UDC 681.3:608.3

特許検索システムPSEARCH/DBにおける大規模全文検索機能の開発

Full Text Retrieve for Huge Data into Patent System "PSEARCH/DB"

中 本 伸 也*(1)
Shinya
NAKAMOTO

抄 録

新日本製鐵エレクトロニクス・情報通信事業部は1994年より特許検索システム"PSEARCH/DB"の販売を開始した。"PSEARCH/DB"は日本国特許庁発行のCD-ROM化された特許公報に対しネットワーク経由での特許検索をWS/PCを用いたオープンシステム上で行う業務システムであり，その概要を紹介した。検索手法として日付，コード類のインデックス検索に加え，新日本製鐵エレクトロニクス研究所開発の全文検索アルゴリズムを実装し多様な検索環境を提供する。

Abstract

The Electronics / Communication Information Business Dept. of Nippon Steel started marketing the patent searching system "PSEARCH/DB" in 1994.The "PSEARCH/DB" is a business system for searching the information data of patents input through networks with using WS/PC on the open system, instead of searching them in the patent gazette filed in a CD-ROM issued by the Patent Office.In the "PSEARCH/DB", an algorithm for searching full texts, developed by the Electronics Research Laboratory of Nippon Steel is packaged in addition to the index searching method with using the date, codes and so, and it thereby provides a variety of patent searching surroundings.

1. 緒 言

製造業において自社の開発技術，知見，ノウハウは，企業活動の基盤となる財産であり，また，技術，知見を権利化した知的所有権は，国境を超えた大競争時代と言われる厳しい企業間競争を勝ち抜くための有効な武器となる。

日本国特許庁では，国内での特許公報の流通の効率化，また，再利用性の向上を目指し，1993年1月より特許公報の電子化媒体での配布を開始した。この特許庁の電子化媒体での公報配布をふまえ，新日本製鐵エレクトロニクス・情報通信事業部システム商品部では全文検索技術を用いた特許公報の検索システムの開発を1993年1月より開始，1993年10月にスタンドアロン型特許公報検索システム"PSEARCH"を，翌1994年11月よりクライアント・サーバ型特許公報データベースシステム"PSEARCH/DB"を販売開始した。本システムで用いられた全文検索技術は，1992年より新日本製鐵で開発された高速文書検索アルゴリズムを利用した全文検索ソフトウェア"NSEARCH"を用いている。

本稿は，"PSEARCH/DB"における特許検索の構造と"NSEARCH"の実装について述べ，また，その実施例として新日本製鐵知的財産部門におけるシステム導入事例について紹介する。

2. 特許検索システム開発の経緯

2.1 特許公報発行形態の変化

日本国特許庁では公開公報で公開特許等を年間約45万件，登録公報(旧公告公報)で登録特許等を年間約15万件公開し，その配布は冊子で行ってきた。これらの膨大な公報から特許情報を検索するために，JAPIO(日本特許情報機構)では特許公報検索データベースを構

*(1) エレクトロニクス・情報通信事業部
システム商品部 マネジャー
東京都渋谷区代々木3-25-3 〒151-8527
大東京火災新宿ビル ☎(03)5352-2348

築し，一般企業，特許事務所等の特許調査用に運用を行ってきている。

この冊子での発行に加え，特許庁では1993年1月より公開公報，1994年1月より公告公報(現登録公報)を，電子化された情報としてSGML(Standard Generalized Markup Language)へ変換した後CD-ROM媒体での配布を開始した。SGMLは文字情報としてのテキストファイルと画像情報としてのイメージファイルから構成され，冊子に比し容積効率が高く，検索情報，表示情報のコンピュータへの取り込みが容易な形式である。そのため，配布媒体のCD-ROM化は特許公報のインハウスデータベース構築を従来に比べ安価に可能にすることとなった。

