UDC 681 . 3 : 608 . 3

Full Text Retrieval for Huge Volumes of Data in Patent System "PSEARCH/DB"

Shinya NAKAMOTO*(1)

Abstract

Electronics & Information Systems Division of Nippon Steel Corp. has started marketing the patent search system "PSEARCH/DB" since 1994. The "PSEARCH/DB" is a business system for searching the information data of patents which had been stored on CD-ROMs as gazette from the Patent Office through networks using WS/PC on the open system. In the "PSEARCH/DB", we used an algorithm developed by Electronics Research Laboratory of Nippon Steel Corp. for retrieving full text in addition to the index searching method using the date, codes and so, to provide a variety of patent searching.

1. Introduction

Development technology, knowledge and know-how in manufacturing industries are assets for supporting the foundation of enterprise activities. The intellectual property rights that assure legal rights of development technology and knowledge is a crucial factor to win global competition among enterprises.

Since January 1993, the Patent Office of the Japanese Government has started distributing electronic media of the patent gazette to improve efficiency of internal distribution and enhance reusability. In January 1993 responding to the patent gazette distribution in electronic media, we, the Systems Product Department, Electronics & Information Systems Division of Nippon Steel Corp. started developing a patent search system using full text retrieval technology which we developed and started marketing as standalone type patent gazette search system "PSEARCH" in October 1993, and the client server type patent gazette database system "PSEARCH/DB" in November 1994. The full text retrieval technology used in the system employs the full text retrieval software "NSEARCH" with the high speed text searching algorithm developed by Nippon Steel Corp. in 1992.

In this paper, the patent search structure in "PSEARCH/DB" and installation of "NSEARCH" will be described, and a system introduction in the Intellectual Assets Department of Nippon Steel Corp. will be introduced as an example of applying them.

2. History of Patent Search System Development

2.1 Change in the patent gazette issue mode

The Patent Office has been issuing 450,000 published patents a year and 150,000 registered patents a year through gazette pamphlets. In order to search the patent information from this enormous number of patent gazettes, JAPIO (Japan Patent Information Organization) has constructed and operates a patent gazette search database for ordinary enterprises and patent attorney's offices.

In addition to the pamphlet issue, the Patent Office has started distributing the published patents since January 1993 and the registered patents since January 1994 on CD-ROMs after converting to SGML (standard generalized markup language) as electronic documents. SGML is composed of a text file with word information and an image file with image information, and has high volume efficiency compared with pamphlets which makes it easy to provide search display information for computers. Therefore, the distribution media compiled in CD-ROMs reduces the cost of the in-house database structure of the patent gazette.

2.2 Utilization of CD-ROM gazette for patent surveillance

In examining utilization of CD-ROM gazette, this paper assumes two types of utilization value of patent information: "rights information" and "technical information". And two types of organizing patent information "daily management" and "accumulative management". The author assumed two types utilization forms as follows.

a) Follow up of recent patent information

b) Retroactive surveillance of past patent information

a) is carried out for extracting the patent information related to the company's own technical fields from the patent gazettes. b) is used for surveying the patent information from the past. For example, the former is used for surveying the company's present business activity fields and the latter is used for technical trend surveillance of its potential business fields under study. Both applications are important key factors when making the technical strategy of an enterprise by surveying other corporations or social trends. With assumed utilization patterns as described above, the department at first commercialized the inexpensive standalone type

*(1) Electronics & Information Systems Div.

patent gazette search system "PSEARCH" for a) usage.

The retroactive surveillance system required heavy investment for constructing a database because of the conventional distribution pattern of the patent gazette and problems in the system technology described later. As conditions of construct in-house database economically become prepared because of patent publication by CD-ROM and progress of information technology, many enterprises began to examine introducing the patent gazette database. Thus, the department decided to commercialize a full-scale client server patent gazette database system "PSEARCH/DB." In the following sections, the author introduces a way commercializing of "PSEARCH/DB" (see **Fig. 1**).

2.3 Progress of IT technology

The patent gazettes are issued on about 140 CD-ROMs annually. This amounts to 70 GB data annually assuming that about 500 MB is stored on one CD-ROM. The number of applications becomes enormous, 600,000 items/year as mentioned above. The conventional technology usually applies general purpose computers and magneto-optic disks, but progress of information technology has made it possible to construct a UNIX based economical distribution system at low cost.

