Very Large Information Integration and Application Platform

CSTP Coordination Program of Science and Technology Projects

Project Director: Shojiro Nishio, Ph. D. (Osaka University, Japan)
Agenda

1. Background and Purpose
2. Very Large Information Integration and Application Platform
3. Four Major Projects of Ministries (METI, MIC, MEXT)
4. Collaboration among Four Projects
5. Conclusions and Future Work

METI: Ministry of Economy, Trade and Industry
MIC: Ministry of Internal Affairs and Communications
MEXT: Ministry of Education, Culture, Sports, Science and Technology
1. Background and Purpose
Background-1: Information Explosion Era

Time spent in searching: 30%

Time spent in intellectual activities (high-value-added industry worker)

161 billion GB (2006)
988 billion GB (2010)

IDC, “Expanding Digital Universe” (Mar., 2007)

- Growth of Web
- Rich Information
- Emerging sensor technologies
Background-2: Problems with Information Retrieval Engines

- Search engines do not analyze and present information from user’s perspective
- Garbage and harmful information may be presented at top of search results

Economy

Education

Local Community

Society

If popular search engines show wrong information at top of search results...

I’m not sure, but I will search for something and use the top search results...
Purpose: New Intelligent Information Retrieval Platform

- Achieving a safe society by using an intelligent platform
2. Very Large Information Integration and Application Platform

2.1 Promotion Scheme
2.2 Overview
2.3 Positioning
2.4 Goal
2.5 Schedule
2.1 Promotion Scheme -1:
Coordination Program of Science and Technology Projects

**Coordination Program of Science and Technology Projects**

- Life Science basic platform
- Emerging and Reemerging Infectious Diseases
- Ubiquitous Network
- Next-Generation Robot
- Biomass Utilization
- Nanobiotechnology
- Hydrogen Use / Fuel Cells
- Research and Development for Counter-terrorism
- Regional Science and Technology Cluster
- Clinical Research and related Translational Research
- Food and Biological Production Research
- Very Large Information Integration and Application Platform
- Research and Development on the safe use of Chemical Substances based on Comprehensive Risk Evaluation
- Platform Development for the Promotion of Nanotechnology Research and Development and its Social Acceptance

**General PTs (Project Teams) for Promoting Sectoral Strategy**

- Life Science PT
- Information and Communication PT
- Environment PT
- Nanotechnology and Materials PT
- Energy PT
- Societal Platform PT
- Regional Science and Technology Cluster PT

**Investigation Committee for Promoting Basic Policy**

**Cabinet Office**

**Minister in Charge of Science and Technology Policy**

**Director-General for Policy Planning**

14 programs
2.1 Promotion Scheme -2: What is Coordination Program of Science and Technology Projects?

A framework mandated to

- Maximize results from projects of governmental organizations
- Strengthen projects by assigning coordinators to lead them
- Make needed adjustments, e.g. eliminating project overlaps
- Utilizing Grants in Aid of Special Coordination Funds for Promoting Science and Technology.
2.2 Overview: New Coordination Program in 2007

Very Large Information Integration and Application Platform

Overview

Develop platform technology for intelligent, next-generation information use

- Safe, appropriate, and secure information systems
- Collection, analysis, and management of various information
- Providing unique information services

Four Targeted Projects

- MEXT: R&D of Ultra-High Performance Database Engine
- METI: Information Grand Voyage Project
- MIC: R&D of Information Credibility Evaluation Technology
- Complementary Project: Sensing Web: Advanced Use of Sensor Information
2.3 Positioning

Cyberspace

Amount of information has explosively increased in Ubiquitous Net society.

Very Large Information Integration and Application Platform

Information Grand Voyage Project
- Classify largely integrated information to be reusable.
- Organize and store classification results.

Sensing Web
- Advantage Use of Sensor Information

Information Analysis

Ultra-High Performance Database Engine
- Improve processing efficiency to a remarkable extent.

