A design of an information retrieval method based on TPO metadata

Information Technology Center (inet-lab)
Ismail Arai
Ismail-a@is.naist.jp

Summary

- □ Introduction
- □ Purpose
  - Requirements
  - Related works
  - Problems
- □ Task
  - Proposal
  - Design
- □ Evaluation
  - Mounting
  - Result
  - Future work
- □ Conclusion

Introduction

- □ Spreading the internet coverage
  - Getting some information by mobile phone, PDA, laptop PC…
  - Access to the internet in Hot Spot
- □ WWW (World Wide Web)
  - Increasing the everyday information
  - Stocking a huge amount of information
  - Full text search is main retrieval method

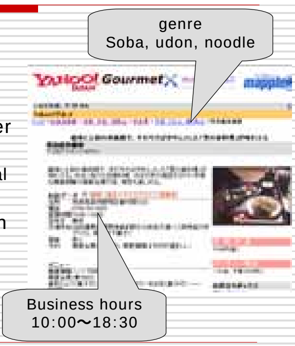
What is information retrieval?

- □ Desired information is changing dynamically with user's circumstance information
  - Weather forecast at specific time in his future address
  - Searching the restaurant information around the current space for lunch
  - Reservation for a concert ticket of one's favorite artist on future off

•User cannot input some keyword by text
•Need to input many keyword every retrieval

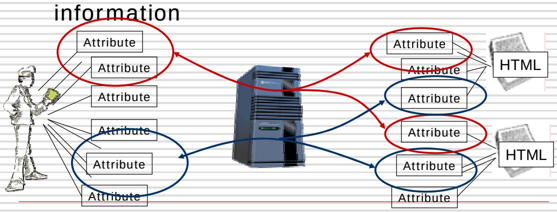
Purpose

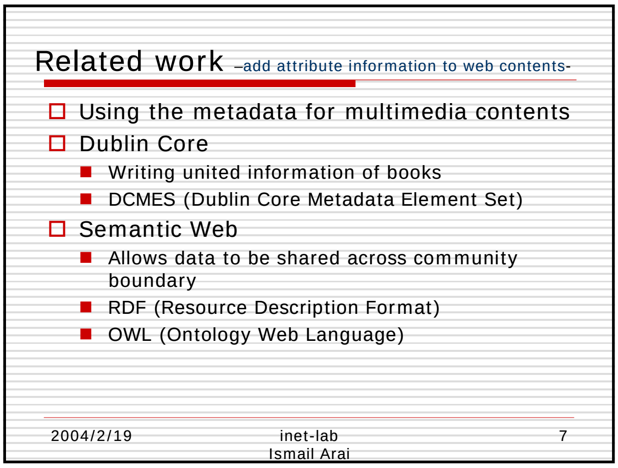
- □ Target
  - Web contents written about mundane life information
- □ Using the information user cannot input by keyword
  - To realize actual retrieval
- □ Save & use the static circumstantial information
  - Cut the cost of inputting keyword



Requirement

- □ Add the circumstantial information to user's query & web contents
- □ Systemize the attribute information
- □ Information retrieval make use of attribute information

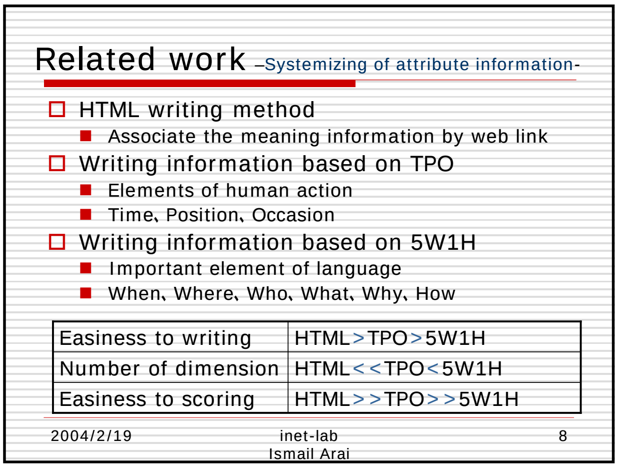




Related work –add attribute information to web contents-

- □ Using the metadata for multimedia contents
- □ Dublin Core
  - ■ Writing united information of books
  - ■ DCMES (Dublin Core Metadata Element Set)
- □ Semantic Web
  - ■ Allows data to be shared across community boundary
  - ■ RDF (Resource Description Format)
  - ■ OWL (Ontology Web Language)

