Key Technologies for Multilingual Information Processing on WWW

Akira Maeda and Shunsuke Uemura
{aki-mae|uemura}@is.aist-nara.ac.jp

Graduate School of Information Science,
Nara Institute of Science and Technology

Abstract

This paper discusses key technologies required to realize a document database which is the multilingual collection of documents typically seen on WWW, and to realize a system which supports easy access to such multilingual information. Specifically, we focus on such techniques as 1) cross-language information retrieval (CLIR), which supports conversion of cultural factors such as units, era names and color names, 2) an algorithm for automatic identification of language and coding system of documents. The goal of our research is to develop a system which supports end-user access to multilingual information by integrating these techniques.

1 Introduction

With the growth of the Internet and WWW in recent years, documents written in various languages are being provided. Although 80% of current Web pages are written in English, it is estimated that over a half of Web documents will be non-English in 2003[1]. Therefore, WWW can be regarded as a huge document database which contains a mixture of documents written in various languages.

However, many problems remain to be solved in order to realize a retrieval system which can handle such multilingual documents in a unified way; e.g. the diversity of document coding systems used in Web pages, the language barrier of a non-native user to formulate a query, and the limitation on inputting the query strings and displaying the search results.

In this paper, we discuss on key technologies we are developing in order to solve some of these problems, namely:

[1]Nua Internet Surveys, http://www.nua.net/surveys/

1. A cross-language information retrieval which supports conversion of cultural factors,

2. An algorithm for automatic identification of language and coding system of documents.

This paper is organized as follows. Section 2 presents some basic information about issues in multilingual information access especially on WWW. In section 3, our system in order to tackle such multilinguality issues is outlined. In section 4, key technologies used in our system are described in detail and section 6 describes conclusions and future tasks.

2 Multilinguality Issues on WWW

2.1 Standards/Protocols

Underlying protocols and standards used on WWW have already been internationalized in some degree.

Initially, HTML was restricted to the ISO 8859-1 character set which is appropriate only for Western European languages. HTML4.0[1] now uses ISO 10646 UCS (Universal Character Set) as the document character set and allows the use of various coding systems which covers most existing character coding systems used on the Internet. Besides, it defines `lang` attribute for specifying the language of the specific bounds of text, and some other issues on i18n are taken into account (e.g. bi-directional text, character references etc.).

HTTP1.1[2] defines the mechanism to indicate the coding system of a page sent from the server to the client by using the `charset` parameter of the `Content-Type` header. Also, it defines the language negotiation algorithm in which a user can specify the preferred language from the multiple language versions of a document.

Unfortunately, these i18n features are not fully implemented on current WWW servers and browsers. The most serious problem is the disuse of `charset` parameter in current WWW documents, which causes incorrect display on browsers and incomplete indexing on search engines.

2.2 Cross-Language Information Retrieval

As an approach to easily access multilingual information, researches on cross-language information retrieval (CLIR) are becoming to be paid much attention[3]. CLIR is a technique to retrieve documents written in one language using a query written in another language. It is effective in cases such as 1) document collection

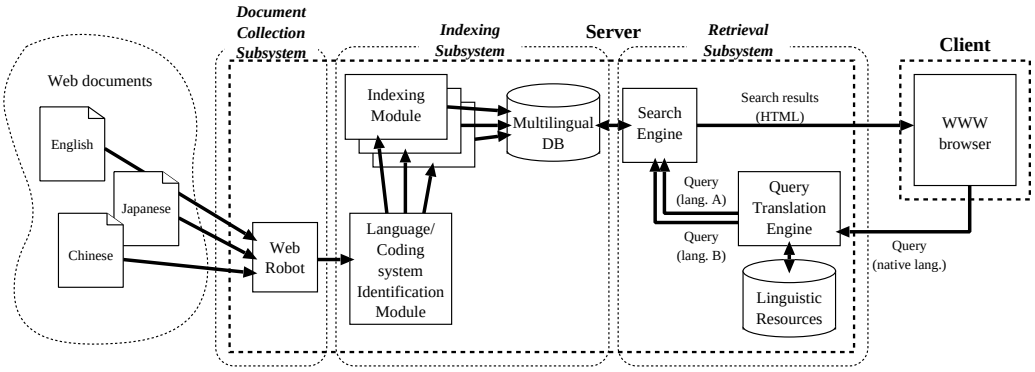


Figure 1: Overview of the system.

itself is multilingual, 2) the user can read the language of document collection, but hard to formulate a query.

