

分散型の学術論文全文データベース

1C-7

芳賀則久 阪口哲男 杉本重雄 田畑孝一

図書館情報大学

1. はじめに

近年、LANに接続されたワークステーションによる分散環境が広まりつつある。我々はこのような分散環境上で、分散型の学術論文全文データベースの構築を進めており、本稿ではその構成と構築手法について述べる。

2. 学術論文の全文データベース

本システムでは、学術論文の全文を図表も含めてデータベースに格納する。格納する学術論文の構成要素を論文名、著者、抄録、本文、図表、参考文献、付録に定め、検索や取り出しの単位とする。図表は画像情報とし、その他の要素は文字情報として格納する。

本システムでは、文字情報として格納されたすべての要素を対象に、利用者が指定した任意の語が出現している学術論文を検索する。

3. データベースの分散配置とその検索

3.1 データベースの分散配置

本システムでは複数の全文データベースを、LANに接続された複数のホスト(ワークステーション)に配置する。利用者は手元のホストからこれらのデータベースを同時に検索する。利用者はデータベースの物理的配置を意識せずに、全てのデータベースまたは名前によって指定したデータベースにおいて論文を検索することができる。

3.2 データベースの検索

本システムでは、利用者が指定した任意の語を文中に含む論文を探す。その際、語が含まれる範囲を論文全体もしくは構成要素によって指定することができる。複数の語を組み合わせる際は、A and B (AとBの両方を含む)、A or B (AとBのどちらかを含む)、A not B (Aを含むがBを含まない)の3種の基本式を組み合わせ指定する。検索結果は

条件に適合する論文の集合となる。システムの利用開始時からのすべての検索結果を保持しているため、その間で集合演算が可能である。

検索結果は、各論文毎に論文名、著者名、本文の長さ、語の出現回数が表示され、論文は出現回数の降順で並べられる。語の出現回数は、検索条件に指定した語の各々の出現回数の合計である。

3.3 分散配置されたデータベースの検索

利用者からはデータベースの物理的配置は見えず、単にシステムが複数のデータベースを取り扱うように見える。全データベースまたは一部のデータベースを対象に指定して検索することができる。対象データベースへの検索要求の分配、検索結果の統合はシステム内部で行われる。

3.4 ユーザインタフェース

本システムはグラフィカルユーザインタフェース(GUI)を備える。検索条件の指定や結果表示の操作などはメニューやボタンによって行う。

検索を複数回行った場合の検索結果は検索履歴として一覧され、任意の結果を参照することができる。結果を参照する際は、それに含まれる論文の一覧が表示される。論文を読む場合は、その一覧の中から論文を選択し、表示ボタンを押すと論文が表示される。検索条件の語の出現位置、論文の先頭、抄録、本文、図表、参考文献、付録のそれぞれの表示ボタンを用意する。

論文表示の際、図表以外の要素に関しては、検索条件で指定された語の部分を色反転して表示し、利用者がその語を見つけるのを容易にする。スクロールやページめくり、ページジャンプにより論文の任意の場所を表示することができる。論文の本文中で図表を参照している部分(図の番号など)も色反転して表示し、これを選択することで参照されている図表を表示することができる。論文は表示するだけでなく、利用者の手元のホストのファイルに保存することも可能である。

A Distributed Full-text Database System for
Retrieving Academic Articles

Norihisa HAGA, Tetsuo SAKAGUCHI, Shigeo
SUGIMOTO, Koichi TABATA

University of Library and Information Science

4. 他システムとの比較

学術論文を対象とする代表的な全文データベースシステムに、カーネギーメロン大学(CMU)のLIS-II[1]とThinking Machines社が中心に開発したWAIS[2]がある。LIS-IIはCMU内部で利用されており、学内ネットワーク上に分散配置されたデータベースとその所在情報を提供するDatabase Name Serviceからなる。検索はこのDatabase Name Serviceから得た所在情報に基づいて特定のデータベースを対象に行う。データベースには論文をページ毎に画像情報として格納しており、検索のためにキーワードが付与されている。

WAISではInternet上にデータベースが分散配置されており、そのデータベースの概要と所在情報を提供する特別なデータベースであるDirectory of Serversがある。検索はまずDirectory of Serversにおいて任意の語を指定し、目的に合うデータベースの所在情報を取り出す。そして、そのデータベースを対象に論文の検索を行う。一般にWAISでは論文は文字情報で格納されており、その全文を対象に索引付けされている。

本システムとLIS-II、WAISを比較すると表1のようになる。本システムは検索する際の対象データベースは複数であり、それらを同時に検索する。一方、LIS-II、WAISでは対象データベースは通常一つであり、WAISでは複数指定することも可能であるが逐次的に検索を行っている。

5. システム構成

本システムは、データベースの統合と利用者との対話を行うクライアントと、データベースの検索を行うサーバから構成される(図1)。利用者一人と対話するクライアントは一つであり、クライアントはデータベース数に応じた数のサーバへ検索要求を分配し、その結果を統合して利用者に提

表1 各データベースシステムの比較

	検索	利用範囲	データ形式		同時対象データベース数
			本文	図表	
本システム	全文	local	文字	画像	複数
LIS-II	キーワード	local	画像	画像	1
WAIS	全文	wide	文字	-	1

示する。クライアントと各サーバはANSIのZ39.50[3]に基づくプロトコルを用いて通信する。システムはUNIXワークステーション上に構築し、データベースは磁気ディスクまたはCD-ROMに格納する。

クライアントはひとつのUNIXプロセスとして実現し、GUIはOSF/Motifに基づいて構築する。

サーバは個々のデータベースの検索を行うUNIXプロセスとして実現する。利用者毎の検索の履歴を保持するために、データベース毎にクライアントの数だけサーバが起動される。全文検索の基本機能はMediaFinder[4]に基づいて構築する。

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 Catalog & User's Guide. O'Reilly & Associates. p.211-226 (1992)
- [3] ANSI Z39.50 Version 2 THIRD DRAFT, 1991.
- [4] MediaFinderSDK使用説明書, ソニー, 1991.

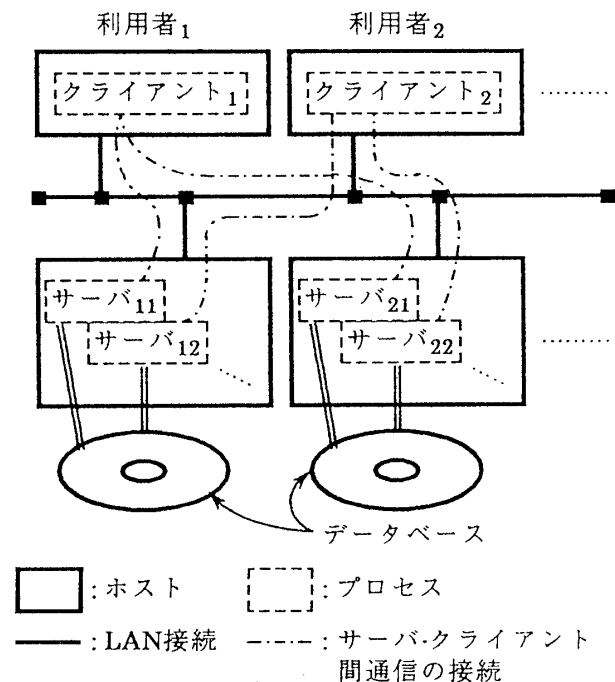


図1 システムの構成