2004/2/19 inet-lab Ismail Arai 7

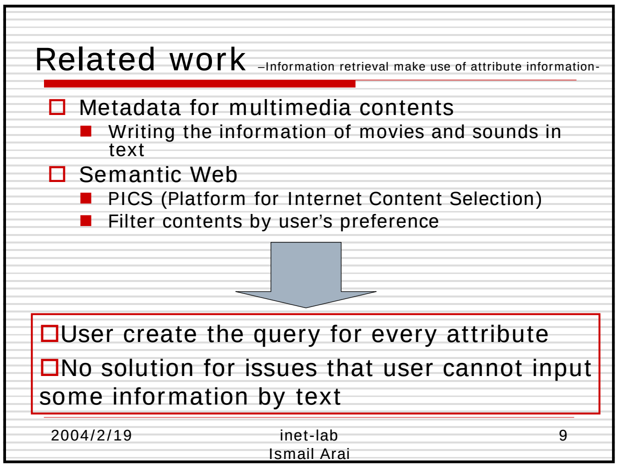


Related work –Systemizing of attribute information-

- □ HTML writing method
  - ■ Associate the meaning information by web link
- □ Writing information based on TPO
  - ■ Elements of human action
  - ■ Time、Position、Occasion
- □ Writing information based on 5W1H
  - ■ Important element of language
  - ■ When、Where、Who、What、Why、How

Easiness to writing	HTML＞TPO＞5W1H
Number of dimension	HTML＜＜TPO＜5W1H
Easiness to scoring	HTML＞＞TPO＞＞5W1H

2004/2/19 inet-lab Ismail Arai 8



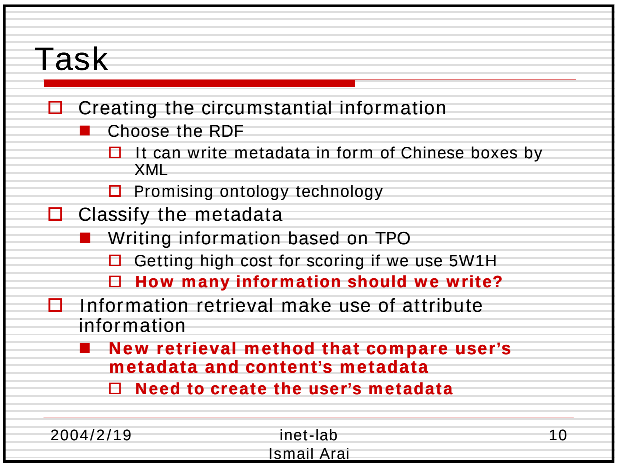
Related work –Information retrieval make use of attribute information-

- □ Metadata for multimedia contents
  - ■ Writing the information of movies and sounds in text
- □ Semantic Web
  - ■ PICS (Platform for Internet Content Selection)
  - ■ Filter contents by user's preference

