WWWブラウザからアクセス可能な多言語全文データベース構築システム

中尾 茂岳*, ミリアン ダルトア*, 前田 亮**, 阪口 哲男*, 杉本 重雄*, ○田畑 孝一*

A System for building a full-text multilingual database accessible from any WWW browser

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

Abstract

Along with the expansion of the Internet the world-wide network has become a reality and informations from every country and in every language are shared. These informations are written in various languages and therefore multilingual display and retrieval possibilities are indispensable functions. According to this background, we believe that a database allowing casual users to retrieve this kind of information is necessary. This paper persents a multilingual full-text database accessible from any WWW browser and allowing display and input without any constraints related to the language environment of the user.

1 はじめに

インターネットの急速な普及によって全世界的な情報ネットワークが現実のものとなり、このインターネットを基盤としてあらゆる地域・言語圏の情報の共有が可能になりつつある。また、これらの情報は多様な言語によって記述されている。ディジタル図書館に代表されるようなインターネット上のデータベースにおいて多言語の表示および検索は不可欠な機能である。このような背景から、必要な情報を選別するための支援となる、特に個人が利用できるデータベースが必要であると考えられる。

現在、このようなニーズに応えて、商用の全文データベースが幾つか存在している。しかし、重装備なので高価、対応言語の限定、プラットホーム (環境) への依存、インストールが複雑、クライアント側の環境に文字フォントが必要、といった制限がある。

そこで本稿では、軽装備でフリーに配布可能、複数言語の混在を許した多言語向き、JAVAの利用によるプラットホーム依存の回避、必要なフォントとコードをネットワークから入手、といった利点を持つ全文検索データベース構築システムの機能、およびその実現方法について述べる [5]。本システムは我々の開発した多言語 HTML (MHTML) 文書ブラウジングシステムを利用して実現される [1,2,6,7]。

2 システム構成

2.1 概観

システム全体の概観は図1のようになる。データベースへの文献の追加は、原文献をISO-2022-JP-2 コード系でデータベースに与える。検索時には Text Input(TI) を利用してascii 文字の入力から多言語の文字列への変換を行い、検索文字列を生成する [6]。TI サー

