情 報 学 基 礎 34－4
（1994．9．14）

LAN上に分散配置する学術論文全文データベースシステム

阪口哲男　杉本重雄　田畑孝一

図書館情報大学

本論文では、著者らが開発したLAN上に分散配置する学術論文全文データベースシステムについて述べる。分散環境はLANにつながれたワークステーション群から構成される。一般に分散環境を指向した情報システムでは、データベースをワークステーションに分散し、ユーザに適切なデータベースを選択させる。本システムでは各データベースサーバに自動的に検索要求を配布するため、ユーザはデータベースを選択する必要がない。学術論文のテキストと共に図表もデータベースに格納する。データベースに格納されているテキストはすべて索引付けされる。本システムでは論文の検索機能と提示機能が統合されている。本システムはUNIXワークステーション上で開発した。

A Full-text Database System Distributed in LAN Environment

Tetsuo Sakaguchi, Shigeo Sugimoto, Koichi Tabata

University of Library and Information Science

This paper describes a full-text database system distributed in LAN environment. A distributed environment consists of workstations connected to a LAN. There are a number of information systems oriented to distributed environment. Databases of those systems are distributed to the workstations and users must choose an appropriate database from them. This system automatically distributes search request among database servers. So, users of this system must not choose databases. Texts and graphs of articles are stored in the databases of this system. All texts in the databases are indexed for full-text searching. Searching databases and displaying retrieved articles are integrated into this system. This system is implemented on UNIX workstation.

1. はじめに

近年、ワークステーションとLANで構成される分散環境が増えつつある。システムの効率の観点から、そのような環境ではデータベースを複数のワークステーションに分散配置することが望ましい。データベースを分散配置した場合、それらのデータベースを検索するための分散型データベースシステムが必要となる。これまでに様々な分散型データベースシステムが企業や大学によって開発されている。

大学における研究室でデータベースを備え、学術論文を格納する場合を想定してみる。各研究室では全文データベースを格納するためのワークステーションを設置し、研究室の研究分野に沿った内容のデータベースを構築する。それらのデータベースを互いに共有すればより有用になる。そのような環境を実現するには、データベースを分散配置することと、LANにつながれたワークステーションからデータベースを利用するためのいくつかの要件が必要である。

本論文ではLAN上に分散配置する学術論文全文データベースシステムとその実現について述べる。本システムは前述のような観点に基づいて設計しており、LAN上に分散配置された全文データベースから学術論文を検索し、提供する機能を備える。また、検索などの操作を容易にするためにユーザフレンドリなインタフェースを備える。

2. 分散環境を指向した全文データベースシステム

2.1 分散環境における全文データベースシステム

分散環境はネットワークを通じて互いにつながれたワークステーション群より構成される。そのネットワークには構内ネットワーク(LAN)、広域ネットワーク(WAN)、そしてインターネットのようなネットワークを相互に接続したインターネットワークがある。分散環境では、一つの仕事を各ワークステーションに分配し、より効率を良くすることができる。いくつかに分けられた仕事の各部分はネットワークにつながれた他のワークステーションより利用することが可能となる。

CD-ROMや磁気ディスクなどの記憶装置の進歩によって、比較的安価に全文データベースをワークステーション上に構築することができるようになった。そのため、分散環境においても全文データベースを容易に構築することができる。単一のワークステーションにすべてのデータを格納している場合、全文データベースのようにデータ量が多くなると検索速度は遅くなる。データを分散して格納し、並列に検索することによって検索速度を速くすることが可能となる。

各ワークステーション上の全文データベースシステムが互いに独立していると、ユーザは自分にとって適切なデータベースを格納しているワークステーションを探して利用しなければならない。このデータベースを探すことはユーザにとって余分な手間になる。そのため、多くの分散環境上の全文データベースシステムでは、データベースを探すための何らかの機構を備えている。

2.2 ネットワークを指向した全文データベースシステムの事例

分散型全文データベースシステムの典型的なものにLIS-IIとWAISがある。以下、それぞれのシステムについて述べる。

(1) LIS-II[1]

LIS-II (Library Information System II)はカーネギーメロン大学のLAN上で稼働している全文データベースの検索システムである。LIS-IIには2種類のデータベースがある。ひとつは学術論文のページ毎のイメージを

格納したデータベースであり、もうひとつはINSPEC抄録データベースである。分散されたデータベースの位置情報をユーザに提供するDatabase Name Serviceと呼ばれる機能を備えている。LIS-IIのユーザはまずDatabase Name Serviceから目的のデータベースの位置情報を得て、次にそのデータベースから論文を検索する。ユーザは検索結果として論文のページ毎のイメージを得る。

