タグ付き文書を対象とした多言語全文検索システム

○阪口 哲男*, 中尾 茂岳†, 前田 亮‡,
杉本 重雄*, 田畑 孝一*

A Multilingual Full-text Retrieval System for Tagged Documents

○ Tetsuo Sakaguchi*, Shigetaka Nakao†, Akira Maeda‡,
Shigeo Sugimoto*, Koichi Tabata*

Abstract

The Internet enables people to share documents written in various languages worldwide. Many documents on the Internet are provided by the WWW. Most of them are markupped with HTML tags. The tags which indicate document elements are very useful for full-text retrieval. The author considers that a full-text retrieval system for tagged multilingual documents is very important to get useful information. This article describes a multilingual full-text retrieval system for tagged documents. It has functions to store and retrieve SGML, XML, and HTML documents. The system handles character code sets both ISO-2022-JP-2 and Unicode for multilingual texts. It is developped with Java for portability. This article also discusses the performance issues of the implemented system.

1 はじめに

インターネットの普及によって、世界各地の様々な文書の共有が可能となっている。それらの文書には英語ではなく他の様々な言語で記述されたものも多く存在する。インターネットでは World Wide Web (WWW) によって文書を提供することが一般的であり、そこでは HyperText Markup Language (HTML) によってタグ付けされた文書が主として用いられている。このような文書が増加している中で、有用な情報を得るために文書の蓄積と検索の重要度が増している。

そのような背景をふまえて、筆者らは様々な言語で記述されたタグ付き文書を対象とした全文検索システムを開発した。本稿ではシステムとその実現について述べる。

2 タグ付き文書の全文検索とその多言語対応

WWW においてハイパテキストの記述に用いられている HTML は急速に普及しており、多くの HTML 編集システムが開発されているほか、ワードプロセッサにおいても固有の文書形式ではなく、HTML 形式で文書の保存が可能なものが増えている。HTML の元になった SGML (Standard Generalized Markup Language: ISO-8879) も文書の電子的な流通や共有のために普及しつつある。また、HTML では文書構造の柔軟な表現が行えないこ

とから、XML (eXtensible Markup Language) が定められ、インターネットを中心としてXML が広まりつつある。

これらのマークアップ言語では文書構造を、文書中にタグを埋め込むことによって表す。タグには開始タグと終了タグがあり、開始タグと終了タグに挟まれた領域がそのタグによって示される文書要素であることを表す。一般的には開始タグは「<タグ名 属性...>」、終了タグは「</タグ名>」のように記述される。このタグによって表される文書要素には、文書のタイトルや作成者、作成日時など文書を特定するための手がかりとなるものも含まれる。

全文検索では、利用者が与えた文字列を含む文書を探すが、文書全体を対象とせずに特定の文書要素に限定することで、より精度の高い検索が可能となる。タグ付き文書ではタグによって文書要素が明示されているので、その情報を含めて索引付けを行っておけば、検索時に利用者が文字列と文書要素を条件として指定することが可能となる。

インターネットでは様々な言語で記述された文書が共有されているが、そこで用いられている文字コード体系には、計算機の機種やオペレーティングシステムに固有のものを除くと、国際標準の ISO-2022 に基づいたものと ISO-10646-1 (Unicode) がある。HTML や XML などの規格では Unicode を標準としているが、日本語、中国語、韓国語を始めとして ISO-2022 に基づいたものがまだ多く用いられている。そのため、文書を蓄積、検索する場合には ISO-2022 に基づいたものと Unicode の両者に対応できる必要がある。

筆者らは、このような観点から 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を設計し、その開発を進めてきた [1]。本稿では実現したシステムと、WWW のみではなく様々な応用システムを構築することができるように汎用性を持たせた API (Application Program Interface) について述べる。

3 タグ付き文書を対象とした多言語全文検索システム

3.1 対象とする文書

本システムでは SGML, XML, HTML に従ってタグ付けされた文書を対象とする。索引付け時にタグ名のリストを与え、そこに含まれているタグによって示される文書要素を検索条件に指定することができる。DTD (Document Type Definition) の解析を行わず、終了タグが略されていない文書を扱う。

文書の記述に用いられる文字コード系としては、ISO-2022-JP-2 と Unicode に対応する。ISO-2022-JP-2 はインターネットにおいて提案された多言語のための文字コード体系であり、ISO-2022 に基づいて定められている。また、従来よりインターネット上で用いられている日本語文字コードの上位互換である。これらとは異なる文字コード体系に対応するために、文字コード変換フィルタを追加することができる。

3.2 システム構成

本システムの構成を図 1 に示す。図で破線の左側が索引付けを、右側が検索を行う。システムに与えられた原文献は、文字コードも含めて一切手を加えずに原文献データベース