Approaches to CLIR can be divided into two categories, i.e. document translation and query translation. Former approach can use existing machine translation (MT) systems and can achieve higher retrieval effectiveness using context information, but it is not feasible for huge document collection such as WWW. In the latter approach, usually user's queries are short and may be a sequence of words, translation ambiguity is a crucial problem[4, 5].

2.3 Text Input/Output

Text input is a crucial issue especially for languages like Japanese, Chinese and Korean (CJK) because CJK has a large number of characters and thus special input method is needed in order to input characters of these languages. Text output is also a crucial issue for most languages other than European languages because the font for foreign languages are usually not installed in computers.

Maeda et al.[6] proposes a solution called MHTML in which users do not need to install input methods or fonts for inputting/displaying foreign language documents.

3 Overview of the System

Figure 1 illustrates the overview of our system.

Our system consists of three subsystems, i.e. *document collection* subsystem, *indexing* subsystem and *retrieval* subsystem.

The *document collection* subsystem collects documents using the Web robot.

In the *indexing* subsystem, collected documents are classified using *language/ coding system identification* module (it is described later), then indexed by language dependent indexing modules, and finally stored in the index database.

In the *retrieval* subsystem, user's query requests are translated into the language the user specified, then documents are retrieved using translated query and returned to the user.

4 Key Technologies

4.1 Automatic Identification of Coding Systems of Documents

4.1.1 Problem of Coding System on WWW

Various languages and coding systems are used in Web documents. 216 coding systems are registered with IANA (Internet Assigned Numbers Authority) at the time of writing.

As a method to add the coding system information for an HTML document, the use of `charset` parameter in `Content-Type` header of HTTP, and the use of META tag of HTML are defined[7]. However, usually these information are not specified, and it causes problems such as incorrect display on Web browsers or incomplete indexing on search engines.

Although multilingual coding systems such as Unicode[8] will diffuse in the future, identification of the language is still needed. Therefore, automatic identification of languages and coding systems can be regarded as an indispensable technique for WWW information retrieval systems. We use the technique in our system in order to choose the suitable language-dependent indexing process.

We propose a statistical method based on one byte code distribution as a simple but efficient algorithm for automatic identification.

4.1.2 Target Languages and Coding Systems

Currently our proposed algorithm supports 4 languages and 9 coding systems as shown in Table 1.

Unit in the table indicates how many bits are actually used in one byte. **Character range** indicates the code range within which graphical characters are assigned.

Although the target coding systems except for ASCII are all multi-byte coding systems, our algorithm treats all coding systems as one byte code for simplicity and efficiency.

Table 1: Target languages and coding systems for the automatic identification.

Coding system	Language	Unit	Character range
ASCII	English	7	33-126
ISO-2022-JP	Japanese	7	33-126
ISO-2022-CN	Chinese	7	33-126
ISO-2022-KR	Korean	7	33-126
Shift_JIS	Japanese	8	33-252
EUC-JP	Japanese	8	33-126, 142-254
EUC-KR	Korean	8	33-126, 142-254
GB2312	Chinese (simplefied)	8	33-126, 161-254
Big5	Chinese (traditional)	8	33-126, 161-254

4.1.3 Training Data

As the training data, we have collected 100K bytes of documents from WWW for each coding systems. One byte code distribution of the training data for each coding systems are shown[2] in Figure 2.

4.1.4 Automatic Identification Algorithm

In the algorithm, firstly the frequency of the code value for each one byte in the document is calculated. Then, identification algorithm which was derived heuristically from the statistical attributes of the frequency distributions of training data is applied. The algorithm is shown in Algorithm 1.

The input is the frequency distribution of the document to be identified and the output is the name of the identified coding system. $freq(x)$ means the frequency of the code value x, $avg_freq(m \ldots n)$ means the average frequency for the range from the code value m to n and $max_freq(m \ldots n)$ means the maximum frequency for the range from the code value m to n. TB, TG and $KANA$ are constants determined from the training data and $TB = 2$, $TG = 2$ and $KANA = 0.02$ were used for the experiment.

