Multilingual Information Processing for Digital Libraries

Akira Maeda
Department of Computer Science
Ritsumeikan University
Background

- Increasing popularity of the Internet in various areas in the world
  - Languages used for Web documents are expanded from English to others
- In a digital library (DL), multilingual information processing is essential
  - Even a small library has multilingual materials
  - Accessed from all over the world
Online Language Populations
Total: 561 Million
(March, 2002)

- English: 40.2%
- Chinese: 9.8%
- Japanese: 9.2%
- Spanish: 7.2%
- German: 6.8%
- Korean: 4.4%
- French: 3.9%
- Italian: 3.6%
- Portuguese: 2.6%
- Dutch: 2.1%

Source: Global Reach
Evolution of non-English-speaking online population

Source: Global Reach
Problems

- Many unsolved problems in order to handle multilingual documents in a DL
- From the user’s point of view, fundamental functions for the general use of a DL are: display, input, and retrieval
- From the system’s point of view, digital documents often lack information of the encoding and the language
  - In web documents, charset parameter is not always attached
    - incorrect display on browsers
    - inaccurate indexing on search engines
Proposed solutions

1. Display and input functions for multilingual text which does not depend on installed fonts and input methods

2. Automatic identification of languages and coding systems of Web documents based on statistics and heuristics

3. Cross-languages IR technique which is suitable for documents in diverse domains
MHTML: Display and input functions for multilingual documents
Motivation

- To enable display/input of foreign language text on an off-the-shelf Web browser
  - Usually, the user have to install fonts or input methods for foreign languages
    - Difficult for novice users
    - Impractical for client PCs in public use
  - A method which does not require installing fonts or input methods on the client
- **MHTML**: Multilingual-HTML
Structure of MHTML document

Header
Font
Text (Internal code string)

書
館
図
#1
#23
#41

図書館

Visualized text
Displaying a document using MHTML

Web server

Japanese HTML document (Shift_JIS)

MHTML server

HTTP request

URL

encoding

HTML conversion

MHTML document

Web browser

MHTML class

parameters

url=URL of the document

encoding=Shift_JIS

Displaying MHTML

Japanese font
후와 가족사랑 츠피상

11

http://mhtml.ulis.ac.jp/
http://www.vt.edu/misc/publish/mhtml.html
Digital library of multilingual old tales

http://www.dl.ulis.ac.jp/oldtales/
Summary

- Multilingual text display/input functions without installing anything
- Simple and efficient
- No need to modify existing HTML documents
- Useful for minor languages
- Useful for displaying characters that are not defined in standard character set
Automatic identification of coding systems and languages of documents
# Target languages and coding systems

<table>
<thead>
<tr>
<th>Coding system</th>
<th>Language</th>
<th>Unit</th>
<th>Character range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO-2022-CN</td>
<td>Chinese</td>
<td>7</td>
<td>33-126</td>
</tr>
<tr>
<td>Shift_JIS</td>
<td>Japanese</td>
<td>8</td>
<td>33-252</td>
</tr>
<tr>
<td>EUC-JP</td>
<td>Japanese</td>
<td>8</td>
<td>33-126, 142-254</td>
</tr>
<tr>
<td>GB2312</td>
<td>Chinese (simplified)</td>
<td>8</td>
<td>33-126, 161-254</td>
</tr>
<tr>
<td>Big5</td>
<td>Chinese (traditional)</td>
<td>8</td>
<td>33-126, 161-254</td>
</tr>
<tr>
<td>EUC-KR</td>
<td>Korean</td>
<td>8</td>
<td>33-126, 142-254</td>
</tr>
<tr>
<td>ISO-8859-1</td>
<td>European languages</td>
<td>8</td>
<td>33-126, 161-254</td>
</tr>
</tbody>
</table>
Proposed identification method

- **7-bit coding systems**
  - can be distinguished from 8-bit codes by checking MSB (Most Significant Bit)
  - Subsets of ISO-2022 can be distinguished by escape sequences

- **8-bit coding systems**
  - identified by analyzing the distributions of character codes
One byte code distributions

- **Shift_JIS**
  - code values: 130, 131 (hiragana and katakana)

- **EUC-JP**
  - code values: 164, 165 (hiragana and katakana)

- **GB2312**

- **Big5**
Vector-distance method
(consecutive two bytes unit)

- Takes the connection of consecutive characters into account

$$\cos D'_{c} = \frac{\sum \sum freq_{c}(i, j) freq_{d}(i, j)}{\sqrt{\sum \sum freq_{c}(i, j)^2} \sqrt{\sum \sum freq_{d}(i, j)^2}}$$