□User create the query for every attribute

□No solution for issues that user cannot input some information by text

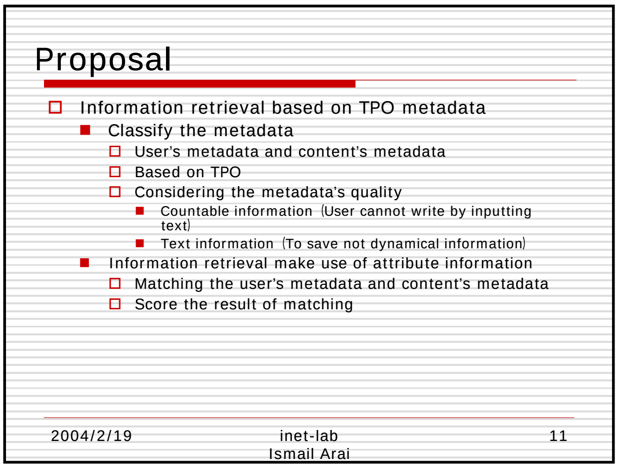
2004/2/19 inet-lab Ismail Arai 9



Task

- □ Creating the circumstantial information
  - ■ Choose the RDF
    - □ It can write metadata in form of Chinese boxes by XML
    - □ Promising ontology technology
- □ Classify the metadata
  - ■ Writing information based on TPO
    - □ Getting high cost for scoring if we use 5W1H
    - □ **How many information should we write?**
- □ Information retrieval make use of attribute information
  - ■ **New retrieval method that compare user's metadata and content's metadata**
    - □ **Need to create the user's metadata**

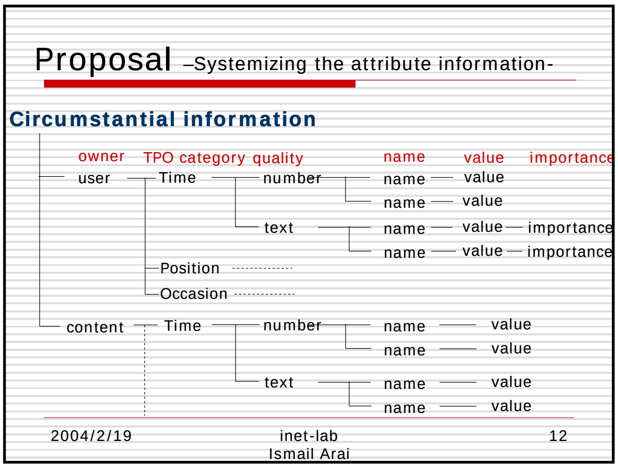
2004/2/19 inet-lab Ismail Arai 10



Proposal

- □ Information retrieval based on TPO metadata
  - ■ Classify the metadata
    - □ User's metadata and content's metadata
    - □ Based on TPO
    - □ Considering the metadata's quality
      - ■ Countable information（User cannot write by inputting text）
      - ■ Text information（To save not dynamical information）
  - ■ Information retrieval make use of attribute information
    - □ Matching the user's metadata and content's metadata
    - □ Score the result of matching