2.2 特許調査へのCD-ROM公報の利用

特許情報の性質として，“権利情報”と“技術情報”としての利用価値，また，情報管理体制としては日々管理と累積管理を想定してCD-ROM公報の利用方法を検討し，本開発で2種の利用形態を想定した。

a) 最新の特許情報の監視調査

b) 過去に溯っての特許情報の遡及調査

監視調査は，週3回発行される特許公報中の自社に関連する技術分野に関る特許情報の調査，遡及調査は過去に溯っての特許調査を行う用途である。例えば，技術情報としての利用を考えると，前者は現在事業化している分野の調査，後者は今後事業化を検討している分野の技術動向調査の用途となる。いずれの用途でも，事業を行う際に他社／世の中の動向を調査することにより企業の技術戦略を策定するが，その重要な判断材料となる情報である。以上の利用形態の想定をもとに，まず監視調査用に安価なスタンドアロン型特許公報検索システム“PSEARCH”の商品化を行った。

遡及調査用システムは，従来の特許公報の配布形態，後述のシステム技術の問題により，そのデータベース構築には膨大な投資が必要であった。特許のCD-ROM化，IT技術の発展によりインハウスデータベースを安価に構築する環境が整ったため，特許公報データベースの導入を検討する企業が増えてきた。これに対し，大規模なクライアントサーバ型特許公報データベースシステム“PSEARCH/DB”の商品化を行うこととなった。以降，“PSEARCH/DB”の商品化を中心に紹介する(**図1**参照)。

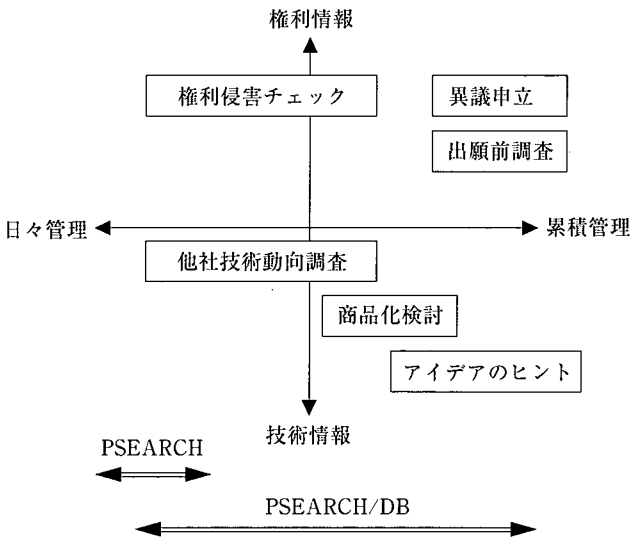

図1 特許調査とシステムターゲット

2.3 IT技術の発展

特許公報は年間約140枚のCD-ROMで発行される。一枚のデータ容量は約500MBとして年間総データ量は70GBとなる。また件数としては前述のように60万件／年が発行される膨大なデータである。従来の技術では，汎用計算機と光磁気ディスクを用いて構築する手法が一般的であったが，以下のようなIT技術の発展によりUNIX系の安価な分散システムでの構築が可能となった。

1) データの蓄積

・CD-ROMオートチェンジャによる大量なCD-ROMの一括管理

・ハードディスクの低価格化による安価な大容量記憶媒体確保

・ハードディスクのRAID構成によるデータの信頼性確保

2) データの検索

・RDB/全文検索技術の発展による検索速度/検索データ量の向上

・マルチCPUマシンによる安価な高速検索環境の実現

・メモリの低価格化による検索環境の価格低下

これらの環境は1993年頃より現実化し，1997年現在では更に安価，高性能な環境となっている。

3. 特許データベースシステム

3.1 特許データの構造

特許庁発行のCD-ROM公報のファイルシステムは，ISO9660準拠のフォーマットであるため，PC-DOS，Windows，UNIX等のOSが可読な形式となっている。