$(i = 32, 65...90, 97...122, 128...255)$
## Results of the Vector-distance method

<table>
<thead>
<tr>
<th>Coding system (language)</th>
<th>Correct rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 byte</td>
</tr>
<tr>
<td>Shift_JIS (Japanese)</td>
<td>99.8</td>
</tr>
<tr>
<td>EUC-JP (Japanese)</td>
<td>99.2</td>
</tr>
<tr>
<td>GB2312 (Chinese)</td>
<td>100.0</td>
</tr>
<tr>
<td>Big5 (Chinese)</td>
<td>100.0</td>
</tr>
<tr>
<td>EUC-KR (Korean)</td>
<td>100.0</td>
</tr>
<tr>
<td>ISO-8859-1 (English)</td>
<td>90.9</td>
</tr>
<tr>
<td>ISO-8859-1 (German)</td>
<td>99.4</td>
</tr>
<tr>
<td>ISO-8859-1 (French)</td>
<td>95.1</td>
</tr>
<tr>
<td>ISO-8859-1 (Italian)</td>
<td>98.2</td>
</tr>
<tr>
<td>ISO-8859-1 (Spanish)</td>
<td>90.7</td>
</tr>
<tr>
<td>ISO-8859-1 (Portuguese)</td>
<td>92.8</td>
</tr>
<tr>
<td>ISO-8859-1 (Danish)</td>
<td>69.8</td>
</tr>
<tr>
<td>ISO-8859-1 (Norwegian)</td>
<td>70.1</td>
</tr>
<tr>
<td>ISO-8859-1 (Swedish)</td>
<td>94.9</td>
</tr>
<tr>
<td>Avg.</td>
<td>92.9</td>
</tr>
</tbody>
</table>
Summary

- Proposed an identification method of coding systems and languages of Web documents
- 98% average correct rate for 12 languages and 10 coding systems
- The method does not require discriminating the boundaries of characters for Asian languages
Query Term Disambiguation for Cross-Language Information Retrieval
Cross-Language Information Retrieval (CLIR)

- A technique to retrieve documents written in a certain language using a query written in another language

Who needs CLIR?

- When users can read several languages
  - Eliminate multiple queries
  - Query in most fluent language

- Monolingual users can also benefit
  - If translations can be provided
  - Inexpensive Machine Translation software
Approaches to CLIR

- Translation of target collection
  - Has the advantage of utilizing existing MT software
  - Not suitable for multilingual, large-scale, and frequently-updated Web collection

- Translation of user’s query
  - Translated queries can simply be fed into existing monolingual search engines
  - Simple dictionary translation introduces ambiguity
    - bank: 銀行 (bank to deposit), 堤防 (dike), 土手 (embankment), 川岸 (riverside) …
Flow of query translation

Query translation module

Query in native language

Query translation

Query disambiguation

To get translations

Bilingual dictionary

Query translation candidates

To get co-occurrence statistics

Monolingual corpus

Translated and disambiguated query

Web search engine
### Mutual Information

<table>
<thead>
<tr>
<th>w1</th>
<th>w2</th>
<th>$\text{COT}_{\text{MI}}(w_1, w_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>database</td>
<td>multimedia</td>
<td>3.37</td>
</tr>
<tr>
<td>database</td>
<td>transaction</td>
<td>2.41</td>
</tr>
<tr>
<td>database</td>
<td>relational</td>
<td>2.01</td>
</tr>
<tr>
<td>database</td>
<td>chair</td>
<td>-2.96</td>
</tr>
<tr>
<td>database</td>
<td>soul</td>
<td>-3.43</td>
</tr>
<tr>
<td>database</td>
<td>iron</td>
<td>-4.62</td>
</tr>
</tbody>
</table>
Disambiguation using Mutual Information

<table>
<thead>
<tr>
<th>Query</th>
<th>Translation candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank</td>
<td>銀行, 貯金箱, 岸, 浅瀬, 土手, 堤防...</td>
</tr>
<tr>
<td>Money</td>
<td>富, 財産, 資産, 通貨...</td>
</tr>
<tr>
<td>Trade</td>
<td>商売, 同業者, 貿易, 交換, 道...</td>
</tr>
</tbody>
</table>

Query: (銀行 AND 通貨 AND 貿易) OR (銀行 AND 資産 AND 商売)
Experiment (MI: \(n\) words COT)
Summary

- CLIR method for DL documents in diverse domains
  - utilizing a Web search engine as a corpus
- Effective for very short queries
- Can easily be extended to other languages
Conclusions

- Proposed some solutions to the problem of multilingual information processing for DL
  - Display, input, and retrieval