

WWWを利用した特許データベースシステム

谷川 哲司

tanigawa@joke.cl.nec.co.jp

NEC 情報メディア研究所

本稿では、筆者らが開発している WWW の仕組みを用いた特許データの利用システムに関して述べる。本システムでは、WWW の CGI 機能を利用し、リアルタイムで HTML 文書ならびに図面データを生成し、WWW ブラウザに表示している。また、CGI 機能を用いてユーザー管理、データ共有、リアルタイムなグラフィック生成等の仕組みを実現している。

Patent Database System using World Wide Web

Tetsuji Tanigawa

tanigawa@joke.cl.nec.co.jp

Information Technology Research Laboratories, NEC Corporation

This paper describes a Patent Database System using World Wide Web. In this system, HTML document and inline Graphics Data are converted from Patent CD-ROM data by CGI (Common Gateway Interface) program. This system uses shared database, and also realizes personal customization and real time graphics generation.

1 はじめに

インターネットの爆発的流行に伴い、ネットワークを用いた情報サービスが急速に普及しつつある[1][2]。特に、マルチメディアデータの表示能力を備えた WWW(World Wide Web)は、現在最も注目されている情報システムであり、この WWW ブラウザ(クライアント)を使用することにより、世界中のテキストデータ、画像データ、音声データ、動画データ等が利用可能になる。WWW ブラウザは、ゲーム機から UNIX サーバーまで、多くの機種・多くの OS に移植されており、誰もが利用する必要があるデータベースへのアクセス用クライアントとしては最適なツールといえる。

一方、平成 5 年より、特許公報が CD-ROM で公開されるようになり、週に 1GB を越える大量のマルチメディアデータが公開されるようになった。これらの、特許公報データにインターネットや社内 LAN を通じてアクセスし、各自の個人用端末で好きなときにじっくり読みたいという要望が技術者の間に高まっている[3-7]。このようなシステムが実現できると、技術者の開発効率改善、ペーパーレス化に伴う環境破壊防止等、その恩恵は計り知れない。

このように、WWW ブラウザを利用して特許データベースを利用するサービスが求められている。そこで、WWW の CGI 機能を利用した特許データベースシステムを試作した。ここでは、特許庁が発行する CD-ROM 公報のデータを、ユーザーからのアクセス発生時に HTML ファイルに変換し、ブラウザ側に表示する機構を採用した。同時に、CGI 機能を利用して、ユーザー管理、情報共有、データ解析等の機能も実現している。

2 WWW と特許データの概要

2.1 WWW

WWW(World Wide Web)は 1990 年に CERN(欧州素粒子物理学研究所)がインターネット上でハイパーテキストファイルの送受信のために開発した情報システムである。文字データのみならず、マルチメディアデータの取り扱いが容易なため、1994 年ごろより全世界に爆発的に広まりはじめた。WWW で利用するハイパーテキストを記述する言語は HTML(Hypertext Markup Language)と呼ばれ、現在の最新バージョンは HTML3.0 である。現在、普及している WWW ブラウザは、ほぼ HTML2.0 以上に対応している[8]。

2.2 CGI 機能

CGI(Common Gateway Interface)は、WWW サーバー上でプログラムを実行させる機構である[9]。一般的には、UNIX サーバー上で、perl, C, sh で書かれたプログラムを実行し、その出力をブラウザに表示させるために利用される。CGI 機能を利用することで、検索等の高度の処理をサーバー側で実現可能になるが、proxy サーバーや WWW ブラウザのキャッシュが無効になるため、通信に時間がかかるという問題がある。また、使い方を誤った場合セキュリティホールになる可能性も指摘されている。

2.3 CD-ROM 公報の形式

CD-ROM 公報内には、テキストデータ・図面データ・インデックス情報等が納められている。CD-ROM 1 枚あたりには 4000 件程度の特許データが格納されており、総容量は約 400

Mバイトである。この CD-ROM は ISO09660(JIS X 0606)に準拠した論理フォーマットになっており、通常は公開公報が週 2 枚、公告公報が週 1 枚のペースで発行されている。各特許データは、テキストファイル(nnnnnnnn.TXT)、イメージファイル(nnnnnnnn.IMG)、ポジションファイル(nnnnnnnn.POS)で構成される。テキストファイルは SGML(ISO 8879)に準拠した形式となっており、またイメージデータは CCITT,T6 の G4 勧告に基づく MMR2 次圧縮されており、TIFF 形式のファイルとなっている[9]。