4.1.5 Experimental Results

We have conducted the experiment of proposed algorithm in terms of accuracy. Total of 2,754 documents were collected from WWW randomly for the test data.

[2]ASCII and ISO-2022-* are not shown here because these are 7 bit coding systems and can easily be distinguished from other 8 bit coding systems, and ISO-2022-* can be distinguished by the occurrence of the specific escape sequences.

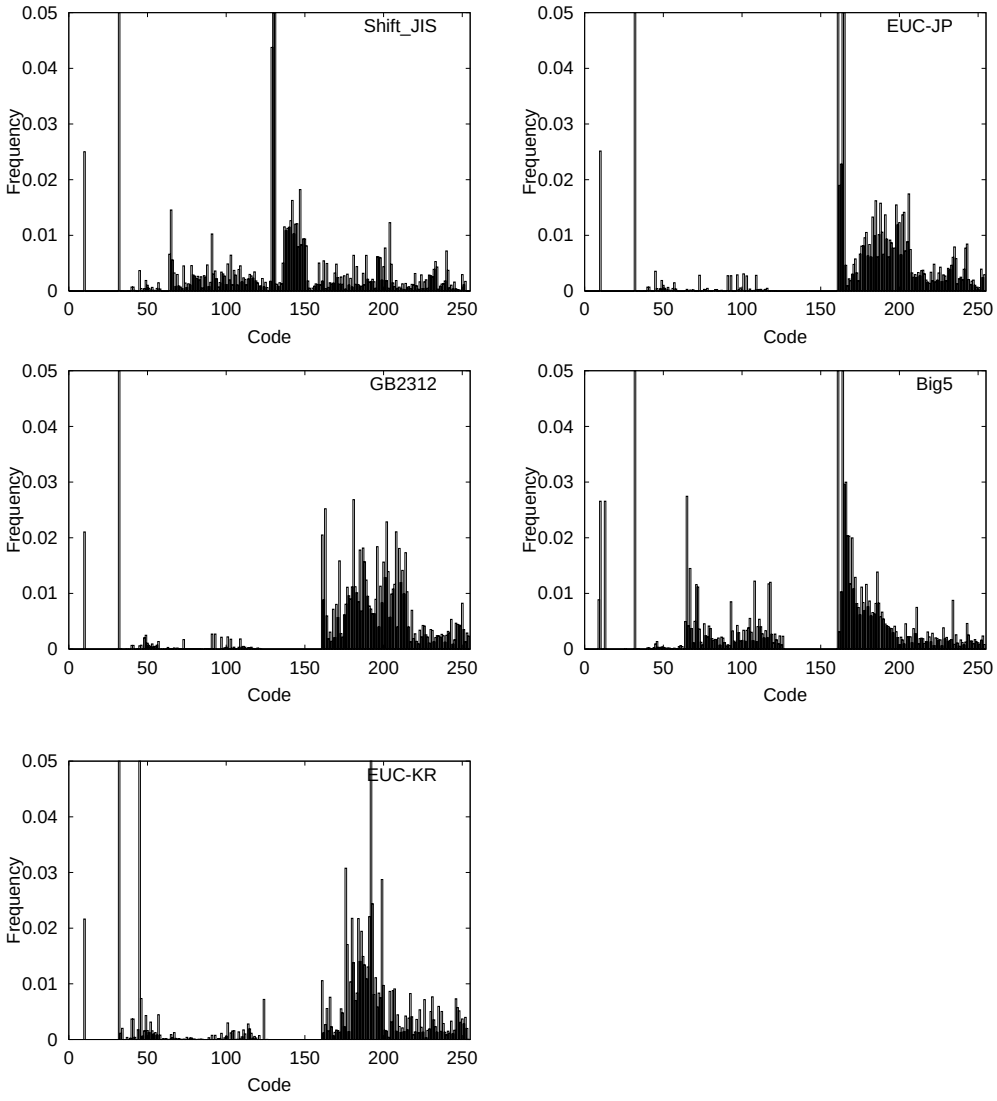


Figure 2: One byte code distribution of each coding systems.