(2) WAIS[2][3]

WAIS (Wide Area Information Servers)はインターネット上の世界的規模の分散型データベースシステムである。WAISはインターネット上に分散している多くの情報サーバとユーザが使用するクライアントから構成される。通常のサーバにはテキストや画像などが格納される。サーバの中にdirectory of serversと呼ばれる特別なサーバが存在し、これには他のサーバの所在情報が納められる。WAISではサーバとクライアントの通信にANSI Z39.50[4]に基づいたプロトコルを用いている。

WAISではユーザは次のような手順でテキストや画像を検索する。(1)適切なサーバを探すためにdirectory of serversに手がかりとなる用語を与える。(2)その検索結果から適切なサーバを選択する。(3)絞り込んだ用語を用いて選択したサーバで検索を行う。(4)その結果からテキストや画像を選択して表示させる。

このようにユーザは世界各地に広がったサーバから自分が求めるテキストや画像を得ることができる。

2.3 LAN上に分散配置する学術論文全文データベースシステム

この節ではLAN上に分散配置する学術論文全文データベースシステムの要件と設計上の留意点について論じる。

前節で論じたように、LIS-IIとWAISでは検索手順が2段階となっている。第1段階はデータベースを探し出すもので、第2段階はそのデータベースから論文を探し出すものである。このデータベースを探し出す段階をユーザに意識させないで済めば、より使いやすいシステムとすることができる。データベース選出段階を隠すためには二通りの方法がある。ひとつは論文をすべてのデータベースで検索する方法であり、もうひとつは同じ用語をデータベースと論文の両方の検索キーに用いる方法である。前者は再現率が高くなるものの、検索時間と計算機の負荷がかかる。後者は負荷は比較的軽く済むが、再現率と適合率を高めるためにデータベースの所在情報を入念に管理する必要がある。本システムでは前者の方法を用いることにする。

すべてのデータベースを検索するには、LANにつながれているすべての利用可能なデータベースサーバを把握する必要がある。利用可能なサーバを登録するために所在情報サーバを準備する。所在情報サーバには利用可能なサーバの表を格納する。この表はデータベースサーバと通信を行った結果に基づいて更新される(図1)。

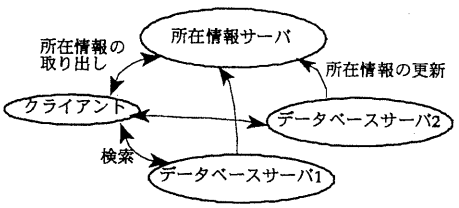

図 1 所在情報の提供

研究者にとって論文の本文と図表を同時に見ることは重要である。従って、本システムでも同一ディスプレイ上にテキストと図表を表示する。LIS-IIでは論文はページ毎の画像として表示されるので、本文と図表を印刷物と同じ配置で見ることができる。この手法ではディスプレイ装置の解像度の

表 1　各システムの特徴

	本システム	LIS-II	WAIS
データベースの内容	全文 （本文、抄録等）	本文、抄録、書誌データ	全文
検索の範囲	全文	抄録	全文
データベースの分散範囲	local	local	global
ユーザの範囲	local	local	global
本文のタイプ	文字列	ページ画像	文字列
図表のタイプ	ビットマップ画像	ページ画像	
同時検索データベース数	1以上	1	1

影響を受けるほか、ユーザは本文中に含まれる用語を目で探さなければならない。一方、現行のWAISではテキストと図表は別々に扱われている。本システムではページ毎の画像は扱わずに、テキストと図表を相互にリンクして格納する。マルチウィンドウ環境を用いることで、これらのテキストと図表を同時に表示することができる。

本システムでは検索条件を自由に指定できるようにするために論理式を取り入れる。検索結果の集合演算もその論理式と同時に実現する。

2.4 事例システムとの比較

LIS-II、WAIS、本システムの特徴を表1にまとめる。

LIS-IIはページ毎の画像のデータベースと検索用に索引付けされた抄録データベースを備える。本システムとWAISでは論文の抄録のみならず全文について索引付けを行う。論文中のテキスト部分と図表部分はLIS-IIと本システムでは論文毎にまとめられているが、WAISでは別々に扱うことになる。