これらのデータは、効率的にアクセス可能なように 1000 件単位ディレクトリ、100 件単位ディレクトリ、10 件単位ディレクトリと階層的に格納されている。このファイル格納形式を図 1 に示す。

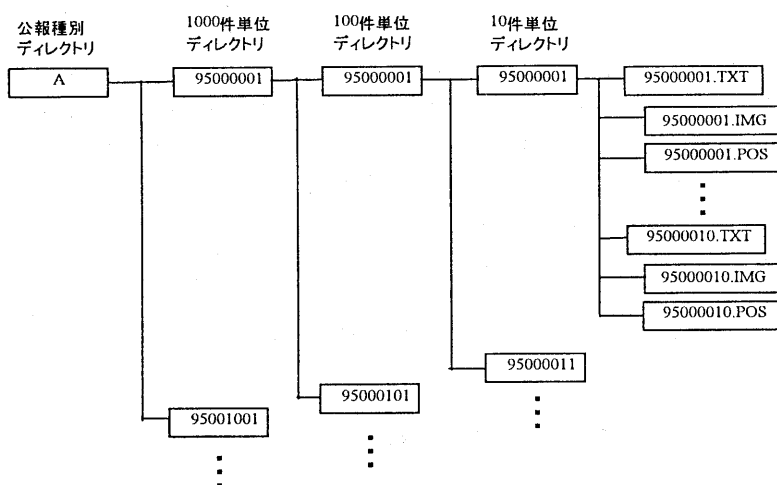


図 1 特許 CD-ROM のデータ格納方式

3 背景と目的

本システムは、特許の専門知識のない一般の技術者・研究者が自席もしくは自宅で安価で効率的に特許データを閲覧できるシステムを目指している。このシステムを開発するにあたり、従来型のシステムや CGI 機能を使わない WWW サーバーの利用も考慮したが、ユーザーの「便利さ」という観点から、CGI 機能を利用したデータベースを採用した。

3.1 従来型特許データベースシステムの課題

スタンドアロン型、もしくはクライアントサーバー型の特許サーバーは高度な処理が可能であり、特許関係の専門家間で広く普及している。特に、平成 5 年 1 月からは CD-ROM 公開公報が発行されており、容易に特許データにアクセス可能な環境が整いつつある。しかしながら、これらのシステムを利用するためには、専用端末／専用ソフトが必要であり、本システムの目的である「一般の技術者・発明者が各自の座席で特許データを調査する」といったパーソナルな用途には向いていない。特に、年間数回程度しか特許システムを利用しない一般ユーザーにとっては、特殊なシステムの操作を覚えること自体かなりの負担である。

3.2 WWW を用いた特許データベースシステム

これに対し、WWW を用いて特許システムを構築する場合は、各ユーザーは普段から使い慣れた、コンピュータ・WWW ブラウザが利用可能であり、導入のためのユーザーの追加投資は比較的少ない。このため、ユーザーインタフェースという観点から、あるいは全社的なシステム導入という現実的な観点から、特許データベースシステムとして非常に有望な方式であると思われる。ところで、WWW を利用して特許サーバーを構築する場合、WWW ブラウザ仕様による制限が大きく、

- SGML 形式で提供される特許データをどのようにして、WWW ブラウザで表示するか。
- TIFF 形式(G4 圧縮)で提供される特許データを、どのようにして WWW ブラウザに表示するか。
- 毎年 60 G バイト近く増え続ける特許データをいかにして格納するか。

といった問題点を解決する必要がある。また、一般向けシステムのため、システム管理者の負担をできるだけ軽減することが望ましい。さらに、HTML 形式の特許データをあらかじめ作成しておき、通常の WWW ブラウザで表示する場合、過去 10 年間のデータの格納に、1T バイトを越える記憶媒体が必要となり、現実的とは言えない。このため、本システムでは、cgi 機能を利用する方式を採用したが、この場合 WWW で一般的になってきたキャッシュによる高速化が利用できず、