if $avg_freq(128 \ldots 255) = 0$ **then**
  if $freq(27) > 0$ **then**
    return "*ISO-2022*"
  else
    return "*ASCII*"
  end if
else if $avg_freq(128 \ldots 160) \neq 0$ **then**
  return "*Shift_JIS*"
else if $avg_freq(166 \ldots 180)/avg_freq(181 \ldots 254) > TB$ **then**
  return "*Big5*"
else if $freq(164) > KANA \ or \ freq(165) > KANA$ **then**
  return "*EUC-JP*"
else if $freq(192) > KANA$ and $max_freq(161 \ldots 175) < KANA$ **then**
  return "*EUC-KR*"
else if $avg_freq(207 \ldots 215)/avg_freq(216 \ldots 254) > TG$ **then**
  return "*GB2312*"
end if

Algorithm 1: The algorithm for automatic identification of coding systems.

Table 2 shows the accuracy of the proposed algorithm for the test data.

Few documents were incorrectly identified in the experiment. The errors are caused by:

- The length of the document (excluding HTML tags) is short,
- Non-ASCII characters (i.e. 2 byte characters) appears extremely few,
- In case of Japanese, the frequency of hiragana (i.e. Japanese phonetic characters) is low (such as a list of person or organization names).

4.2 Cross-Language Information Retrieval

4.2.1 Query Translation

Existing CLIR approaches require a parallel corpus or a comparable corpus for the disambiguation of translated query term, but these corpora are not readily available. Furthermore, bilingual dictionaries may not be readily available for a particular language pair (i.e. minor languages). Thus our approach focuses on a method which does not depend on available language resources as much as possible. For the disambiguation of translated query terms, we use co-occurrence statistics of two words in the target language corpus. The advantage of our approach is that it does not require rarely available language resources like a parallel corpus or a comparable corpus.

Table 2: Accuracy of the proposed algorithm.

Coding system	**Accuracy**
ASCII	100.0%
ISO-2022	100.0%
Shift_JIS	99.8%
EUC-JP	96.6%
GB2312	96.9%
Big5	98.1%
EUC-KR	98.0%
avg.	98.5%

4.2.2 Query Term Disambiguation

Translation candidates of the source query obtained from the dictionary are disambiguated by following method.

We use *mutual information measure*[9] as the metric for the significance of word co-occurrence tendency. The co-occurence tendency MI of two words w_1 and w_2 is defined as follows:

$$MI(w_1, w_2) = \log_2 \frac{\frac{f(w_1, w_2)}{N}}{\frac{f(w_1)}{N}\frac{f(w_2)}{N}} \tag{1}$$

where N is the total number of paragraphs in the corpus, $f(w)$ is the number of occurences of the word w, and $f(w_1, w_2)$ is the number of times both w_1 and w_2 are fall in a paragraph.

The terms actually used for the translated query can be determined as follows:

1. Calculate the co-occurence tendency MI between all translation candidates derived from distinct source language term pairs,
2. Calculate the average of MIs for all combinations of translation candidates,
3. The term set that has the maximum average MI is selected as the appropriate translation.

Figure 3 shows an example of how the translation candidates are disambiguated in this method.

Note that at least two words are necessary for the source language query in order to apply this method, thus one word query cannot be disambiguated.

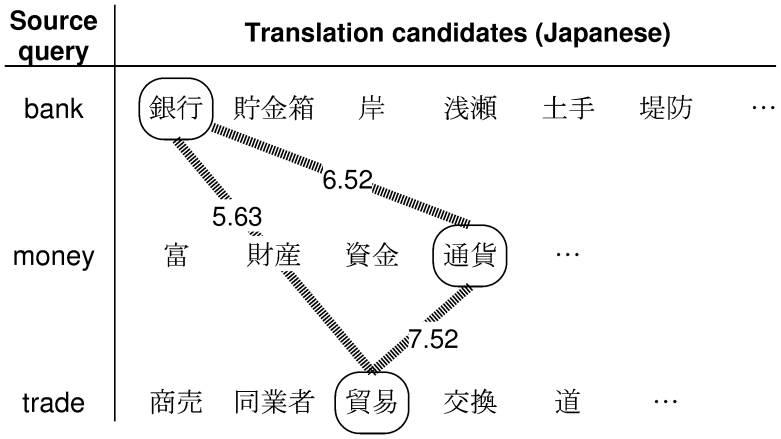


Figure 3: Example of query term disambiguation method.