1) Data accumulation
   - Integral management of extensive CD-ROMs using a CD-ROM autochanger.
   - Economical large volume storage media available due to cost reduction of hard disks.
   - Data reliability assumed due to RAID structure in hard disks.
2) Data search
   - Increase in speed/volume of data search due to progress of RDB/full text retrieval technology.
   - Economical and high speed retrieval environment due to multi-CPU machines.
   - Cost reduction of search environment due to cost reduction of memory.

This environment described above has become available around 1993 and has been improving toward further cost reduction and high performance in 1997.

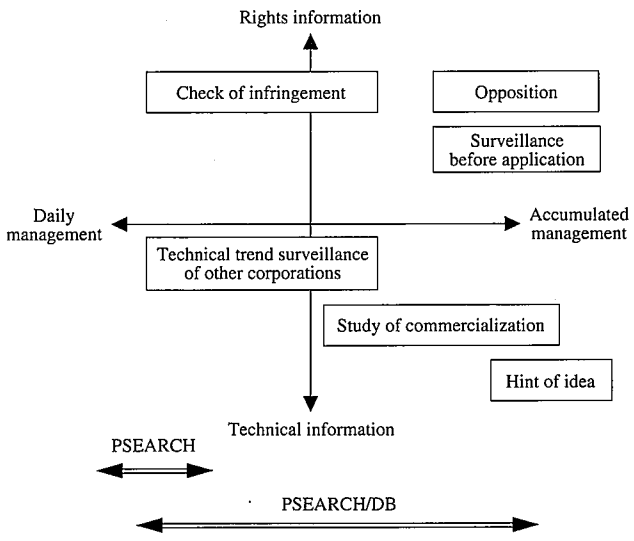

Fig. 1 Patent surveillance and system target

3. Patent Database System

3.1 Structure of patent data

The file system in CD-ROM gazette issued by Patent Office format uses in accordance with ISO 9660 based format and it can be therefore read on several OSs such as MS-DOS, Windows, and UNIX.

3.2 Required functions

Required functions can be roughly classified into four kinds as follows.

1) Retrieval function
   - Index retrieval function

   Retrieval by codes such as date, No., and patent classification.
   - Full text retrieval function

   Function to search terms among text of the patent.
   - Analytical function of retrieval equations

   Function to analyze the retrieval equations unique in the patent searching, and to delivering the results to the retrieval function mentioned above.
2) Display function of the number of surveillance results
   - Function to display the number of surveillance results
3) List display function
   - Function to display a list of surveillance results
4) Output function
   - Display function

   Function to display searched gazette contents on the screen.
   - Print function

   Function to output searched gazette contents into printer.

Application functions required by clients are can be simplified as shown above.

3.3 Operating environment

The following technologies are used as concrete operating environments assuming an open system environment by the workstation (WS)/personal computer (PC).

1) Database server
   - Hardware : NSSUN workstation
     : RAID-5 disk
     : CD-ROM autochanger
   - OS : Solaris 2.X
     : Solaris 1.X
   - Search system : Oracle (RDB:relational data base)
     : NSEARCH (full text search)
   - CD-ROM file system
     : File system for Tracer Corp.'s CD-ROM autochanger
   - Development language : C SQL yacc
   - Communications protocol : TCP/IP
2) Retrieval client (for MS-Windows)
   - Hardware : PC-AT and its compatible
   - OS : MS-Windows 3.1
   - Development language : Visual C++
   - Communications protocol : TCP/IP (use Winsock)
3) Retrieval client (for WWW)
   - Browser : Latest version of Netscape Navigator
   - Development language (for preparing cgi-bin)
     : C Perl

3.4 Basic structure

3.4.1 General

As shown in **Fig. 2**, the logical application structure consists of four tiers: 1) retrieval application, 2) communications applica-