- データ変換にかかる時間の短縮、通信にかかる時間の短縮が課題として残る。

4 特許データベースシステムの構成

本節では、特許データベースシステムの構成を述べる。

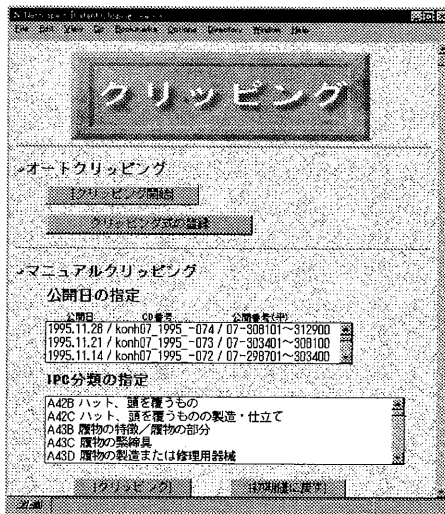


図 2. クリッピング画面



図 3. 概要表示画面

4.1 メニュー構成

本システムでは、データへのアクセスする手段として、キーワード検索、クリッピングという2つの方法を用意している。これらは、個人が登録している検索条件ならびにテンポラリな検索条件を用いて、検索を行う。検索結果は、HTMLのFORM機能を用いて一覧表示され、さらにここで選択した特許に関し、概要表示、全文表示、解析表示を行うことが可能である。図2、図3にクリッピング画面、概要表示画面を示す。

4.2 データベースの構成

特許データベースは、CD-ROM 公報のディレクトリ構造をそのまま利用して構築している。ただし、ハードディスク上にデータを格納する場合は、テキストファイルは圧縮をしている。この方式の最大の利点は、CD-ROM チェンジャーをデータベースとして利用可能なことにある。

実際にブラウザに表示するデータは、このデータをHTMLファイルに変換し、gif形式の画像データ(縮小画像)を組み込んだものになる。

4.3 ユーザー管理機構

本システムでは、個人IDを用いてユーザー管理を行っている。各ユーザー毎に、カスタマイズファイル、ログファイル、検索ログファイル、クリッピングログファイルを作成し、各ページにアクセス発生時にシステムログとは別に詳細なデータ収集を行っている。これらのデータは、過去の検索履歴の利用、個人環境カスタマイズ、グループでの情報共有等に利用する。

5 ユーザーインターフェース・解析機能の強化

WWWのユーザーインターフェースの強化方法としては、HotJava等を用いたインタラクティブな方式が提案されている[10]。しかしながら、より多くの発明者・技術者が利用可能な特許システムを実現するという目的のためには、既存のブラウザで利用可能な方式を採用する必要がある。今回の開発にあたっては、第一にこの汎用性を重視し、機能面の弱点をCGI機能で補うという方針で行った。ユーザーインターフェースの強化点としては、

- ブラウザ別カスタマイズ機能
- グラフィックス機能

を行っている。

5.1 ブラウザ別カスタマイズ機能

現在、我々が開発中の特許データシステムの検索結果を表示可能なWWWブラウザは、確認できているだけでも10種類以上存在し、現在も増え続けている。これらの中にはHTML仕様に独自の拡張を加えているもの、最新のHTML仕様をいち早く採用しているものも存在する。本システムでは、より多くのブラウザに対応するため、デフォルトのCGIの出力コードは、HTML2.0に準拠したもの限定している。しかしながら、より高機能なブラウザを使用している場合に

は、ブラウザから受け取るデータによりブラウザの種類を推定し、これらのネイティブな拡張機能を積極的に利用している。図5に推定結果をユーザーに通知している画面を示す。

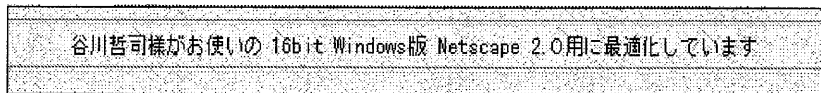


図5. 推定したブラウザの表示例(ブラウザの一部)