Information Credibility Evaluation Technology
- Necessary refinement of the integrity of stored information
- Offering of guaranteed credibility to user

Various Applications
- Business
- Life
- Education
- Healthcare, etc.
2.4 Goal:

Fiscal 2011 Very Large Information Integration and Application Platform

P1: Information Grand Voyage
Next generation information analysis services
- API for home appliance information provision services
- API for car navigation information provision services
- API for music content services
- API for xxx service

P2: Information Credibility Evaluation Technology
API for trustworthy and credibility analysis
- Trustworthiness and credibility analysis services

P3: Ultra-high performance database
Search, analysis, and classification of content
- DB for analyzed and classified content

P4: Sensing Web
Search, analysis, and classification of content

Existing search services
- API for xxx search engine
- API for Web search engine

API for handheld search engine

Ubiquitous Net Society
- Different information from all over the world

Information “melting pot”
- Huge volume of disorganized multimedia content on the Internet

Transmission of specific information
- Diverse sensor and graphic information
- Various databases

Huge volume of disorganized multimedia content on the Internet
2.5 Schedule

Information Grand Voyage Project
- Next-generation information retrieval and analysis
  - Start 2007
  - 3 years

Information Credibility Evaluation Technology
- Evaluation of information credibility and trustworthiness
  - 4 years

Ultra-High Performance Database Engine
- Improve processing efficiency to a remarkable extent
  - 5 years

Sensing Web: Advanced Use of Sensor Information
- Advanced use of sensor information
  - 3 years

Further results obtained by collaboration among projects
3. Four Major Projects of Ministries (METI, MIC, MEXT)

3.1 Overview
3.2 Information Grand Voyage Project
3.3 Information Credibility Evaluation Technology
3.4 Ultra-High Performance Database Engine
3.5 Sensing Web: Advanced Use of Sensor Information
Very Large Information Integration and Application Platform

3.1 Overview

- **Information Grand Voyage Project** (METI: 3 years)
  - Protocol and platform technology needed to achieve service collaboration
  - Real-time processing technology of multimedia information
  - Technology for large-scale Information collection and resource management
  - Integrated processing technology for environmental and behavioral information
  - Safe management technology for privacy information

- **Information Credibility Evaluation Technology** (MIC: 4 years)

- **Ultra-High Performance Database Engine** (MEXT: 5 years)

- **Sensing Web** (complementary project: 3 years)
  - Started in FY2007
  - Ubiquitous Network (Information use in the real world)
  - Internet (Information use in cyberspace)

- **Information Analysis and Management**

Copyrights Reserved 2008, ©CSTP Coordination Program of Science and Technology Projects
3.2 Information Grand Voyage Project

METI: Ministry of Economy, Trade and Industry
http://www.igvpj.jp/index_en/
3.2 Information Grand Voyage Project (METI)

**Background**

- **Growth of Web**
- **Rich Information**
- **Emerging sensor technologies**

**Objectives: Info-plosion driven aspirations**

Applying “Information Explosion” to achieve a richer social environment:

- Build the infrastructure for next-generation information retrieval / analysis technology to help users gather and analyze a huge quantity and variety of information on the Web and elsewhere with ease, convenience, and success.

- By fiscal 2011, make it possible to strategically exploit mega data, central to the era of information explosion.
Information Fusion by Information Grand Voyage Project

Web Information
- Corporate & Individual Web sites
- Images, Videos and Music
- Blog, SNS
- A variety of DBs

Non-Web Information
- Spatial information
- Purchase history information
- Audio-visual Information
- Healthcare Information

Greater innovation and Richer info-based society

Next-generation information retrieval / analysis technology
Commitment of the Information Grand Voyage Project

Triune Approach

Demonstration with model services *
- Verify effectiveness and feasibility of the next-generation information retrieval / analysis technology using 10 model services

Establish mechanism for independent innovation through practical use of information

Industrial implementation and strengthening competitive power

Establish and revise regulations and environment
- Establish and revise regulations on privacy and copyright protection.
- Create mechanism for smooth distribution of intellectual property
- Prepare an environment for development and demonstration

Technology Development *
- Develop next-generation information retrieval / analysis technology
- Achieve versatile and common retrieval / analysis technology (provide a common technology)

* More detail description next slide
## Model Services in FY2007

### A  Future-oriented personal services considering privacy

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A-1</td>
<td><strong>My life assist service</strong> (Behavior-based chain search using cell phone)</td>
</tr>
</tbody>
</table>
| A-2 | **Cooperative service using contactless IC card for traffic**  
               (Sharing sensor information with PASMO) |
| A-3 | **Profile passport** (Application of behavior history using game machines) |