3.2 要求機能

要求される機能は以下の4種に大別される。

1) 検索機能

・インデックス検索機能
　日付，番号，特許分類等のコードでの検索

・全文検索機能
　特許本文中の語句で検索する機能

・検索式解析機能
　特許検索独特の検索式を解析し，前述の検索機能に渡す機能

2) 結果件数表示機能

・検索結果の件数を表示する機能

3) 一覧表示機能

・検索結果の一覧を表示する機能

4) 出力機能

・表示機能
　検索された公報内容をディスプレイ上に表示する機能

・印刷機能
　検索された公報内容をプリンタに出力する機能

クライアントから見たアプリケーションの機能としては上記のような非常に単純な機能となる。

3.3 動作環境

具体的な動作環境としては以下のような技術を使用し，ワークステーション(WS)/パーソナルコンピュータ(PC)でのオープンシステム環境を前提とした。

1) データベースサーバ

・Hard Ware　：NSSUNワークステーション
　　：RAID-5ディスク
　　：CD-ROMオートチェンジャー

・OS　：Solaris2.X
　　：Solaris1.X

・検索システム　：Oracle（RDB：Relational Data Base）
　　：NSEARCH(Full Text Search)
・CD-ROMファイルシステム
　　：Tracer社製CD-ROM オートチェンジャ用
　　ファイルシステム
・開発言語　　：C SQL yacc
・通信プロトコル：TCP/IP

2)検索クライアント(Windows用)
・ハードウェア　：PC-AT及びその互換機
・OS　　：MS-Windows3.1
・開発言語　　：Visual C++
・通信プロトコル：TCP/IP(Winsockを使用)

3)検索クライアント(WWW用)
・ブラウザ　　：NetscapeNavigator2.X
・開発言語(cgi-bin作成のため)
　　：C Perl

3.4 基本構成

3.4.1 全体

論理的アプリケーション構成は**図2**のように，1)検索アプリケーション，2)通信アプリケーション，3)各サービスアプリケーション，4)各資源の4層構成，また物理構成はクライアント・サーバの2層構成を採った。通信はsocket I/Fを使用し，検索サーバとOracle/NSEARCH間の通信はそれぞれsocket経由で，SQL及びNSEARCH検索プロトコルで検索が行われる。

3.4.2 検索部

検索機構はRDB(Oracle)及び全文検索(NSEARCH)を用いる。特許情報検索は一般に検索式によって行われるため，検索キーによって前記2種の検索機構を使い分ける必要がある。検索式の例を式(1)に示す。

IC=(A01K?+A01L?)* TXPA="新日本製鐵"　……(1)

式(1)の検索キーは，
IC(国際特許分類：Oracleに検索データ有)
TXPA(出願人の全文検索：NSEARCHに検索データ有)
が使用されている。検索実行時に式(1)は式(2)，(3)，(4)の単項検索式及び(5)の組み合わせ式に展開される。

IC=A01K?　……(2)
IC=A01L?　……(3)
TXPA="新日本製鐵"　……(4)
(1)=((2)+(3))*(4)　……(5)

(2)，(3)式の検索結果の論理和と(5)式の論理積が求める解である。

PSEARCH/DBでは検索キーによって行う検索を切り替える機構を採る。(2)，(3)式はOracleで検索を行い，(4)式をNSEARCHで全文検索，(5)の組み合わせ式を検索サーバで実行し解を得る手法を取った。この手法により，検索者は検索機構の差を意識することなく検索式の作成が可能となる(**図3**参照)。

3.4.3 表示データ

オートチェンジャーを用いた場合，CD-ROMのドライブへのかけ換えを行うためデータ取り出し時間は最大40秒程度を要する。この時間はクライアントで表示命令を行って実際に表示が行われる時間となる。PSEARCH/DBでは，ディスク上に表示データの一部を持たせることによって初期レスポンスタイムの向上をはかる手法を採った。