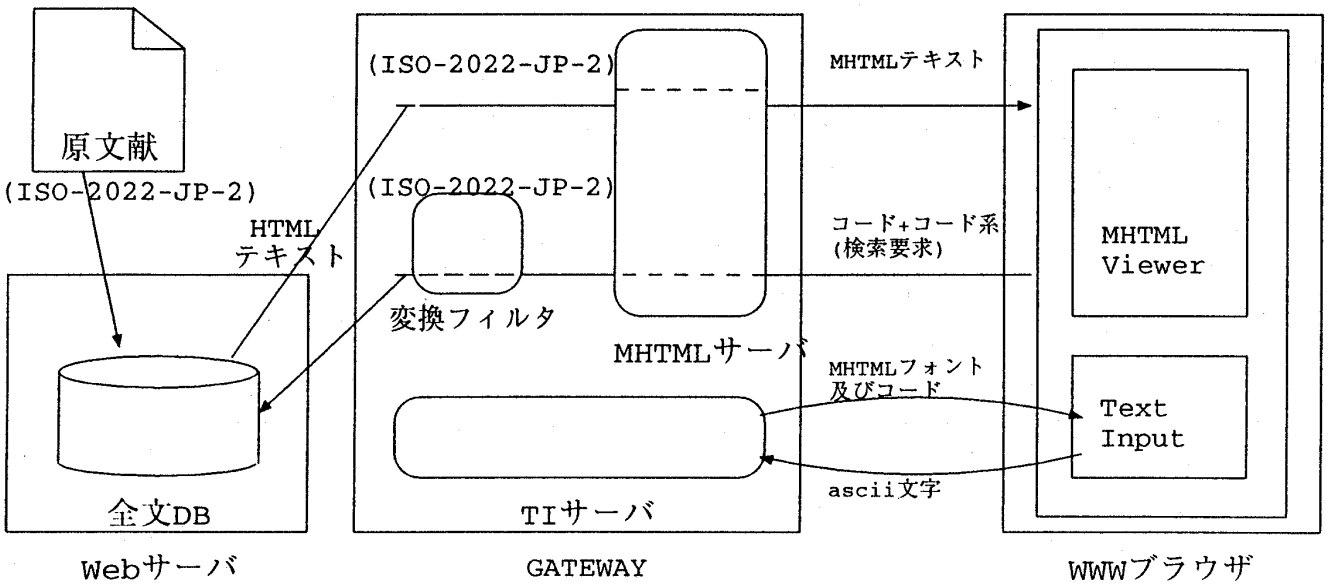

図 1: システムの概観

バからは表示用のフォントと共に検索に使用する文字コードも受け取る。次に生成した文字列にタグや関係演算子による条件づけを指定し、検索の実行によって MHTML サーバを通して検索要求とそのコード系をデータベースに送る。途中、文字コードを ISO-2022-JP-2 に変換する。検索結果は HTML の形式で返され、 MHTML サーバによって MHTML 形式に変換し、ブラウザに文書タイトルの一覧の検索結果を得る。結果の一覧表示からは該当する文書の全部、あるいはタグの指定による一部を閲覧することができる。また、文書は MHTML によって文字化けすることなしに表示されるが、利用者が必要とすれば、原文献データベースから変換の行われていない原文献を提供することもできる。

2.2 データベース

データベースの対象となる文書は SGML で書かれた文書である。新規のデータベース作成時にはデータベースで扱う SGML の文書形式の定義 (タグ) の指定を行なう (図 2)。

データベースに追加する原文献はその文字コードを ISO-2022-JP-2 とする。原文献は文献番号を与えられ、(1) そのままの文字コードで原文献データベースに保存される；(2) 原文献はフィルタを通して内部コードに変換され、 N-gram 方式でインデクシングされる。検索用のインデクスは原文献をフィルタを通して内部コードに変換したものから生成する。内部コードは ISO-10646-1(Unicode) を採用する。

検索時は、受け付けた検索要求の文字列をフィルタに通して内部コードに変換し、変換した文字列に対して検索エンジンが処理を行い、検索結果を HTML 形式で生成する。

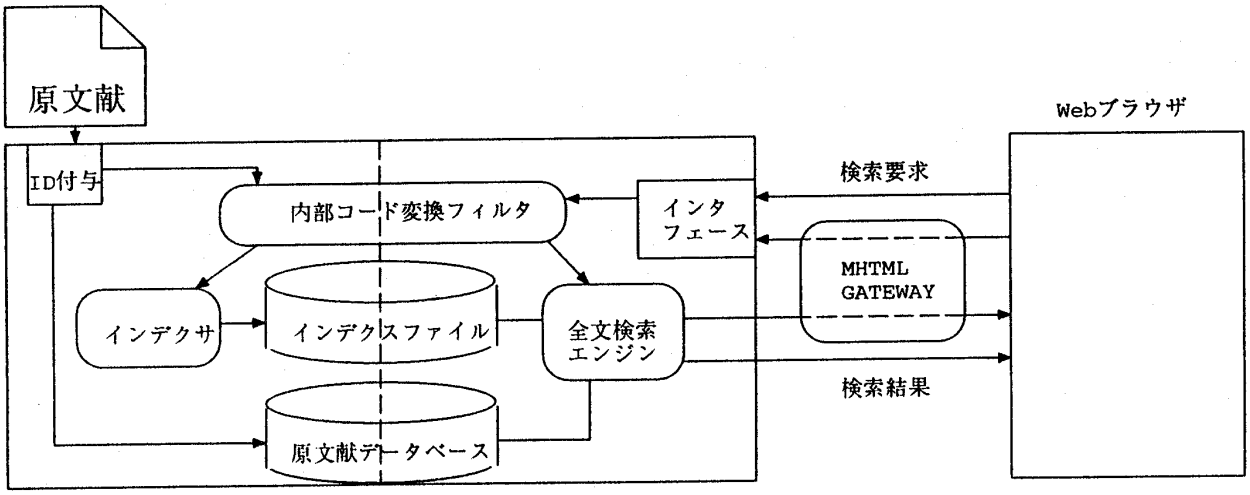

図 2: データベースの内部構成

3 実現方法

3.1 データベース

検索用インデクスの形式は 2 種類あり、一つが文字を対象としたもの、他方が SGML タグを対象としたものである。

文字対象のインデクスは各文字を識別子としてインデクスファイルを用意し、出現した字の文書番号、原文献ファイルの先頭からの文字位置、同じく先頭からのバイト位置、その字を内包するタグの情報を一つのセットとして、該当する文字が出現する度にインデクスファイルに追加する。