### B  Next-generation Web services giving new values

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B-1</td>
<td><strong>Laddering search service</strong> (Interactive search)</td>
</tr>
<tr>
<td>B-2</td>
<td><strong>Sagool TV</strong> (Internet movie search)</td>
</tr>
<tr>
<td>B-3</td>
<td><strong>View search Hokkaido</strong> (Next-generation image search)</td>
</tr>
</tbody>
</table>

### C  IT services based on new social infrastructure

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C-1</td>
<td><strong>Time-space information mining service</strong> (Real-time information retrieval)</td>
</tr>
<tr>
<td>C-2</td>
<td><strong>Health Monitoring and assistance Service</strong> (Medical sensor)</td>
</tr>
<tr>
<td>C-3</td>
<td><strong>New integrated safe flight support system</strong> (Collection and analysis of safety information)</td>
</tr>
<tr>
<td>C-4</td>
<td><strong>Mega research</strong> (Interview-type questionnaire)</td>
</tr>
</tbody>
</table>
Example of Model Service: A-1 My Life Assist Service (1)

Cell phone assists you in real life!

- **My Life Services**
  - Recommendation service
  - Behavioral analysis service
  - Living info. storage service
    - info. of behavioral activities

- **Service providers**
  - Contents/Services providers
  - Outing assist service providers
  - Mining providers

1. NW access status
   - Time, attributes, taste and targets
   - PC, cell phone and information appliances
   - Households

2. # of passengers and attributes
   - People's flow and traffic line trend, In-car NW access trend
   - Railways & buses
   - # of passengers and attributes
   - People's flow and traffic line trend, In-car NW access trend
   - FeliCa and in-car NW

3. People's flow and traffic line to each exit in station
   - Flow line trend, # of migrants and attributes NW access in station
   - Stations
   - Indoor location Wireless LAN
   - Analyze people’s flow via camera (each exit)

4. Behavioral feature in department store
   - Relationship between attributes and consumption
   - Response to sales
   - Department stores
   - RFID tag
   - POS
   - FeliCa
   - POS

5. Relationship between pre NW access and actual visit
   - Behavioral feature on the way
   - Restaurants
   - FeliCa
   - POS

6. Statistic and real time consumer attitude feature of in-store POS data
   - POS
   - Stores

7. Biological data such as heart beat rate data and walking distance data
   - GPS and camera
   - Walking

Very Large Information Integration and Application Platform

September, 2008

Copyrights Reserved 2008, ©CSTP Coordination Program of Science and Technology Projects
Example of Model Service: A-1 My Life Assist Service (2)

Preview channel

You can get contents related to your everyday behavior and your potential desires (It starts by opening the folding cell phone)

One day, I got off work on time and opened my cell phone on the way home ...

Well, I have worked overtime and came home late for several days ...

Surprising proposal that hasn't appeared in his recent activity and matches his preference

Propose a new activity by predicting her next behavior from her past data

Click!

List of restaurants in Yokohama
For relaxation
For stress release
For healing ...
Linked to information

I did a good job this week. I’ll give myself a reward!

Tokyo station now
Hallo! Why don’t you come by XX?
Technology Development: Preparation of Collaboration Platform

**Definition**

1. Place to share achievements of Information Grand Voyage Project
2. Platform to accelerate R&D of the next-generation information-access technology and enhance competitive edge

**Collaboration Platform (place)**

- **IGV intellectual property bank**
- **Rich content**
- **Common technology library group**
  - Common technology to services
  - Common technology to infrastructures

**Community**

- **Stakeholder participants**
- **Information Great Voyage Project members**
- **Information Great Voyage Consortium members**
3.3 Information Credibility Evaluation Technology

MIC: Ministry of Internal Affairs and Communications

3.3 R&D of Information Credibility Evaluation Technology (MIC)

**Current Problems**

- Judging credibility of information is left to users based on experience, intuition, etc.
- In the large volume of information flowing through the Internet, valuable info is mixed into unknown, harmful, and incorrect info: obtaining needed info is difficult!
- Phishing and various disinformation scams cause real damage!

**Project Goal**

To create an environment where anyone can manage and use text, audio, and graphic information on the Internet with ease and security by establishing technologies to detect false and deceptive information and thus supply credible information.