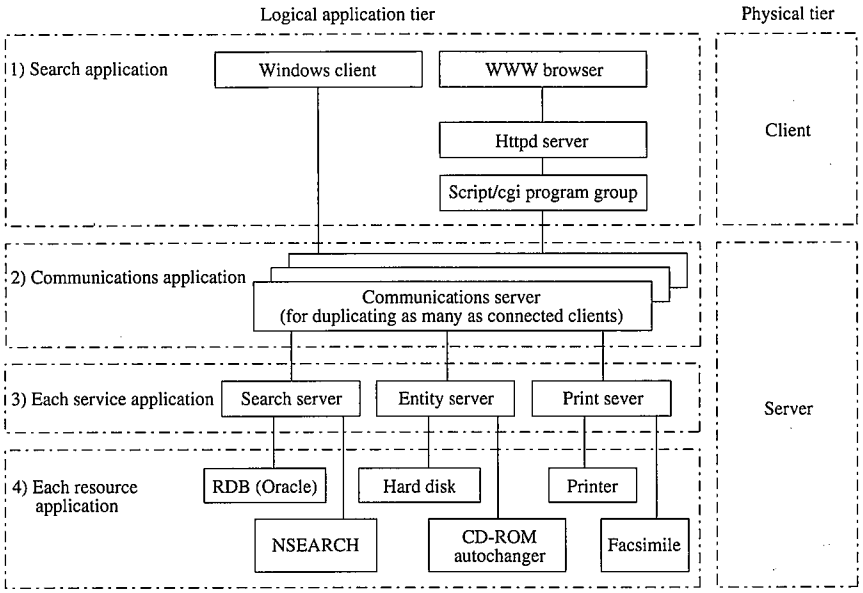

Fig. 2 Entire system configuration

tion, 3) respective service application, and 4) respective resource application. Its physical structure has two tiers: client and server. Communications are performed using a socket I/F, and communications between the retrieval server and Oracle/NSEARCH are performed via socket in accordance with SQL and NSEARCH protocol, respectively.

3.4.2 Retrieval section

The retrieval mechanism uses RDB (Oracle) and full text retrieval (NSEARCH). The patent information is usually searched by a retrieval equation, so that the above two search mechanisms should be separately used according to the search key. An example of the search equation is shown in (1).

IC = (A01K? + A01L?) × TXPA = "新日本製鐵" ····· (1)

新日本製鐵 means Nippon Steel Corp. The search keys in Equation (1) use;

IC (International Patent Classification: search data available in Oracle)

TXPA (full text search for applicant: search data available in NSEARCH)

When searching, Equation (1) is expanded into a single term search by Equations (2), (3), and (4) and into an Equation combining (2), (3), (4) and Equation (5).

IC = A01K? ····· (2)

IC = A01L? ····· (3)

TXPA = "新日本製鐵" ····· (4)

(1) = ((2) + (3)) × (4) ····· (5)

Logical sum of Equations (2) and (3) and logical product of Equation (5) are target solutions.

PSEARCH/DB employs a switch mechanism for data search using a search key. For Equations (2) and (3), surveillance is performed by Oracle, for Equation (4), by NSEARCH for full text search, and for a combination Equation with Equation (5), surveillance is performed to find a solution by the search server. This method enables an operator to prepare a search equation without being aware of differences between search mechanisms (see **Fig. 3**).

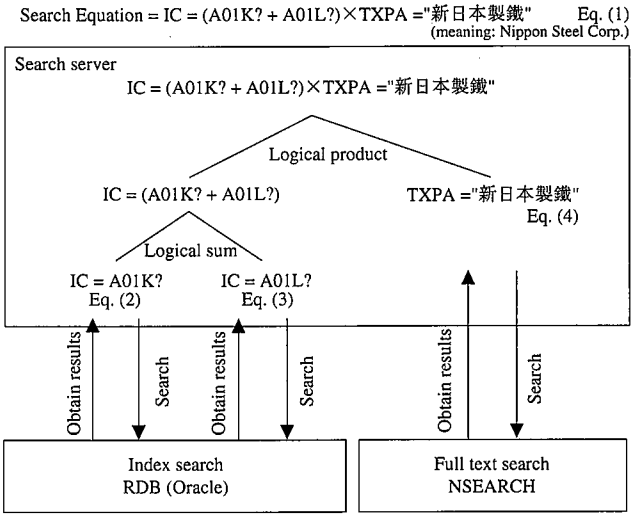

Fig. 3 Execution of search equation

3.4.3 Display data

When the autochanger is used, it takes 40 seconds to extract data at maximum for replacing CD-ROM. This is the time required to display based on the indication order by the client. PSEARCH/DB employs a method to enhance initial response time by making a hard disk hold a part of the display data.

3.5 Installation

The installation above mentioned was carried out on the server. The database structure is composed of the retrieval data retrieving RDB/full text, and the display data. A definition file for the structure information is prepared and each retrieval/display section operates based on information in the definition file. The full text retrieval file cannot be put into one file due to constraints of the OS that one file has to be under 2 GB. The system is installed, therefore, to retrieve the dividing files, and make plural full text retrieval files, correspond to a RDB instance. **Fig. 4** shows an example of the files divided according to the retrieval period.