SGML タグ対象のインデクスはタグ名を識別子としてインデクスファイルを用意し、開始タグと終了タグを一つの組として、そのタグの出現した文書番号、その開始タグのファイルの先頭からのバイト位置、同じく終了タグのバイト位置を一つのセットとしてインデクスファイルに追加する。

検索時においては検索文字列をインデクスを作成するときと同じ変換を行い、各文字のインデクスファイルを走査する。

データベースは JAVA 言語を用いて開発を行っている。

3.2 データベース API

データベースを WWW ブラウザから利用するための API を用意する。この API によってデータベースへの原文献の追加・削除、検索画面等の WWW ブラウザからデータベースを利用するための雛形を提供する。この API は CGI を用いて実現する。

4 おわりに

このシステムのプロトタイプができあがり、現在試験中である。本システムで多言語表示に用いている多言語 HTML ブラウザ機能についてはそのソースコードを 1997 年 9 月よ

り無償で公開・配布しており、現在までに25ヶ国、110人以上の人によってダウンロードされている。(URL: http://mhtml.ulis.ac.jp)。また、この多言語ブラウザ機能を利用した昔話の多言語電子テキストコレクション (URL: http://www.DL.ulis.ac.jp/oldtales/) については、そのホームページでの募集に(これまでに)応じた16人のボランティアが参加し、コレクションの充実が図られている [3,4]。そして、これには世界中で30以上のサイトからハイパーリンクが張られている。

参考文献

[1] Sakaguchi,T., Maeda,A., Fujita,T., Sugimoto,S., Tabata,K.. A Browsing Tool for Multi-lingual Documents for Users without Multi-lingual Fonts. 1st ACM Intrenational Conference on Digital Libraries (DL'96), pp.63-71 , 1996

[2] Sugimoto,S., Maeda,A., Sakaguchi,T., Tabata,K., Fujita,T.. Experimental Studies on Software Tools to Enhance Accessibility to Information in Digital Libraries. Journal of Network and Computer Application, Academic Press Vol.20, No.1, pp.25-43, 1997

[3] Dartois,M., Maeda,A., Fujita,T., Sakaguchi,T., Sugimoto,S., Tabata,K. . Building a Multi-lingual 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. Research and Advanced Technology For Digital Libraries (First European Conference, ECDL'97), Lecture Notes in Computer Science, Springer, No.1324, pp.215-231, 1997

[4] Dartois,M., Maeda,A., Sakaguchi,T., Fujita,T.,Sugimoto,S., Tabata,K. . A Multilingual Electronic Text Collection of Folk Tales for Casual Users Using Off-the-Shelf Browsers. D-lib Magazine, October 1997, ISSN 1082-9873(URL:http://www.dlib.org/)

[5] 阪口哲男, 中尾茂岳, 太田 純, ミリアン ダルトア, 前田 亮, 杉本重雄, 田畑孝一. ディジタル図書館における多言語情報アクセス, 情報処理学会、電子化知的財産・社会基盤研究グループ (情処Gr研報 Vol.98, No.EIP-3), pp.5-12, 1998年1月.

[6] 前田 亮, Myriam Dartois, 太田 純, 藤田岳久, 阪口哲男, 杉本重雄, 田畑孝一. クライアントにフォントを必要としない多言語HTML文書ブラウジングシステム, 情報処理学会論文誌、Vol.39, No.3, pp802-809, 1998.

[7] Akira Maeda, Myriam Dartois, Takehisa Fujita, Tetsuo Sakaguchi,Shigeo Sugimoto, Koichi Tabata. Viewing Multilingual Documents on Your Local Web Browser, Communications of the ACM, Vol.41, No.4, pp.64-65,1998.

*図書館情報大学 **奈良先端科学技術大学院大学

*University of Library and Information Science

**Nara Institute of Science and Technology