Extension of MHTML to Text Input and Text Search Functions in Multiple Languages on Off-the-shelf Browsers

Shigeo Sugimoto, Shigetaka Nakao, Myriam Dartois, Jun Ohta, Akira Maeda*, Tetsuo Sakaguchi, Koichi Tabata

University of Library and Information Science
Tsukuba, Ibaraki, Japan
{sugimoto, nakao, myriam, jun, saka, tabata}@ulis.ac.jp

*Nara Institute of Science and Technology
Ikoma, Nara, Japan
aki-mae@is.aist-nara.ac.jp

Abstract

The World Wide Web (WWW) covers all over the world. However, browsing function for documents in multiple languages is not widely available yet for casual users. From the viewpoint of digital libraries, functions to display and input multilingual texts are obviously crucial. Multilingual HTML (MHTML) is a browser technology for multilingual documents on WWW. The authors developed a display function of multilingual documents based on the MHTML technology. This technology was intended as a simple, light, easy-to-use and inexpensive technology to view multilingual documents via the Internet. The authors have extended it to text input function in multiple languages on off-the-shelf browsers. The new technology gained by this extension is quite useful to create an environment for end-users of digital libraries where they are able to view and search multilingual documents from any off-the-shelf browser. In addition to the extended MHTML technology, this paper shows an SGML-based full-text retrieval system which is developed using the extended MHTML. It has a user interface to input queries and to display results in multiple languages.

Keywords

Multilingual Document Browsing, Off-the-Shelf WWW Browsers, Multilingual Texts Display and Input, Text Retrieval in Multiple Scripts

1 Introduction

The Internet and the World Wide Web (WWW) are very important infrastructure for digital libraries. English is widely accepted as a common language on WWW for global communication, but on the other

hand, there are a huge amount of documents written in non-English languages on WWW. It is obvious that functions to access information in foreign languages as well as English are crucial for WWW and digital libraries. From another viewpoint, since libraries are inherently multilingual, library information systems have to cope with multilingual library information. In the case of library information systems handling Chinese, Japanese and Korean (CJK) texts, display and input functions for a large set of characters containing non-standardized characters is one of the key technologies to build a digital library.

The authors have been working on a browser technology named Multilingual HTML (MHTML) to display multilingual documents on an off-the-shelf WWW browser even if the browser has no fonts required to display the documents[4][5][7]. Since a document browsing function is realized as a Java applet, users are required to have only an off-the-shelf browser which is capable to run Java applets, i.e., a browser applet running on a browser. We have applied the MHTML technology to a gateway service to view foreign documents and to a multilingual electronic text collection of folktales[2][3]. We have extended MHTML to realize a text input function in multiple languages. We have implemented a Japanese text input server which sends a user interface applet to input Japanese words/characters from a remote client without any Japanese functions. Since the extended MHTML technology is designed independently of languages, it is extensible to other languages. We have also applied the extended MHTML to an SGML-based text retrieval system, which is a full-text database for documents written in multiple languages and has a user interface built based on the MHTML technology.

2 Display and Input of Texts in Multiple Languages

A display function of HTML documents and a text input function on a client are the most basic functions required to access information on WWW. However, these functions for texts in foreign languages are not always provided on a client. The MHTML project had initially started to realize a light, easy-to-use and ubiquitous environment to browse WWW documents written in multiple languages on an off-the-shelf WWW browser. We developed a viewing function to display multilingual documents on a client where fonts for multilingual texts are not necessarily installed. Key aspect of MHTML was that, even if a standard character code set for multilingual texts is widely accepted it is not practical to assume that end users of digital libraries can afford a complete set of fonts for the code set.