WAISの情報サーバはインターネット全体に分散している。LIS-IIと本システムのデータベースサーバはLAN上でのみ分散配置されている。本システムでは検索要求を各データベースサーバに同時に分配している。この方法はネットワークの遅延時間の影響を受けるため、インターネットで分散される場合には適用できない。

LIS-IIとWAISは2以上のデータベースを逐次に検索するが、本システムでは並列して検索を行う。データベースの数が増えるに従って逐次に検索する時間は増加する。そのため本システムでは検索するデータベースの数が増えるほど検索時間の点で有利になるが、サーバとなるワークステーションの平均的な負荷は増加する。

3. LAN上に分散配置する学術論文全文データベースシステム

3.1 システムの概要

LAN上に分散配置する学術論文全文データベースシステムは複数のデータベースサーバ、所在情報サーバ、検索クライアントから構成される(図2)。データベースサーバはLANにつながれたワークステーションに分散して配置される。各々のデータベー

スサーバには全文データベースが格納される。全文データベースには格納されている論文の主題領域を表す名前が付けられる。データベースに格納されているテキストは全文検索のために索引付けされ、画像はテキストとリンク付けされる。所在情報サーバは利用可能なデータベースサーバの所在情報を管理する。検索クライアントはユーザと対話し、データベースサーバをアクセスする。

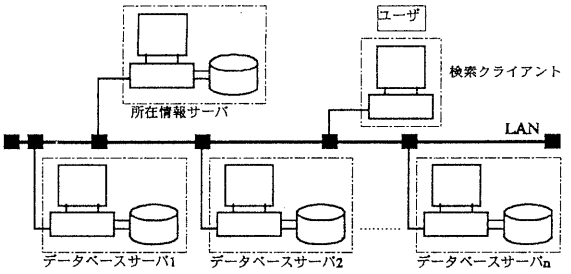

図 2　システムの概観

ユーザが検索クライアントに用語を与えてそれを含む論文を検索するように要求すると、検索クライアントはまず利用可能なすべてのデータベースサーバの所在情報を所在情報サーバから得る。次に指定された用語を含む検索要求を各々のデータベースサーバに送る。データベースサーバは検索が終わると結果集合を検索クライアントに送るので、検索クライアントはそれをユーザに提示する。ユーザがその結果の中から論文を選ぶと検索クライアントはその論文の内容を転送するようにデータベースサーバに要求し、転送されてきた内容を表示する。

3.2 データベース

データベースサーバ上の全文データベースには論文のテキストと図表が格納される。論文のテキストは、タイトル、著者名、抄録、本文、参考文献、付録、書誌データの7つに分けられて格納される。これらはすべて文字単位で索引付けされる。論文中の図表はテキストとリンク付けされ、ビットマップ画像として格納される。全文データベースには格納されている論文の主題領域を表した名前が付けられる。

所在情報データベースは所在情報サーバによって管理される。所在情報データベースには利用可能なデータベースサーバ上にある全文データベースの所在情報が格納される。所在情報はデータベースを特定するために、データベースの名前とそのデータベースを提供しているサーバのネットワークアドレスまたはホスト名の組で表される。個々の所在情報はデータベースサーバからの要求によって更新される。

3.3 データベースサーバと所在情報

データベースサーバは論文の検索と転送の機能を提供する。データベースサーバに対して送られる要求には大きく分けて検索、集合演算、論文転送の3種類がある。

検索要求には検索の条件を表す式が含まれる。条件を表す基本式は用語と範囲を指定する論文の部分名である。論文の部分名はタイトル、著者名、抄録、本文、参考文献、付録、書誌データのいずれかであり、省略可能である。省略した際は前述のすべて、すなわち全文を範囲として指定したことになる。データベースサーバはこの指定された用語を指定された範囲に含む論文を探し出す。条件を表す複合式は基本式あるいは複合式をAND、OR、またはNOTの論理演算子で連結したものである。各演算子の意味は以下のようになる。

A AND B: Aを含みかつBを含む論文

A OR B:　Aを含むかまたはBを含む論文

A NOT B: Aを含むがBを含まない論文

検索の結果としては探し出された論文に関する情報の集合が返される。その情報は論

文毎のその用語の出現数、大きさ、論文のタイトルから構成される。

集合演算の要求は事前に行われた検索結果に対して集合操作を行うことを意味する。集合演算の結果は検索要求の結果と同様の形式である。集合演算にはAND、OR、NOTがある。ANDとORはそれぞれ集合の積と和を求める。NOTはand-not、つまり一方の補集合との積を求める。