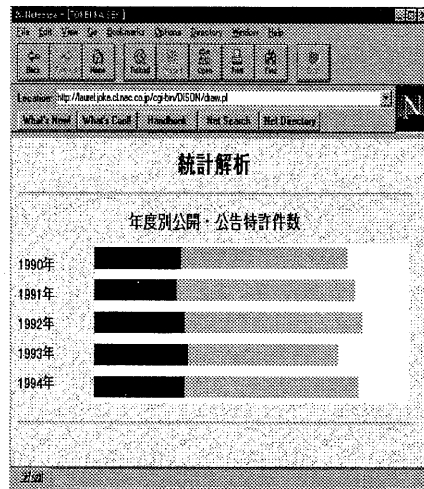


図6. 簡易グラフィックス機能

5.2 簡易グラフィックス機能

ユーザーと特許サーバーのインタラクティブな対話を実現するためには、双方向の通信が必要である。ユーザーからサーバーへの方向のコミュニケーションとしては、HTMLのFORM機能を用いて実現することが可能であるが、逆方向(サーバーからクライアント)に関しては主流といえる一般的手法はまだ存在しない。そこで本システムでは、CGIプログラム中で2Dもしくは3D画像を動的に描画し、クライアントに表示するという手法を開発した。画像描画は、perlプログラムで書かれたグラフィックス関数を呼び出すことで実現しており、簡単に使いこなすことが可能である。また、グラフィックス機能をサポートしたブラウザであれば、問題なく利用可能である。この機能は、解析機能から利用可能であり、実際の利用イメージを図6にまた解析機能と連携する場合のシステムを図7に示す。

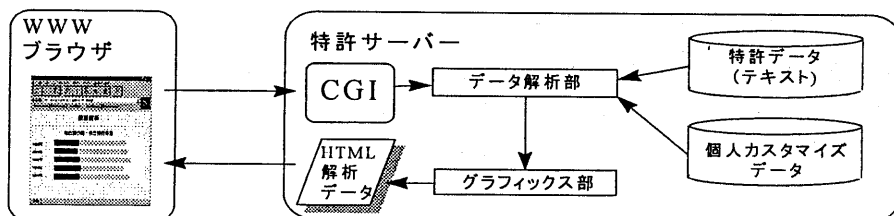


図7. 解析機能と簡易グラフィックス機能の連携

6 まとめ

WWW を利用した特許データベースシステムについて述べた。本システムでは、WWW の CGI 機能を利用して特許データを汎用の WWW ブラウザで表示可能な形式に変換する。同時に、CGI の機能を利用してハイパーリンク情報の埋め込み、ユーザー管理とユーザー登録情報の共有化といった機能を実現した。また、HTML2.0 対応のブラウザでのユーザーとシステムのインタラクティブな対話の実現を目指して、リアルタイムな画像データ生成といった試みを行った。

謝辞

本研究を行うにあたり、貴重な助言をいただいた、NEC 情報メディア研究所 村木一至部長、福島俊一主任、NEC 研究開発技術本部 会森清課長、NEC 情報システムズ 植木久夫氏、ATR 安藤真一氏に感謝いたします。

参考文献

- [1] 木庭袋圭佑、益丘竜介: "World Wide Web によるメディア統合", 情報処理学会研究報告(データベースシステム)、100-21、(1994.10.14).
- [2] 早川和宏、鶴巻宏治、浜田洋: "ユーザー履歴に基づく WWW サーバーの類似検索"、情報処理学会研究報告(情報メディア)、21-2、(1995.5.12).
- [3] 山田武樹: "CD-ROM データをインターネットへ", パテント, 1995, Vol48, No.10, pp.37-39, (1995/10).
- [4] 秋山敦: "新事業展開への提言 2-(1) — CD-ROM 公報を利用した新事業展開に関する考察—", パテント, 1995, Vol48, No.10, pp.19-24, (1995/10).
- [5] 秋山敦: "新事業展開への提言 2-(2) — CD-ROM 公報を利用した新事業展開に関する考察—", パテント, 1995, Vol48, No.11, pp.44-50, (1995/11).
- [6] 笠井浩毅: "CD-ROM 公報と特許情報", パテント, 1995, Vol48, No.10, pp.19-24, (1995/10).
- [7] 松倉秀実: "インターネットと弁理士(1) —インターネット手始め—", パテント, 1995, Vol48, No.9, pp.36-43, (1995/9).
- [8] 吉村信他: "インターネットホームデザイン", 翔泳社, (1995).
- [9] 特許庁: "CD-ROM 公報仕様"
- [10] 神場知成: "The Krakatoa Chronicle: WWW 上のエージェント機能を利用した、対話型パーソナル新聞", 情報処理学会研究報告(マルチメディア通信と分散処理)、73-3、(1995.11.30).