Text input function is crucial as well as the display function for browsing documents in multiple languages. The text input function is generally defined as a mapping to a character code or a code string from a user action on an input device, i.e., a single keystroke, a combination of keystrokes, a sequence of keystrokes, a mouse click, and so on. Since inputted texts are usually displayed on a screen, the text input function requires a display function as well. In the case of an ordinary Japanese text input function, for example, a Japanese word or phrase expressed in phonetic characters (Hiragana, Katakana or alphabets) is converted to an appropriate Japanese word or phrase expressed in Kanji, Hiragana, Katakana and/or alphabets. The character code string emitted from the function is displayed on a screen using font locally installed. The phonetic expression in al-

phabets, i.e., transliteration, can be used to input texts from a conventional ASCII keyboard. The mapping function can be located in the client or in a server connected via a network, but the font has to be locally provided. In addition, users have to set up their local environments in accordance with the requirement to input texts such as connection to the mapping function and font installation. It is difficult for an end user who casually accesses foreign documents to set up his/her environment.

3 MHTML

3.1 Basic Concepts

An MHTML server and an MHTML object are the key components of the MHTML technology. The MHTML server fetches a document from a document server, converts it into an MHTML object and sends the object to a client with an applet to display the object on the client. The MHTML object contains the text string of the source document, and a minimum set of font glyphs required to display the text. (The character string in the object is internalized object by object, so that it can not be re-converted to the source character code string.) Since the applet can display all of the characters contained in the source document using only the glyphs sent from the server, the client need no font for foreign languages. Figure 1 shows an MTHML object. The repertoire of the languages of the server primarily depends on the set of fonts stored in the font bank because conversion of documents into ISO-2022-JP-2[6] standard is usually straightforward.

MHTML also has the advantage that it can display a document which contains a non-standard characters. These non-standard characters occasionally appear in Japanese texts and they are called Gaiji in Japanese. A Gaiji can be assigned a code and a glyph locally, but the code has no meaning outside the local machine. An MHTML server can display any character if it is given its glyph and code. There are two ways to make a Gaiji displayable on a client, (1) to add new glyphs to a font file in the font bank, and (2) to add a new font file, which contains glyphs for a set of Gaijis, to the font bank.

3.2 Extension of MHTML – Text Input

Text search is a primary function for information access, i.e., text search in a database and in a document displayed on a screen. Text input function in multiple languages is indispensable to realize text search function in multiple languages. By slightly extending the MHTML technology, we have gained a framework for text input in multiple languages from an off-the-shelf browser. As illustrated in Figure 2, the extended MHTML object contains an identifier of character encoding and character codes in addition to the components of a basic MHTML object shown in Figure 1. A source character code string can be reproduced from an extended MHTML object by replacing every character in the internalized text string by its corresponding source character code. The character encoding identifier is required to make the reconverted text conformant with the ISO-2022-JP-2 standard. Conversion to Unicode[8] is also possible. Based on the framework, we have developed a Japanese text input function for an off-the-shelf browser running on a client which has no Japanese text environment.

Figure 3 shows an outline of a Japanese text input server based on MHTML. The Japanese input server has been implemented using a text input software called Wnn and a user interface applet defined