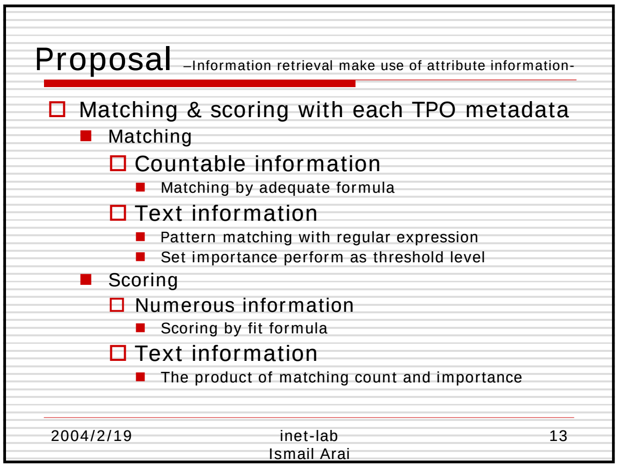
2004/2/19 inet-lab Ismail Arai 11



Proposal –Systemizing the attribute information-

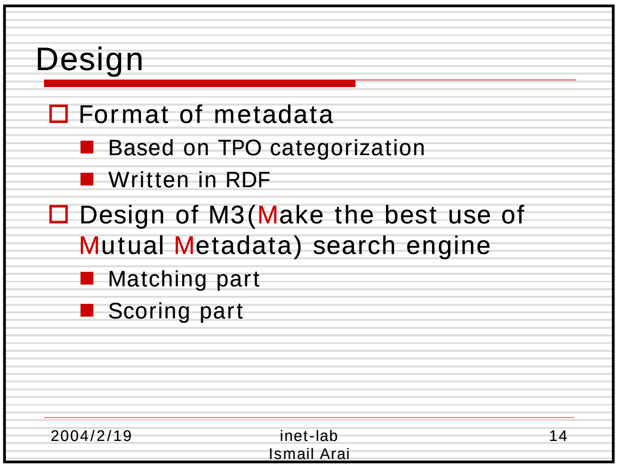
Circumstantial information

2004/2/19 inet-lab Ismail Arai 12



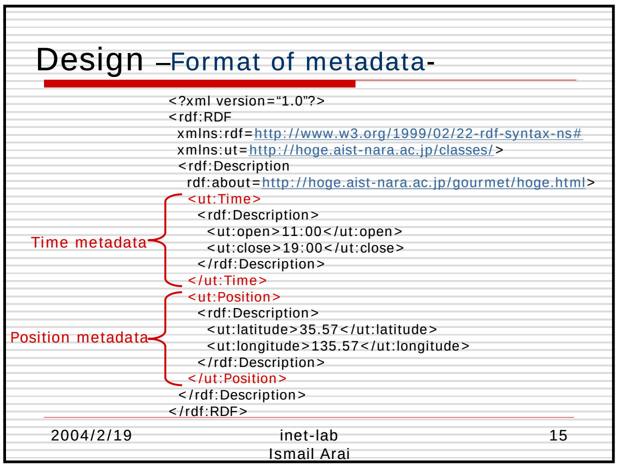
Proposal –Information retrieval make use of attribute information-

- □ Matching & scoring with each TPO metadata
  - Matching
    - □ Countable information
      - Matching by adequate formula
    - □ Text information
      - Pattern matching with regular expression
      - Set importance perform as threshold level
  - Scoring
    - □ Numerous information
      - Scoring by fit formula
    - □ Text information
      - The product of matching count and importance



Design

- □ Format of metadata
  - Based on TPO categorization
  - Written in RDF
- □ Design of M3(Make the best use of Mutual Metadata) search engine
  - Matching part
  - Scoring part