Understand the meaning of information over the Web and analyze its relationship with other information

Inquiry and verification of credibility

- Correct information
- Harmful
- Disinformation

(1) Provide only credible information
(2) Display level of credibility

This information is 95% disinformation or malicious. Display this?
Overview

Research and Development on Evaluation Technologies for Information Credibility

- Develop technologies to analyze information on the web and to predict the probability of its credibility.
- Offer analysis results so users can evaluate information as “good” or “bad” based on the user’s own judgment criteria.

Digital information in pile of diamonds and dirt

Text, image, audio, and video information

Knowledge Base

Deep analysis using natural language processing including analysis of content and meaning of text information

Analysis of text, image, audio, and video information using approaches in data analysis

Issue A: Technology for Web content analysis

Offer information that serves as a guide for the user in judging credibility of information

Issue B: Temporal analysis technology for the content and meaning

Derivation

Surface feature

Style

Lexicon

Temporal transition analysis

Opinion A

Opposition

Opinion B

Opinion C

Opposition
Issue A: Technology for Web content analysis

Develop technologies to analyze the credibility of large volumes of web-based text, graphic, audio, and video information using methods in data analysis.

Technology to analyze graphic, audio, and video information

- Develop a cross-media information analysis engine to verify inconsistencies in graphic, audio, and video information, and their explanatory text information.

Technology to analyze text information

- Develop a surface-feature analysis engine to analyze the quality of text from the surface features of web text such as blogs; such features include typographical errors, omitted letters, lexical diversity, and sentence readability.
- Develop a sender analysis engine to analyze the features of the topics most frequently raised by the sender based on sending history, etc.
**Issue B-1: Temporal analysis technology for meaning and content (analysis of content and social evaluation of information)**

Natural Language Understanding and Evaluation technology to extract opinions latent in the collection of text on the Web and logical relationships between opinions

**Natural Language Understanding and Evaluation technology**

- **Develop the technology for building large-scale lexical knowledge** to use in capturing the logical relationships between opinions

- **Develop a logical relationship-recognition engine** to identify the logical relationship between any two opinions

- **Develop a opinion-analysis engine** to analyze and sort information related to fields and viewpoints of interest to the user is interested and to present logical relationships between opinions as well as their importance

- **Develop a summary-creation engine** to present a general view of the opinions in the fields of interest to the user
**Issue B-2: Temporal analysis technology for the content and meaning (analysis of content and social evaluation of information)**

Temporal analysis technology to analyze the temporal transition in opinions, their mutual relationships, and process used to transmit the information

Temporal analysis technology

- Establish a formula to handle opinion transition quantitatively, and establish a way to express the transitions
- Establish a temporal clustering method to output opinion transition and relationship type relevant to needed information. Develop information propagation phase-judgment engine to determine the phase of the clustered information in the information propagation process: Whether the opinion is established, has spread out rapidly to many persons, or is already out of date

Digital information in pile of diamonds and dirt

- Text, image, audio, and video information
- Knowledge Base
- Deep analysis using natural language processing that includes analysis of the content and meaning of the text information

Issue B: Temporal analysis technology for the content and meaning

Offer information that serves as a guide for the user in judging credibility of information
3.4 Ultra-High-Performance Database Engine

MEXT: Ministry of Education, Culture, Sports, Science and Technology

http://www.mext.go.jp/english/index.htm
3.4 R&D of Ultra-High-Performance Database Engine (MEXT)

**Background: Towards Database Engine for Info-plosion Era**

- Growth of amount of Data: so called ‘Info-plosion’ (Information Explosion)
- Digital shadow is becoming substantial large
- Traceability system introduced for various applications: PLM, GAP, Green IT, etc.
- Our target: huge, not mid-range database

**Change**

- Computing Environment drastically changing
- Increasing number of cores: New Moore’s Law
- Storage system revolutionized by SAN technology
- Database Engine has only slightly changed

**Project Goals**

- Creation of technology providing ultra-high performance advantage for large-scale database management systems:
  1. Establish innovative execution principle “Out of Order Database Execution Principle”
  2. Design and implement database platform software based this principle
  3. Verify its effectiveness through actual use
Performance Comparison: Existing and OoODE Database Engines