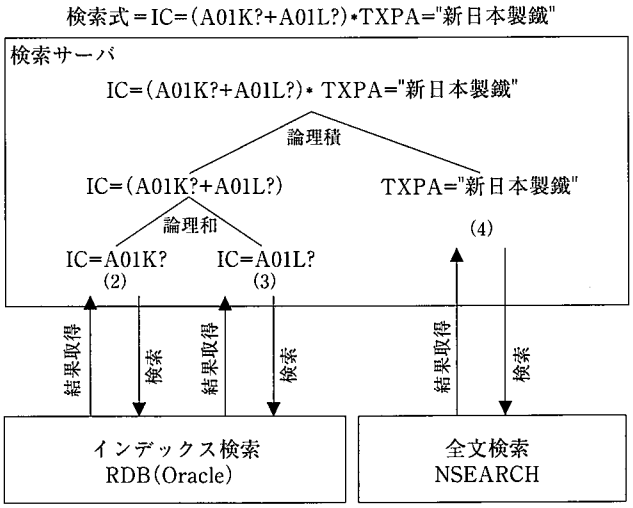

図3　検索式の実行

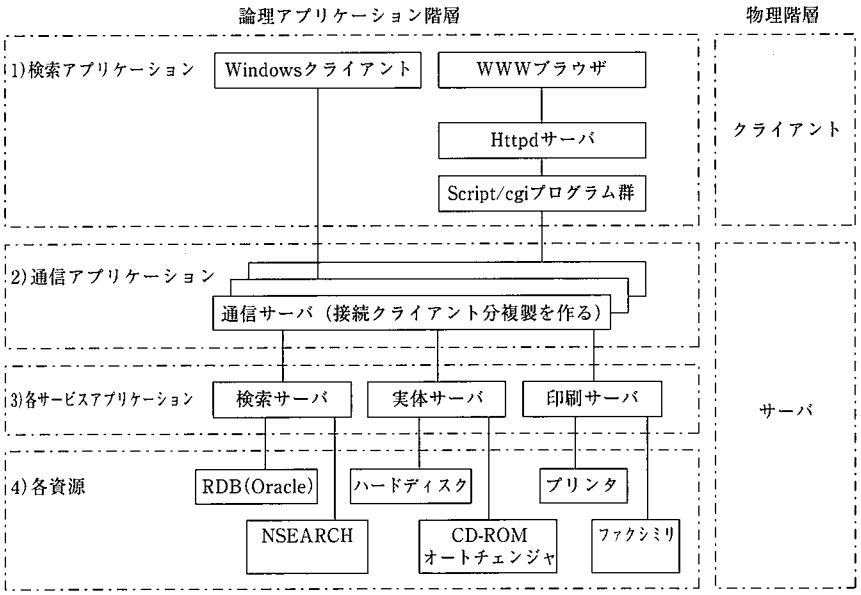

図2　システム全体構成

3.5 実装

以上の実装をサーバ上で行った。データベース構成はRDB/全文検索を行う検索データと表示データから構成される。これらの構成情報の定義ファイルを作成し，各検索/表示部は定義ファイルの情報によって動作する構成を採る。また，1ファイル2GB以下というOSの制限から全文検索用ファイルは1ファイルにできないため，ファイルを分割して検索を行う実装を行い，RDBの1インスタンスに対し全文検索ファイルが複数対応する形式を採った。**図4**の例では検索期間でファイルを分割している。

4. 特許検索システムの導入例

PSEARCH/DBは1994年の販売開始以来様々な導入事例があるが，ここでは1997年9月現在最も大規模な導入事例として新日本製鐵知的財産部門での導入事例について紹介する。

4.1 ネットワーク構成

サーバ機を中央に配置，各製鉄所(拠点)から既存のNS-INS(社内専用回線網)によるTCP/IPネットワーク経由でサーバにアクセスする集中管理のネットワーク構成を採った。(**図5**参照)

4.2 データベース配置

データベースサーバは以下の機器で構築した。

- CPU　：NSSUN-SP20
　　：NSSUN-SP1000 4CPU Model
- DISK　：NSSUN CL2000 4GB Disk × 20
- OS　：Solaris1.2(NSSUN-SP20)
　　：Solaris2.5.1(NSSUN-SP1000)
- CD-ROM AutoChanger
　　：Pioneer DRM5004X(4 Drive Model) × 2台
- 検索システム　：Oracle V7.2.2
　　：NSEARCH V3