Period
1993
1994
1995
1996

Key abbreviation	Key name	Search method	Data type
AN	Application number	Oracle	char (9)
PA	Publication number	Oracle	char (9)
IC	IPC	Oracle	char (11)
TXAB	Abstract	NSEARCH	full text
TXIN	Inventor	NSEARCH	full text

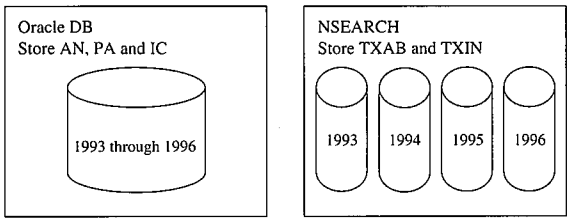

Fig. 4 Database installation

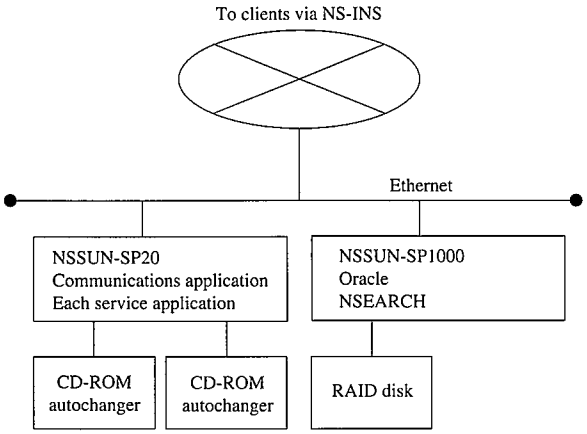

Fig. 6 Server system structure

4. Example of Patent Search System Introduction

PSEARCH/DB has been introduced into various ways since its sales in 1994. Among them, this paper describes an example of introduction in the Intellectual Assets Department of Nippon Steel Corp., which is the largest case as of September 1997.

4.1 Network structure

A centralized network management system was constructed placing a server at the center. The server is accessed from each steelworks via TCP/IP network by existing NS-INS (in-house specialized network) (see **Fig. 5**).

4.2 Database arrangement

The database server was constructed of the following devices.

- CPU : NSSUN-SP20
  : NSSUN-SP1000 4CPU Model
- Disk : NSSUN CL2000 4GB Disk × 20
- OS : Solaris 1.2 (NSSUN-SP20)
  : Solaris 2.5.1 (NSSUN-SP1000)
- CD-ROM autochanger
  : Pioneer DRM 5004X (4 Drive Model) × 2 sets
- Retrieval system : Oracle V7.2.2
  : NSEARCH V3

Fig. 6 shows the system configuration.

As of September 1, 1997, the gazettes shown in **Table 1** are registered for operation.

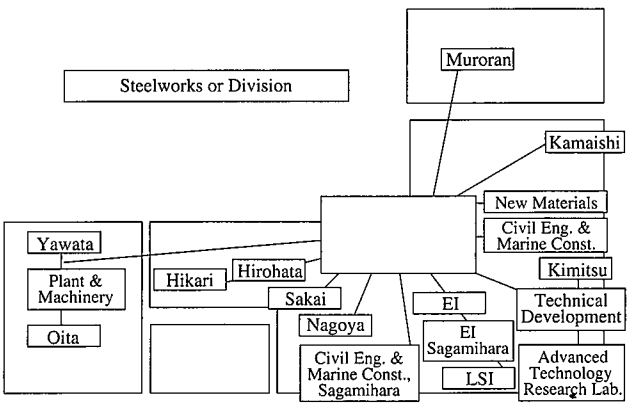

Fig. 5 In-house system network structure in Nippon Steel Corp.

Table 1 Kind and number of registered gazettes

Gazette kind	Stored period	Number of items (round numbers)
Opened publication gazette	from January 1993	1,934,000 items
Registered utility model gazette	from July 1993	40,000 items
Announced gazette	from January 1996	21,000 items
Reannouncement	from January 1996	2,000 items
Publication gazette	from January 1994	388,000 items
Registered gazette	from May 1996	195,000 items
Total		2,580,000 items

5. Conclusions

In this paper, the auther described the introduction of client server patent gazette database system "PSEARCH/DB" as an example of the system structure based on the full text retrieval technology "NSEARCH" developed by Systems Product Department, Electronics & Information Systems Division of Nippon Steel Corp. This system has been widely distributed inside/outside the corporation as a business package, and has started operating as the patent gazette database system inside Nippon Steel Corp. The patent is expected to provide utilization benefits as the data accumulation period becomes longer. It is now the fifth year since the patent gazette began to be issued on CD-ROMs, and it will continue to be issued in electronic media in some way. The author will further enhance this system in cooperation with others inside/outside Nippon Steel Corp. to apply experience and knowledge from this system structure to other fields.