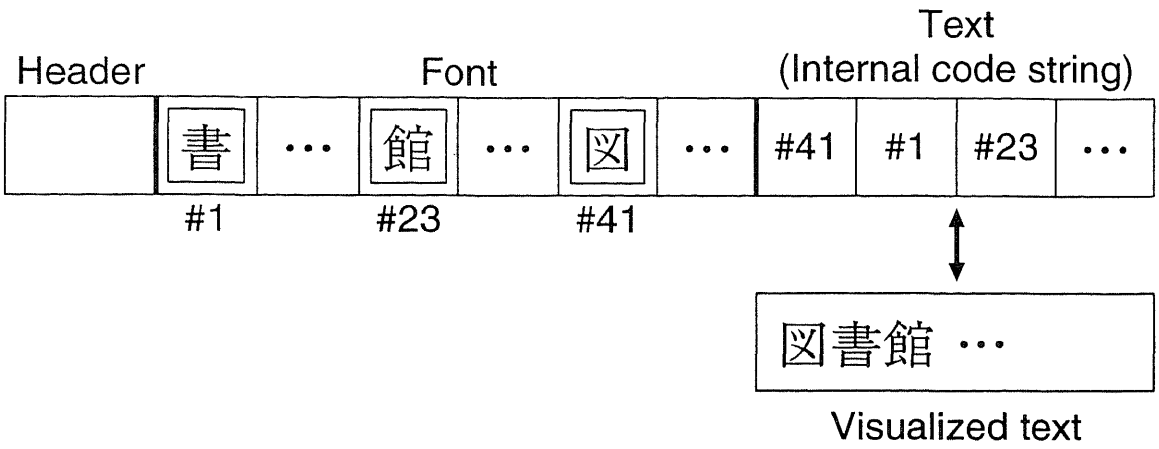

Figure 1: MHTML object

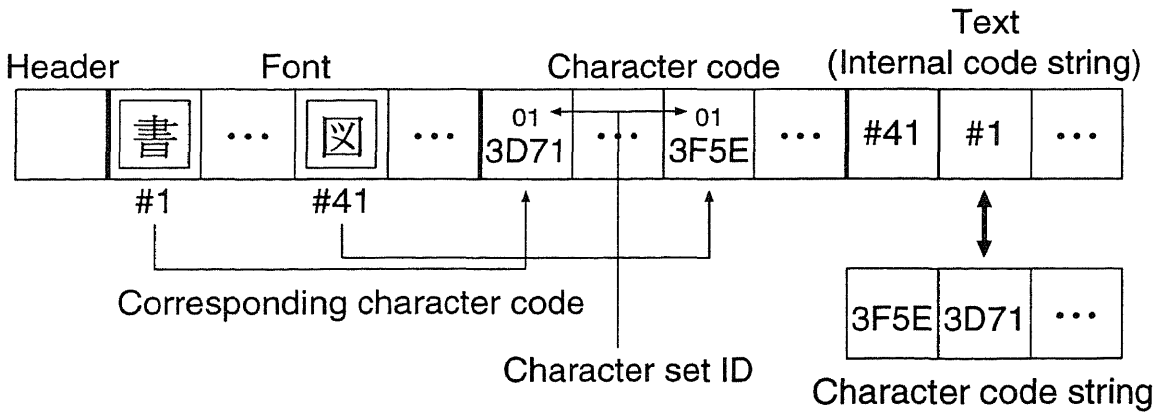

Figure 2: Extended MHTML object

based on the extended MHTML. The text input server (TI server) is located between a client and a WWW server. The TI server receives a Japanese word written in a transliterated form and produces a list of Japanese words. Figure 4 shows a user interface applet to input Japanese texts. This text input applet has a text input field to type a Japanese word or phrase in transliterated form. The CONVERT button means to send the inputted string to the Japanese text input server. A list of words is returned from the TI server and is displayed on the left of the input field. A word selected by a mouse click on the list is displayed in the top line. Users can append additional words/characters to form a complete word or phrase for retrieval. The SUBMIT button means to send the character code string displayed in the top line to a text search function which uses the string for retrieval. The Japanese texts displayed on the applet are sent as an extended MHTML object.

4 An SGML-based Text Retrieval System

We have experimentally developed an SGML-based text retrieval system which is designed to receive texts encoded in multiple scripts. Currently, it can store and retrieve Japanese and ASCII texts, but it is extensible to texts scripted in any character codes. Its user interface is designed using MHTML in order to provide users with ubiquitous accessibility to texts scripted in multiple languages. Figure 5 shows the user interface of the text retrieval system. The window at the bottom is a text input window shown in figure 4. The window at the top shows a list of hits gained by a retrieval. A user can display a source text by clicking on an element of the list.

Figure 6 shows the outline of the system configuration. It receives a text encoded in a regional standard and converts it into the ISO-2022-JP-2 standard. The text is converted into a Unicode-based text to create the index. And, on one hand, the text is stored as it is. ISO-2022-JP-2 standard has multiple character code spaces for character sets of regional languages and defines switching protocol between the spaces. This feature is quite advantageous to use an existing document encoded in a regional standard in the internationalized environment. However, this encoding scheme is disadvantageous to make a text retrieval function simple. On the other hand, since the flat character space given in Unicode is advantageous for a simple text retrieval function, Unicode is employed as the basic encoding scheme to build the index. The index is created based on N-gram. Its user interface is created using the extended MHTML. Thus, this system provides a framework for retrieving texts encoded in multiple encoding scheme (i.e., multilingual documents) with ubiquitous user interface for retrieval.