転送の要求はそれまでに検索した結果に含まれる論文を検索クライアントに転送することを意味する。その際、転送すべき論文を特定するために検索結果に含まれている論文の識別子を指定する。この要求を受け取ったデータベースサーバはその論文をクライアントに送る。

データベースサーバはそのデータベースの所在情報を定期的に所在情報サーバに送る。ある一定時間以上所在情報が送られてこなかった場合、所在情報サーバはそのデータベースサーバが停止したと判断し、その所在情報を所在情報データベースから削除する。新しいデータベースが追加されたときは、その所在情報が所在情報サーバに送られ、所在情報データベースに追加される。

3.4 検索クライアント

本システムにおける検索クライアントの役割は、ユーザとの対話とデータベースサーバのアクセスである。検索クライアントはユーザの操作を解釈し、サーバへ送る要求を作成する。また、サーバから要求に対する返答を受け取り、ユーザに見やすい形式に変換して提示する。

ユーザインタフェースをより使いやすくするためにGUIを採用する。GUIを用いればユーザはウィンドウ上のボタンやメニュを操作したり、検索結果中の論文をマウスを用いて選ぶことができる。すなわち、ユーザはコマンドを覚えたり検索結果の識別子をメモにとったりする必要がなくなる。

4. システムの実現

4.1 システムの構成

本システムはイーサネットに接続されたUNIXワークステーション上で開発した。開発環境を以下に示す。

ハードウェア:	ソニーRISC NEWS 3000シリーズ
OS:	NEWS-OS4.2.1R
プログラミング言語:	C
GUIツールキット:	X11R5, OSF/Motif Widget
検索ライブラリ:	ソニーMediaFinder (DPS, SDK)

本システムではMediaFinder[5]の全文検索機能を利用している。ネットワーク機能についてはUNIX BSDソケットライブラリを用いて実現する。

本システムの構成を図3に示す。データベースサーバ、所在情報サーバ、検索クライアントはワークステーション上で機能する。どのサーバやクライアントも同じワークステーション上で動作させることが可能である。サーバとクライアントはLANを通じて通信する。各構成要素の詳細については次節以降に述べる。

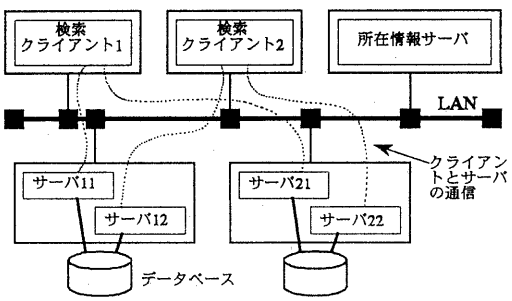

図 3 システムの構成とサーバ・クライアント間通信

4.2 データベースサーバ

個々のデータベースサーバはそれぞれUNIXのプロセスとして実現する。検索結果となる論文集合を保持するため、サーバのプロセスは各検索クライアント毎に生成する。サーバの検索部はMediaFinder SDKを用いて実現する。そのため、全文データベースはMediaFinder DPSで索引付けする。 MediaFinder SDKは検索結果としてメモリアドレスを返すので、本サーバではそれらを識別子に変換してクライアントとの通信に用いる。

各クライアントに対して新たにサーバプロセスを生成するため、UNIXのインターネットスーパデーモンと似た働きをするマスタサーバプロセスを準備する。このマスタサーバプロセスは所在情報サーバに所在情報を送る。

4.3 所在情報サーバ

所在情報サーバもデータベースサーバと同様にUNIXのプロセスとして実現する。本システム中には唯一の所在情報サーバが存在する。所在情報サーバはシステム管理者の操作あるいはワークステーションの自動リブート手続きによって起動される。

データベースサーバから送られる個々の所在情報が内部のデータベースに格納される。検索クライアントから要求がくれば、内部データベース中の所在情報を送る。

4.4 検索クライアント

検索クライアントはMotif Widgetを用いたGUIを備える。検索クライアントはGUI部と通信部に分けられる。GUI部はユーザと対話し、サーバに対する要求を作り出す。通信部はデータベースサーバと所在情報サーバとの通信を担う。GUI部によって生成された要求をサーバとのプロトコルに従って変換し、サーバへ送る。そしてサーバから返された結果を変換しGUI部に渡す。GUI部は渡された結果をウィンドウ上に表示する。