Drastic improvement in performance through Out of Order Execution

Existing database engine

New-type database engine (OoODE*)

Synchronous I/O issuance

Large-volume asynchronous I/O issuance

* OoODE: Out of Order Database Engine
OoODE (Out of Order Database Engine)

Development Component

- Application is constant
- Out of Order DBMS
  - Implementation of new function
- OS
- Storage
  - Improvement of existing function

Performance Prospects

Relative performance

- 10 times of performance improvement is planned
- 100 times of performance improvement is planned

Initial Experiment: Improvement of Performance through Large-Volume I/O Issuance Mechanism

Results of Initial Experiment

TPC-H benchmark Q3 MySQL trial environment (1CPU)

Execution time [sec.]

<table>
<thead>
<tr>
<th>No. of threads</th>
<th>Execution time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>800</td>
</tr>
<tr>
<td>2</td>
<td>400</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
</tr>
</tbody>
</table>

Example of execution tree in Out of Order Execution

- Aggregation function
- Selection (SHIPDATE)
- Selection (ORDERDATE)
- NL coupling (ORDERKEY)
- NL coupling (CUSTKEY)
- Secondary Index
- Primary Index
- CUSTOMER
- ORDERS
- LINEITEM
3.5 Sensing Web: Advanced Use of Sensor Information

Complementary Project: Kyoto University

http://www.kyoto-u.ac.jp/en

3.5 Sensing Web: Advanced Use of Sensor Information

Background

- Web has grown worldwide: now a huge knowledge base
- Sensor network, however, are not utilized widely:
  - Problem of protecting privacy in information: Main target of sensor networks has been non-personal, such as weather monitoring
  - Variety of sensors in use: Different procedures required for operation

Purpose

Creating “Sensing Web”

- Sensing Web offers a Web-like environment for freely using sensor information by applying privacy-protection technology (e.g. filtering)
- Real observed information can be used to create new multimedia services
Difference between Web Information and Sensor Information

- Web information requires human effort for information input
- Sensor Information uses direct observation for automatic input
Content Extraction of Sensor Information for Societal Use

Sensing Web

1. Technology for privacy information management (Access Management)
2. Technology for content extraction of sensor information (Information Sharing)
3. Observation real-world content-presentation technology (Utilization of Information)

- Content bus
  - Sensor search
  - Location
  - Sensor response
  - Access restriction
  - Data retrieval

- Presentation system
- Sensor node
- Clickable map
- Privacy structuring
- Content extraction system
## Framework of the project

### 1. Technology for privacy information management (Access Management)

**Responding to problems unique to sensor information**
- Privacy structuring and access management
- Extraction method of information from streaming data

### 2. Technology for content extraction of sensor information (Information Sharing)

**Mechanism for sharing sensor information**
- Description method of sensor information including time-space context
- Description method of requirements from user
- Search and detection methods of required sensor

### 3. Observation real-world content-presentation technology (Utilization of Information)

**Search and presentation of distributed sensor information**
- Collection method of information from largely distributed sensors
- Information processing from sparse observation environment over a large area
4. Collaboration among Four Projects
Collaboration among Four Projects

Information Credibility Evaluation Technology (MIC)  
Information Grand Voyage Project (METI)

Ultra-High Performance Database Engine (MEXT)  
Sensing Web: Advanced Use of Sensor Information (complementary subject)

Goals

- To build a next-generation information retrieval / analysis platform enabling users to collect and analyze needed information from diverse sources inside and outside the Web easily, effectively, and reliably.
- To develop technologies for strategically using super-large-volume data in this age of information explosion by fiscal 2011.
5. Conclusions and Future Work
5. Conclusions and Future Work: How do we advance this Coordination Program?

Moving the coordination program into the future

Pioneering new fields in an age of information explosion requires application of information.

- Contribute to an innovative environment by establishing a platform technology that can be applied over the Web using next-generation information retrieval / analysis technology focusing on media content.

- Increase the values of various projects by collaboration among them and adoption of existing technologies like Google to create new services and expand markets.

Discussions among specialists in information retrieval, database, and language processing.

- A “highly sophisticated” information retrieval / analysis technology must be established within three-year development period of Information Grand Voyage Project.

- Opportunities must be provided for information exchange such as symposia and discussions with experts.