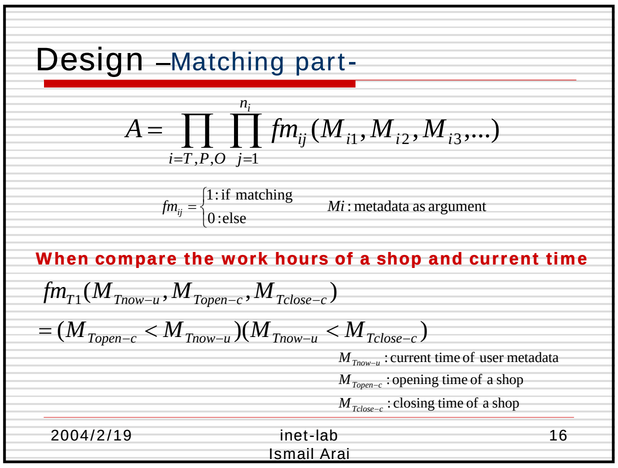


Design –Format of metadata-

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
  xmlns:ut=http://hoge.aist-nara.ac.jp/classes/>
  <rdf:Description
    rdf:about=http://hoge.aist-nara.ac.jp/gourmet/hoge.html>
    <ut:Time>
      <rdf:Description>
        <ut:open>11:00</ut:open>
        <ut:close>19:00</ut:close>
      </rdf:Description>
    </ut:Time>
    <ut:Position>
      <rdf:Description>
        <ut:latitude>35.57</ut:latitude>
        <ut:longitude>135.57</ut:longitude>
      </rdf:Description>
    </ut:Position>
  </rdf:Description>
</rdf:RDF>
```

Time metadata

Position metadata



Design –Matching part-

$$A = \prod_{i=T,P,O} \prod_{j=1}^{n_i} fm_{ij}(M_{i1}, M_{i2}, M_{i3}, \ldots)$$

$$fm_{ij} = \begin{cases} 1 : \text{if matching} \\ 0 : \text{else} \end{cases} \qquad Mi : \text{metadata as argument}$$

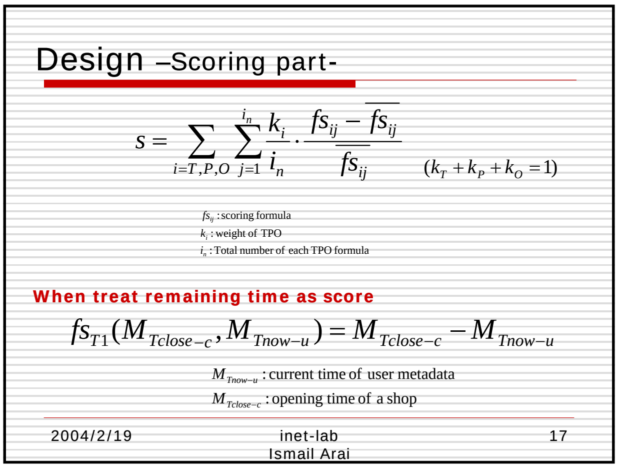
When compare the work hours of a shop and current time

$$fm_{T1}(M_{Tnow-u}, M_{Topen-c}, M_{Tclose-c})$$
$$= (M_{Topen-c} < M_{Tnow-u})(M_{Tnow-u} < M_{Tclose-c})$$

M_{Tnow-u} : current time of user metadata

$M_{Topen-c}$: opening time of a shop

$M_{Tclose-c}$: closing time of a shop



Design –Scoring part-

$$s = \sum_{i=T,P,O} \sum_{j=1}^{i_n} \frac{k_i}{i_n} \cdot \frac{fs_{ij} - \overline{fs_{ij}}}{\overline{fs_{ij}}} \qquad (k_T + k_P + k_O = 1)$$

fs_{ij} : scoring formula

k_i : weight of TPO

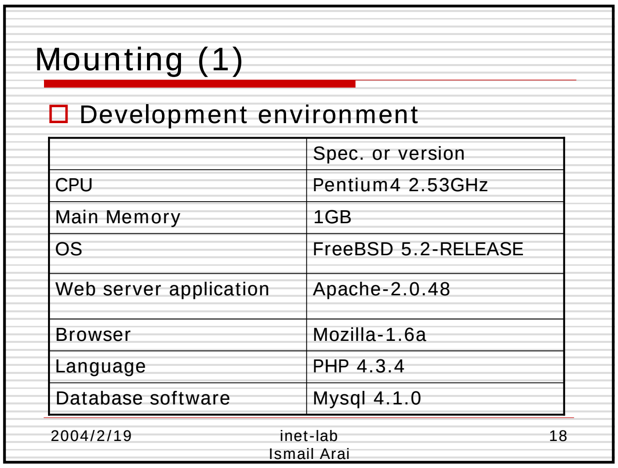
i_n : Total number of each TPO formula

When treat remaining time as score

$$fs_{T1}(M_{Tclose-c}, M_{Tnow-u}) = M_{Tclose-c} - M_{Tnow-u}$$

M_{Tnow-u} : current time of user metadata

$M_{Tclose-c}$: opening time of a shop



Mounting (1)

- □ Development environment

	Spec. or version
CPU	Pentium4 2.53GHz
Main Memory	1GB
OS	FreeBSD 5.2-RELEASE
Web server application	Apache-2.0.48
Browser	Mozilla-1.6a
Language	PHP 4.3.4
Database software	Mysql 4.1.0

Mounting(2)

- Target contents
  - 2,300 of restaurant information in Osaka from Yahoo gourmet
- Extracted metadata

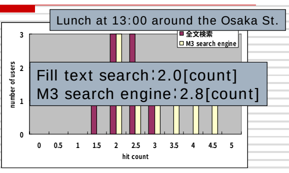
TPO	Property
Time	opening time, closing time, shop holiday
Position	latitude, longitude
Occasion	name, budget average, genre, purpose, menu, credit card, comment

Evaluation

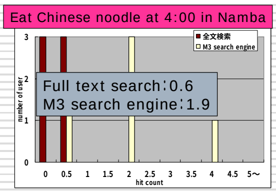