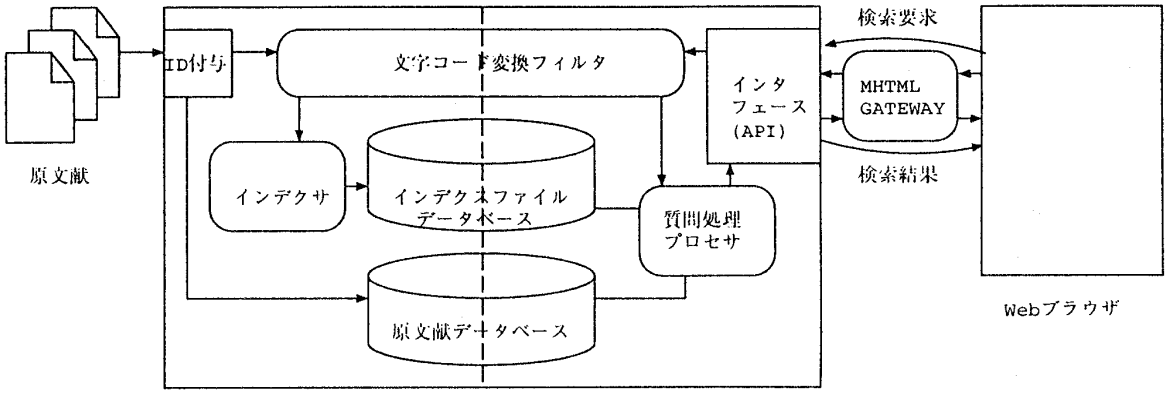

図 1: システム構成

に格納する。同時に文字コード変換を行いインデクサによってインデクスファイルを生成する。インデクスは、全文検索のための N-gram インデクスとタグ位置のインデクスから構成される。

API を介して応用プログラムから検索要求を受け取ると、質問処理プロセサにより検索を行う。検索要求中の文字列も索引付け時と同様に文字コード変換を行う。また、検索の結果として文献を取り出すときは、原文献データベースから取り出す。これにより、文字コード変換による文字の入れ替わりを防いでいる。

API は現在 Java 言語のクラスライブラリの形態で提供しており、次のようなクラスを提供している。

- `DB` (データベースの検索、原文献の取り出し)
- `DBManager` (データベースの管理、文献追加、削除など)
- `Query` (検索質問式の構成)
- `Result` (検索結果集合)

4 システムの実現と利用者インタフェースの実装例

システムの実現は Java を用いて行った。Java を用いることで機種やオペレーティングシステムへの依存性をなくし、システムの可搬性を高めることを狙っている。実際の開発は SPARCstation Ultra1 (200MHz, 256MB-RAM, Solaris2.5.1) 上で JDK 1.1.6 を用いて行った。

また、単に Java による API を準備するのみではなく、UNIX 向けのコマンドラインインタフェースを作成した。これを CGI (Common Gateway Interface) と MHTML ブラウザ [2] と組み合わせ、文字フォントが備わっていない環境でも使用可能な WWW 上の全文検索システムの利用者インタフェースの実装を行った。

実現したシステムの機能確認のために、日本昔話のディジタル文庫 (URL: http://www.DL.ulis.ac.jp/oldtales/) の、日本語、中国語、フランス語、英語の物語を実際に索引付けし、検索を行った。また、性能評価のために図書館情報大学附属図書館の OPAC データ 1 万件を SGML 化したもの (約 22MB) を索引付けした。生成されたインデクスファイルは約 78MB となり、索引付けには約 130 分かかった。そして、検索の応答時間はおよそ 2〜6 秒となった。

一般に Java のプログラムは仮想計算機によって実行されるため、他の言語よりも性能が劣るとされる。しかしながら、現在多くの Java の処理系では実行時コンパイラ (Just-In-Time Compiler) 技術などの利用により、繰り返し演算の性能向上が図られている。また、今回の実現では当初 JDK に標準で提供されているファイル入出力クラスを用いたが性能が低かったので、ファイル入力のバッファリングを行うクラスを Java で新たに定義して用い、検索速度を 10〜20 倍に向上させた。このことから Java は歴史が浅く、様々な組み込みクラスの最適化がまだ十分にされていないものがあると考えられる。

5 おわりに

本稿では、多言語のタグ付き文書を全文検索するシステムとその実現について述べた。Java で開発しているが、本システムはファイル入出力が中心であるため、適切なバッファリングを行うことによって実用上十分な性能を出すことができたと考えられる。また、可搬性を考慮して Java を用いて開発したが、UNIX 以外の MS-Windows や OS/2 などの環境でのテストを行う必要がある。今後は、ドキュメントの整備や、Java クラスライブラリとしてのパッケージ化など、インターネットを通じた配布を目標に開発を進める。

参考文献

[1] 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. 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. 情報知識学会第 6 回 (1998 年度) 研究報告会講演論文集. p.61–64. 1998.

[2] S. Sugimoto, A. Maeda, M. Dartois, J. Ohta, S. Nakao, T. Sakaguchi, K. Tabata. Experimental Studies on an Applet-based Document Viewer for Multilingual WWW Documents – Functional Extension of and Lessons Learned from Multilingual HTML. Proc. of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL '98), Lecture Notes in Computer Science 1513, p.199–214. 1998.

---

* 図書館情報大学 University of Library and Information Science
† 同上 (現在、会社員)
‡ 奈良先端科学技術大学院大学 Nara Institute of Science and Technology