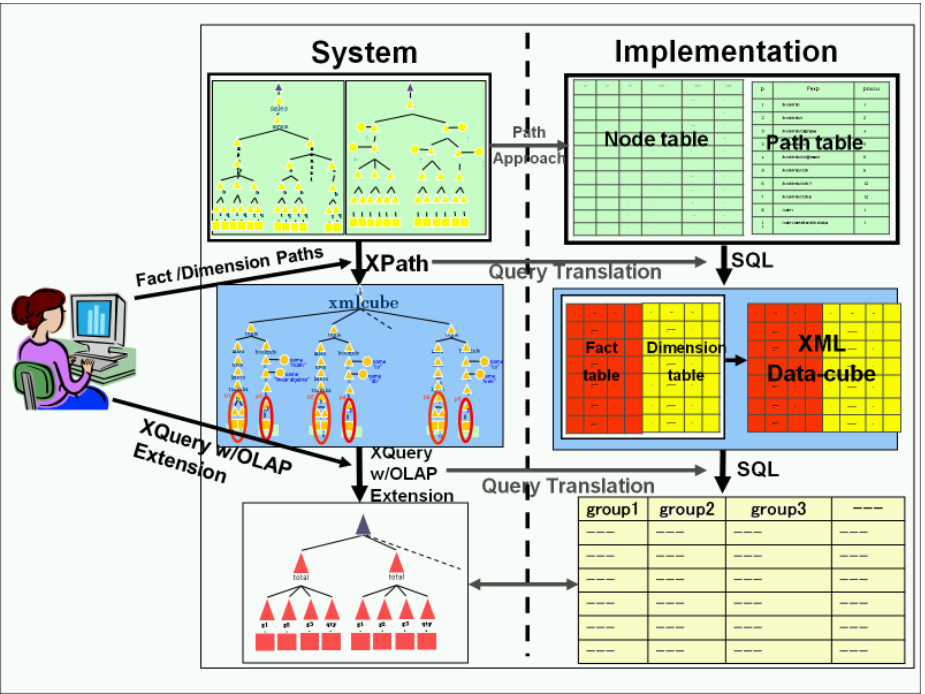
➔**Interactive analysis of XML data (XML OLAP)**

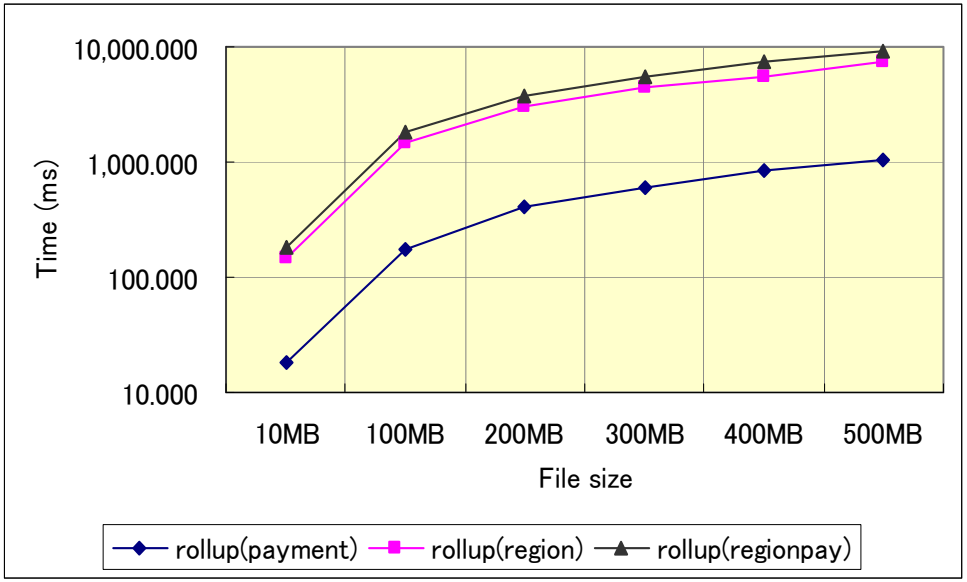
Approach

- Definition of facts/dimensions in XML
- Definition of XML datacube
- Mapping XML datacube to RDBMS
  Making use of relational storage

Results

- Prototype system implementation
- Performance evaluation





Funds

- Grant-in-Aid for Scientific Research from Ministry of Education, Culture, Sports, Science and Technology (~$0.6 million; past 3 years)
  - Grant-in-Aid for Scientific Research A
  - Grant-in-Aid for Scientific Research on Priority Areas (Infoplosion Project)
  - Grant-in-Aid for Exploratory Research
  - Grant-in-Aid for Young Scientists
- JST CREST Project (in collaboration with OS group) (~$0.3 million; past 3 years)
- From industry

Publication and Awards

- Refereed Papers
  - 2004: 17 (Journal 10, Conference 7)
  - 2005: 14 (Journal 7, Conference 7)
  - 2006: 18 (Journal 7, Conference 11)
  - 2007: 20 (Journal 9, Conference 11) (As of August 2007)
- Awards
  - 2 Best Paper Awards (DBSJ, IEICE)
  - Young Researchers' Award (DBSJ)
  - 2 Achievement Awards (IPSJ Fellow, IEICE Fellow)
  - 15 Students' Awards

Future Plan

- Research and development of technologies for data engineering infrastructure
- Scientific databases
  - Collaboration with Global Environmental Science Group
  - Other datasets, data mining, …
- Reinforcement of collaboration with other groups and divisions to tackle new research issues in data engineering

Thank you.