4.5 動作例

ここでは例によって本システムを説明する。図4にディスプレイ全体のイメージを示す。"gui"という名のウィンドウが４つある。それらのうち右上のウィンドウには利用可能なデータベース名がすべて表示されている。左上のウィンドウは検索条件を入力するためのものである。左中程にあるウィンドウは検索結果の論文集合を表示している。右下のウィンドウは検索結果より選び出した論文を表示している。"his"という名のウィンドウは検索履歴を表示している。各行はそれぞれ論文集合を意味する。"sjx"というウィンドウは条件入力の際に使用するソニーの仮名漢字変換システムである。

検索条件は図5のように指定する。ここで反転表示されているものはそれぞれの用語を探す際の範囲指定である。範囲指定や"AND"のような論理演算子はウィンドウに備えられたボタンによって入力される。

検索結果が表示されているウィンドウにおいて、各行は論文を表している。いずれかの行がマウスで選択されて表示を指示するボタンが押されると、新たなウィンドウが出現して指定された論文が表示される。

5. おわりに

これまでに述べたように、本システムは次のような特徴を備えている。

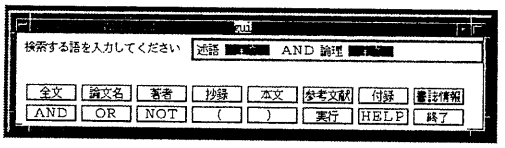

図 5 検索条件入力ウィンドウ

(1) ユーザはデータベースの物理的所在を知る必要がない。
(2) 論文のテキストと図表の両方がユーザに提供される。
(3) データベースの検索と論文の表示がひとつのシステムに統合されている。

すべてのサーバに検索要求を配ることは全体の負荷を高くするが、単一のデータベースで構成する場合に比べて各サーバが備えるデータベースの大きさが小さくなるので、個々のサーバの負荷は高くはならない。この手法はインターネットのような広域ネットワークでは通信における遅延時間が長くなるため適用できないが、LANでは実用に耐えると考えられる。

現在実現しているシステムはデータベースサーバと検索クライアントの通信に独自のプロトコルを用いている。これをANSI Z39.50のような標準化されつつあるプロトコルを用いることで、例えばWAISなどの他の情報システムをも利用できる統合環境を備えたシステムを実現することができると考えられる。

参考文献

[1] Troll, D. Information Technologies at Carnegie Mellon. Library Administration & Management. Vol.6, No.1. p.91-99 (1992)

[2] Krol, Ed. "Searching Indexed Databases: WAIS" , The Whole Internet User's Guide & Catalog. O'Reilly & Associates. p.211-226 (1992)

[3] Kahle, B. Wide Area Information Server Concepts. Version 4, Draft, 1989 ("wais-concepts.txt" distributed to the Internet)

[4] ANSI Z39.50 Version 2 Third Draft, 1991.

[5] MediaFinder SDKユーザズマニュアル, ソニー, 1991

[6] Philips, G, L. Z39.50 and the Scholar's Workstation Concept. Information Technology and Libraries. Vol.11, No.3. p.261-270 (1992)

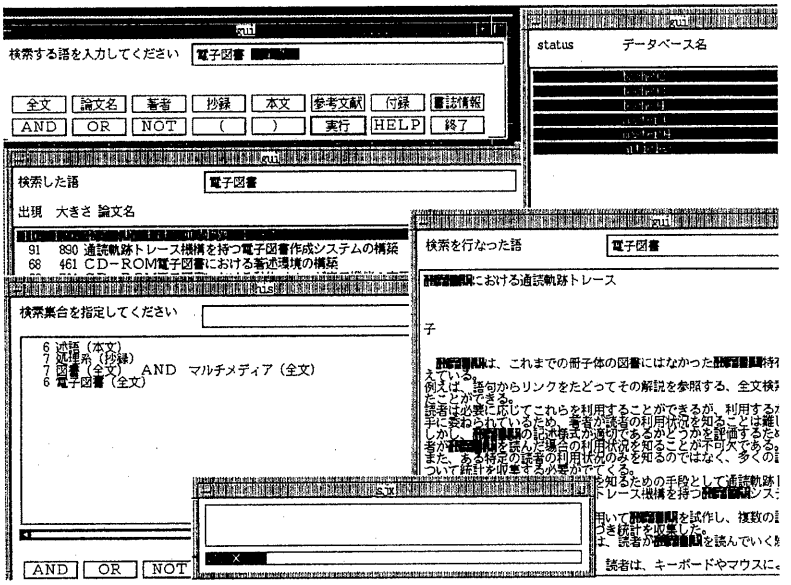

図 4 システムの画面全体イメージ