5 Conclusion

The technology implemented as MHTML is quite simple. The research of MHTML was started to realize a simple, light, easy-to-use and inexpensive environment to read and write foreign texts in the WWW environment. The functions implemented in the research have proved the feasibility of such environment. We believe that the framework realized in MHTML has potential to change paradigm of text input and output in a distributed environment.

The authors have applied the MHTML technology to build user interfaces for an electronic text collection, an OPAC and a full-text retrieval system. They are also collaborating with the internationalization

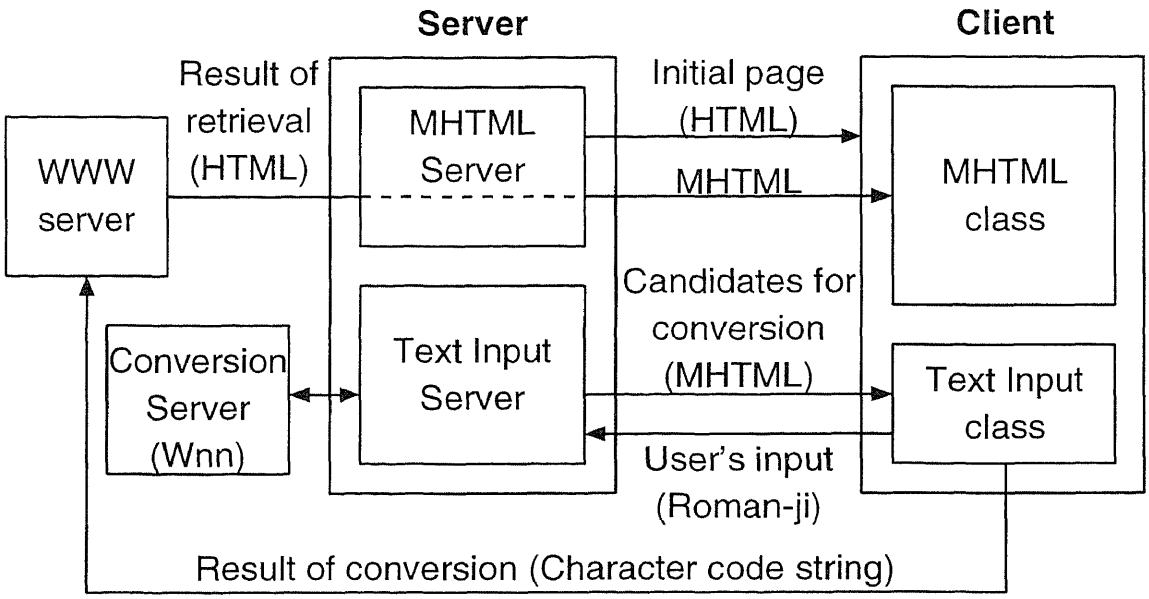

Figure 3: Text Input Server

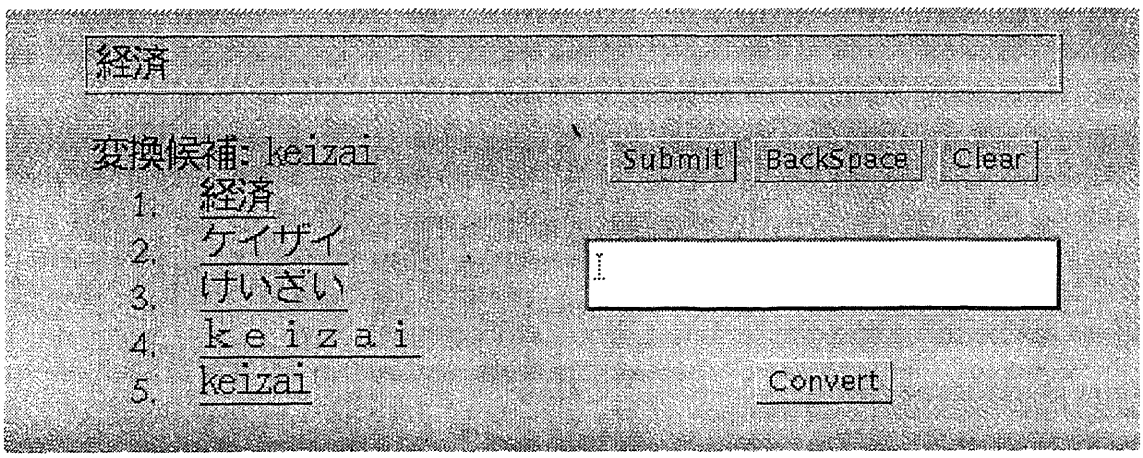

Figure 4: Text Input Applet

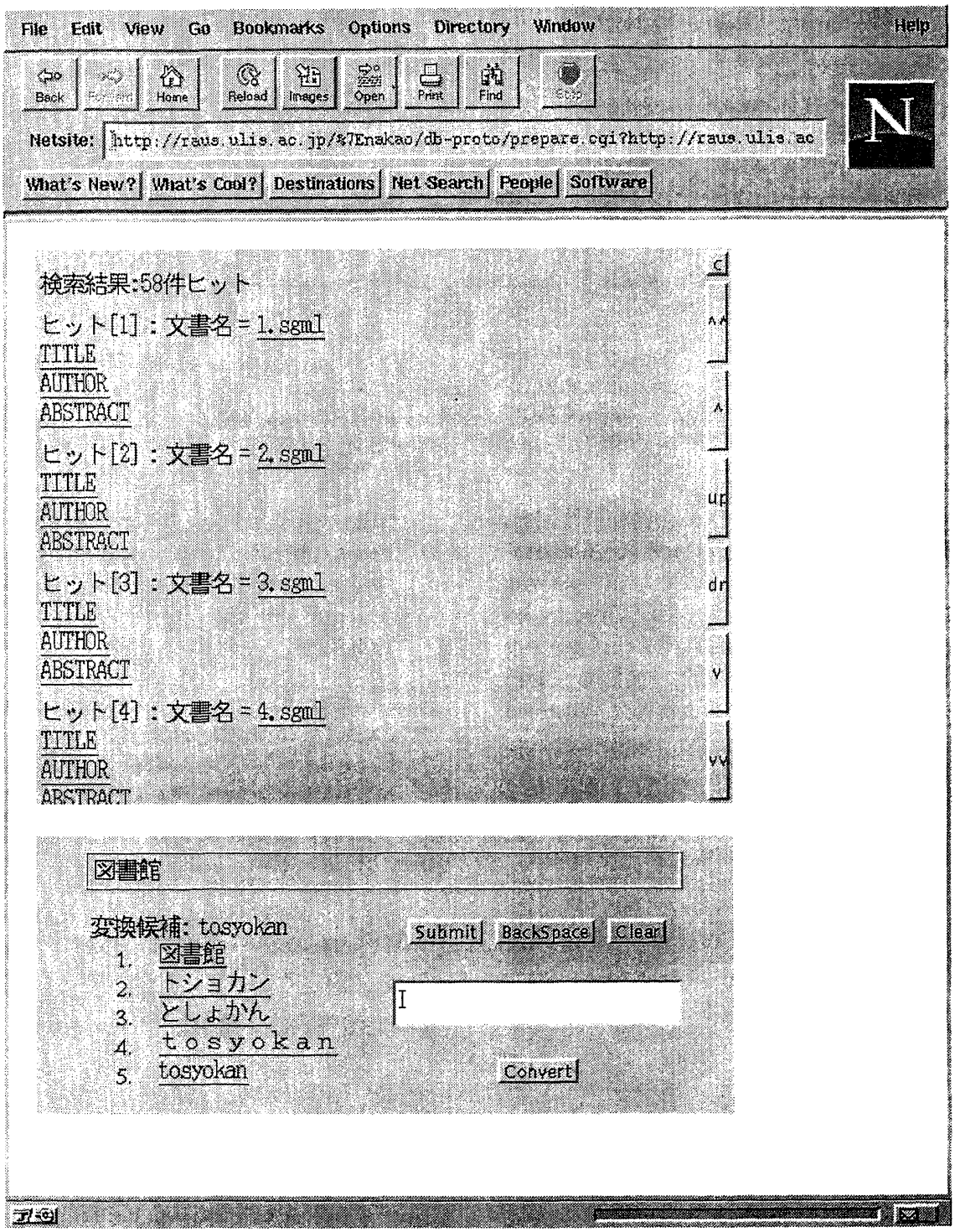

