2M−8

重なった対象領域を持つ複数データベースに対する日本語横断検索システム

谷 幹也　久保 加奈子　市山 俊治

NEC関西C&C研究所

1 はじめに

重なった対象領域を持つ複数の文献データベースに対して情報検索を実行する場合、データベースの適切な選択とそれぞれのデータベースに固有の検索式の生成を行なう必要がある．本報告では、1) 各データベースに含まれる対象領域毎の件数、詳細度、網羅性、言語、収録年度などからなるデータベース特徴知識を用いて、日本語質問文の意味構造からデータベース及び検索方針を選択する．2) 選択されたデータベースのそれぞれに対して、概念マッピング情報からなる対象領域知識を用いて検索式を生成し、検索結果の統合処理を行なう．の2つを特徴とする日本語横断検索システムを提案する．現在当社で開発を進めている自然言語インタフェース構築キット"IF-Kit"[谷 91a]を利用して評価を行なっている．

2 対象領域の重なるデータベース

主題分野が等しいが、その分野に関する網羅性がa) 主題領域,b) 情報源タイプ, c) 収録時期,d) 収録雑誌,e) 収録件数,f) 収録言語,g) 収録項目などの点で異なるデータベースは多数存在する．本報告書では、主題領域が「情報学」の文献検索において、対象データベースの網羅性がa) ～ f) の点で異なる図1. のデータベース群を元に説明を行なう．

	データベースa	データベースb	データベースc	データベースd	データベースe	データベースf
主題領域	情報学	情報学	情報学	情報学	情報学	情報学
情報源タイプ	図書	特許	雑誌論文	雑誌論文	雑誌論文	雑誌論文
収録時期	1940～現在	1980～現在	1975～現在	1975～現在	1945～現在	1975～現在
収録雑誌	-	-	A, B	C, D	B, D	A, E
収録件数	1万	4万	5千	7千	3万	2万
収録言語	日本語	日本語	日本語	日本語	日本語	日本語
収録項目	書誌事項	書誌事項	書誌事項+抄録	書誌事項+抄録	書誌事項+抄録	書誌事項+抄録

図1. データベース特徴情報

2.1 サーチャによる検索の現状

ユーザが情報検索の専門家(サーチャ)に依頼して検索を行なう場合、サーチャはユーザから寄せられた検索要求に対して、次のような手順で検索式の作成を行なっている．

1) 主題分析
2) 概念分析
3) 検索語の選択
4) 検索式の生成

この「主題分析」の中で検索要求の中から取り出した主題に近いデータベースを選択するわけだが、この時データベースを選択する基準として、前章のa) ～ f) の基準及び以前の検索経験から網羅性とコストをつり合わせるように、データベースの選択を行なう．

2.2 狙い

本システムは、サーチャが行なっているデータベースの選択基準及び選択ルールを組み込むことによって、対象領域の重なっているデータベース群に対して適切なデータベースの選択を行ない、それぞれのデータベースに適応した知識を利用することで自動的に検索式の生成を実行するシステムである．以下に実際の処理の流れを説明する．

3 システムの構成

対象領域の重なった複数のデータベースの横断検索は図2. のように行なう．

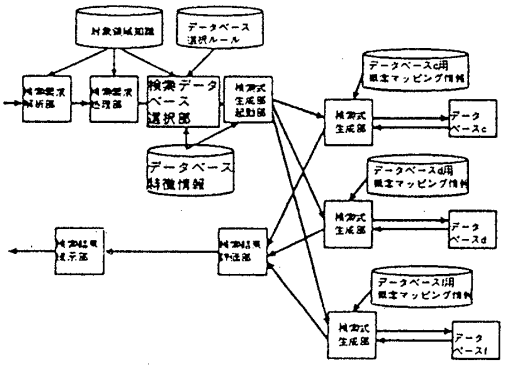

図2. 処理のながれ

処理の流れを例を用いて説明する．検索要求として、「情報検索に関する雑誌論文をもれなく示せ。」が与えられた場合、以下の1から6の手順により検索を行なう．

1. 検索要求解析部

日本語で入力された検索要求を辞書、解析ルールを用いて解析し、図3. の概念構造を作成する．

The Japanese Information Retrival System for cross-domain databases. Mikiya Tani , Kanako Kubo and Shunji Ichiyama
Kansai C&C Research Lab., NEC Corp.