システム構成を**図6**に示す。

1997年9月1日現在**表1**に示す公報を登録して運用を行っている。

表1 登録した公報の種別と件数

公報種別	蓄積期間	件数(千件単位四捨五入)
公開公報	1993年1月～	1,934件
登録実用新案公報	1993年7月～	40件
公表公報	1996年1月～	21件
再公表	1996年1月～	2件
公告公報	1994年1月～	388件
登録公報	1996年5月～	195件
計		2,580件

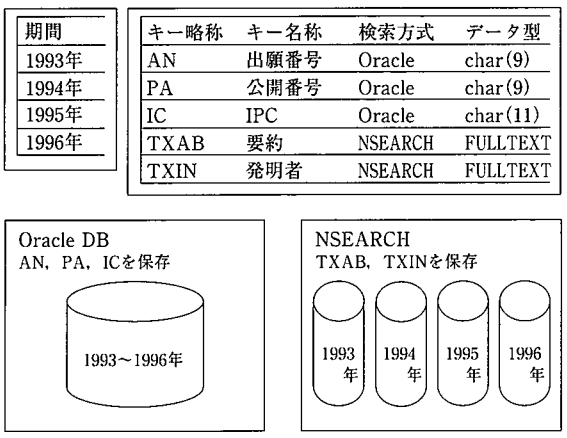

図4 データベース実装

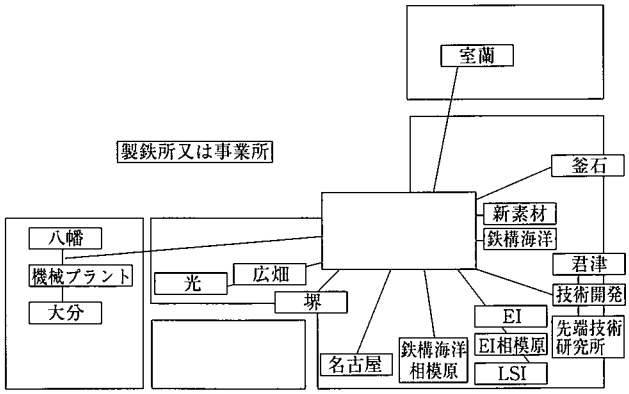

図5 新日本製鐵社内システムネットワーク構成

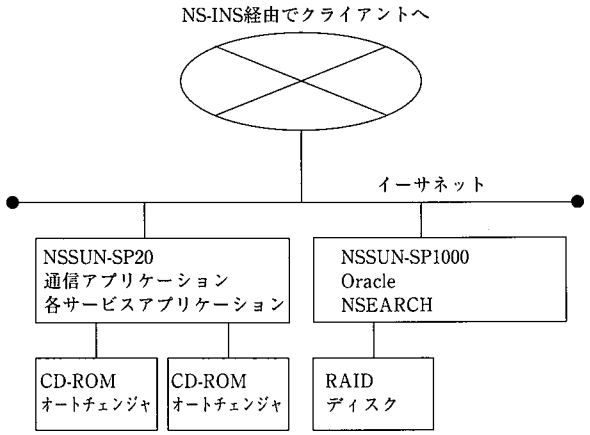

図6 サーバシステム構成

5. 結　言

新日本製鐵エレクトロニクス・情報通信事業部システム商品部における全文検索技術"NSEARCH"を基盤としたシステム構築の例として，クライアント・サーバ型特許公報データベースシステム"PSEARCH/DB"の実装及びシステム導入例を紹介した。本システムは業務パッケージとして広く社内外に販売され，現在新日本製鐵内の特許公報データベースシステムとしても運用を開始している。特許情報は蓄積期間が長期にわたるほどシステム利用効果が期待できるシステムである。現在CD-ROM公報が発行されて5年目であるが今後とも特許公報の発行形態は何らかの形で電子化された媒体で継続されるであろう。本システムを，新日本製鐵内外，各方面の協力を得て更に発展させ，また本システム構築で得られた検索システムの知見を他業務へ活かしていく所存である。