Figure 5: User Interface of the SGML-based Text Retrieval System

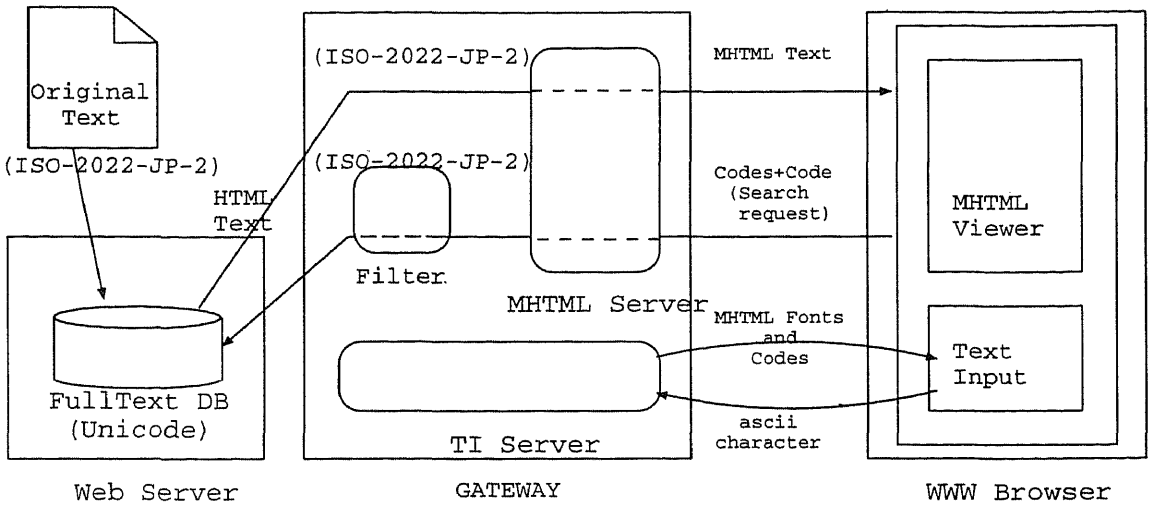

Figure 6: Overview of an SGML-based Full-text Retrieval System for Multlingual Documents

group of Dublin Core[1] to apply MHTML for multilingual user interface for a metadata database. These application systems require rather simple text-based user interfaces but not fancy ones. To extend the repertoire of languages, MHTML requires fonts. Fonts of public domain are required to extend services for not-for-profit services.

References

[1] Baker, T. and Weibel, S.; Dublin Core in Thai and Japanese: Managing Universal Metadata Semantics, Digital Libraries, no.11, pp.35-47, 1998 (in Japanese)

[2] Dartois, M., et al.; A multilingual electronic collection of folk tales for casual users using off-the-shelf browsers, D-lib magazine, 1997, `http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html`

[3] Dartois, M., et al.; Building a multilingual electronic text collection of folk tales as a set of encapsulated document objects: An approach for casual users to browse multi-lingual documents on the fly, Proceedings of ECDL'97, pp.215-231, 1997

[4] Maeda, A., et al.; Viewing Multilingual Documents on Your Local Web Browser, CACM, vol.41, no.4, pp.64-65, 1998

[5] Maeda, A., et al.; A Multilingual HTML Document Browsing System for Clients without Multilingual Fonts, Transactions of IPSJ, vol.39, no.3, pp.802-809, 1998 (in Japanese)

[6] Ohta, M. and Honda, K.; ISO-2022-JP-2; Multilingual Extension of ISO-2022-JP, RFC 1554, 1993

[7] Sakaguchi, T., et al.; A Browsing Toll for Multi-lingual Documents for Users without Multi-lingual Fonts, Proceedings of DL'96, pp.63-71, 1996

[8] The Unicode Consortium; The Unicode Standard, Ver.2.0, Addison-Wesley, 1996