- Accuracy of retrieval
  - Availability of systemizing metadata
  - Adequacy of matching and scoring
- Arriving time
  - Cut the cost of retrieval by make best use of user metadata

Evaluation -correct count of top 10 results-

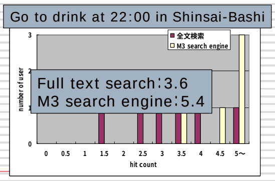
- How many results is correct above top 10 results?


Lunch at 13:00 around the Osaka St.

Fill text search：2.0[count]
M3 search engine：2.8[count]


Eat Chinese noodle at 4:00 in Namba

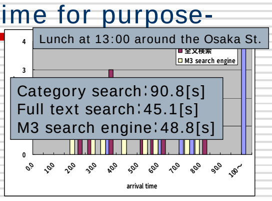
Full text search：0.6
M3 search engine：1.9


Go to drink at 22:00 in Shinsai-Bashi

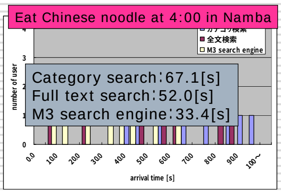
Full text search：3.6
M3 search engine：5.4

Evaluation –Arrival time for purpose-

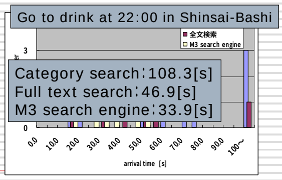
- Measure the time from display the top page to deciding the destination


Lunch at 13:00 around the Osaka St.

Category search：90.8[s]
Full text search：45.1[s]
M3 search engine：48.8[s]


Eat Chinese noodle at 4:00 in Namba

Category search：67.1[s]
Full text search：52.0[s]
M3 search engine：33.4[s]


Go to drink at 22:00 in Shinsai-Bashi

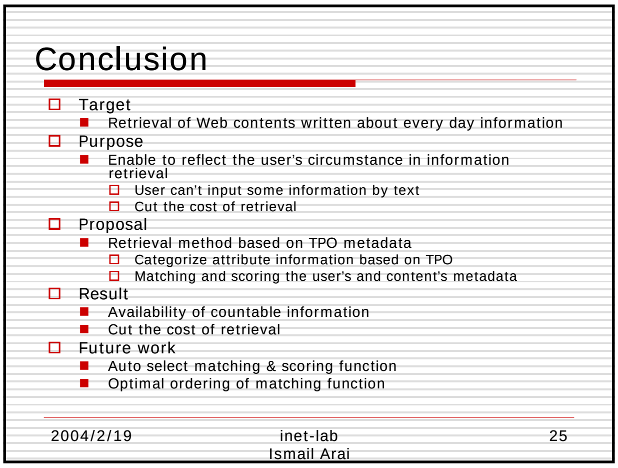
Category search：108.3[s]
Full text search：46.9[s]
M3 search engine：33.9[s]

Results

- Availability of countable information
  - M3 search engine got higher accuracy than full text search
- Cut the cost of retrieval
  - Save the information of user's query
  - Shorter arrival time

Future work

- Auto select matching & scoring function
  - There are some inadequate function
  - Factor of decreasing accuracy
- Optimal ordering of matching function
  - A sharp matching function should be used early
  - Factor of slow retrieval



Conclusion

- Target
  - Retrieval of Web contents written about every day information
- Purpose
  - Enable to reflect the user's circumstance in information retrieval
    - User can't input some information by text
    - Cut the cost of retrieval
- Proposal
  - Retrieval method based on TPO metadata
    - Categorize attribute information based on TPO
    - Matching and scoring the user's and content's metadata
- Result
  - Availability of countable information
  - Cut the cost of retrieval
- Future work
  - Auto select matching & scoring function
  - Optimal ordering of matching function

2004/2/19 inet-lab Ismail Arai 25