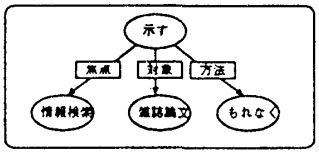

図 3. 中間構造

2. 検索要求処理部

作成された概念構造から検索式を生成するのに必要な情報と DB 選択、結果評価に必要な意図情報を表 2. の情報を利用して切り出す. ここでは「全件検索」が意図情報となる.

表 2. 意図情報

意図	クルーワード
全件検索	全部, すべて, 洗いざらい, もれなく,...
件数指定	X 件, 一つ,2,3...
時間指定	早く, すぐ...

3. 検索 DB 選択部

作成された概念構造と意図情報から図 4. で表すようなネットワーク型のデータベース特徴知識及びルールを利用して、DB の選択を行なうと共に、結果評価部へ意図情報を伝達する.

この際、概念構造の対象となっている「情報検索」と階層関係にある概念「情報管理」「文献検索」「DB 検索」を扱っている雑誌は A,B,C,E の 4 つであり、図 1. のデータベースの特徴情報から、なるべく少ないデータベースで A,B,C,E の 4 つの雑誌を網羅できるようにデータベースを選択し、(c,d,f) のデータベースについて検索式生成部を起動する.

4. 検索式生成部起動部

選択されたデータベースの組と概念構造から、それぞれのデータベースに対して検索式を生成する知識、ルールを持った検索式生成部を立ち上げ、概念構造を伝達する.

5. 検索式生成部 [谷 91a]

それぞれのデータベースに対応した検索式を生成し、検索を実行して、検索結果を結果評価部へ伝達する.

6. 結果評価部

意図情報から検索結果待ちの是非を決定する. 意図「全件検索」であるため、表 3. の基準から全ての検索の結果が返ってくるのを待ち、検索結果を比較して同じものを除去し、ユーザへ提示する.

表 3. 検索結果処理基準

意図	動作
全件検索	全ての検索結果を待ち、重複分は融合
件数指定	(指定件数< α の場合) 最初に返ってきた検索の中で条件を満たせば出力 (指定件数> α の場合) 全ての検索結果が入ったのち、優先順位順に件数
時間指定	指定時間までに返ってきたものを優先判断

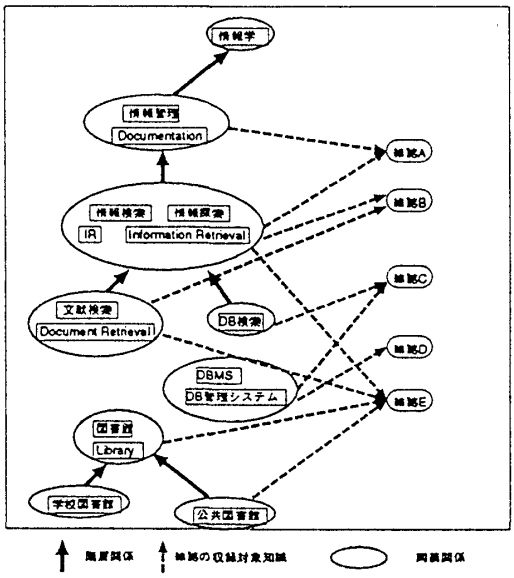

図 4. データベース選択知識

4 おわりに

従来、複数のデータベースに対して一律の検索を実行する横断検索システムかあるいはサーチャを介した検索を行なっていた複数データベースの検索を、サーチャのノウハウであるデータベース間の優先順位を反映させることで、効率的に検索を行なえる横断検索システムを提案した. 本システムは、各データベースの特徴情報と (対象領域概念, 情報源) のマッピング知識を用いることで、日本語の情報要求に対して、重なった対象領域を持つ複数データベースから適切なデータベースを選択し、それぞれのデータベースに対して適した検索式を生成するものである. そのため、ユーザはデータベースの違いを意識することなく効率的に検索を実行できる.

今後は、インプリメントしたシステムに対する評価を行なうともに、データベース選択ルールの拡張、対象領域知識の構築手法の確立を行なっていく予定である.

[参考文献]

[谷 91a] 谷幹也, 飯野香, 山口智治, 市山俊治 : 自然言語インタフェース構築キット:IF-Kit, 信学技法 NLC91-62, 1991.

[谷 91b] 谷幹也, 久保加奈子, 市山俊治, 会森清 : 情報検索におけるサーチャの知識を用いた自然言語からの検索式生成, 44 回情処全大 3G-9, 1992.