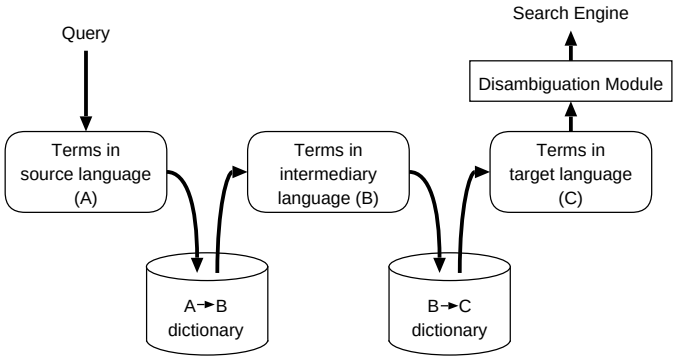


Figure 4: Term translation through the intermediary language.

4.2.3 Translation through Intermediary Language Dictionary

In case that a dictionary which directly translate a query term into the objective language is not available, we consider translation through the intermediary language (e.g. English) (Figure 4). Although it is clearly inevitable that the retrieval effectiveness remarkably drops in such a case, we suspect that the needs for such translations still exist.

4.2.4 Conversion of Cultural Factors

Existing CLIR approaches fundamentally assume that ordinary words or phrases are used for a query term and do not consider translation of cultural factors such as units and era names. Besides, an expression peculiar to a language or a culture

such as color names cannot always be translated simply into objective language. For example, existing CLIR techniques cannot handle a query such as "Find a personal computer which became on sale in 1999 and costs not more than 1,000 dollars." These conversion can be regarded as an extension of the thesaurus expansion technique for CLIR.

Since the present HTML documents are not sufficiently structured, units such as a date or a money value appeared in a document can only be handled as a mere string. However, when a better structured markup language such as XML come into wide use in place of HTML, queries such as the example stated above will acquire more importance.

Considering the implementation of these conversion of cultural factors, measurement units, era names and color names are static and can easily be converted using a formula or a mapping table. However, the exchange rates for currency units change dynamically and cannot be converted in such a way. In this case, it may be possible to obtain the dynamic conversion table of exchange rate from external WWW site (e.g. http://www.oanda.com/). If the conversion of the past exchange rate is needed as the query example stated above, the exchange rate of the past may be obtained from such a WWW site.

5 Conclusions

In this paper, we observed on two essential techniques required to realize a system which can handle multilingual document collection in a unified way, namely CLIR and automatic coding system identification. Although many problems are remain to be solved, we believe that these techniques can improve multilingual information access on WWW in some aspect.

Future work include:

- Evaluation of the proposed CLIR technique,
- Addition of languages for automatic language/coding system identification algorithm,
- System implementation.

References

[1] Raggett, D., Hors, A. L. and Jacobs, I.(eds.): *HTML 4.0 Specification*, W3C Recommendation, World Wide Web Consortium (1998).
`http://www.w3.org/TR/REC-html40/`.

[2] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T.: Hypertext Transfer Protocol – HTTP/1.1, RFC 2616, Network Working Group (1999).

[3] Klavans, J. L. and Schäuble, P.: NSF–EU Multilingual Information Access, *Communications of the ACM*, Vol. 41, No. 4, p. 69 (1998).

[4] Oard, D. W.: Alternative Approaches for Cross-Language Text Retrieval, *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval* (1997).

[5] Grefenstette, G.(ed.): *Cross-Language Information Retrieval*, The Kluwer International Series on Information Retrieval, Vol. 2, Kluwer Academic Publishers (1998).

[6] Maeda, A., Dartois, M., Fujita, T., Sakaguchi, T., Sugimoto, S. and Tabata, K.: Viewing Multilingual Documents on Your Local Web Browser, *Communications of the ACM*, Vol. 41, No. 4, pp. 64–65 (1998).

[7] Yergeau, F., Nicol, G., Adams, G. and Düerst, M.: Internationalization of the Hypertext Markup Language, RFC 2070, Network Working Group (1997).

[8] Unicode Consortium, T.: *The Unicode Standard, Version 2.0*, Addison-Wesley (1996).

[9] Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol. 16, No. 1, pp. 22–29 (1990).