広域ネットを含むバックエンドデータベース仮想化の一手法

4R－6

池田 崇博　野村 直之
NEC 情報メディア研究所

1.はじめに

オンラインで手に入れることのできる情報には、インターネット上で公開されている情報や、パソコン通信サービスで提供されている情報をはじめ、ローカルなグループ内で共有しているデータベースにある情報や、個人のパソコン上のファイルまで、さまざまな種類のものがある。こうした、あらゆる情報源の中から、必要とする情報を収集するためには、一般に以下のような手順が必要になる。

手順1.データベース検索式の設定
手順2.検索対象データベースの選定
手順3.対象データベースへのアクセスと検索実行
手順4.検索結果の解釈

旧来のデータベース検索においては、これらすべてを逐次手動で実行しないかぎり、情報を入手できない。例えば、データベースを検索するにあたって、検索したいことがらを式の形に表せなければ、検索を実行できないし、検索結果として返されたデータの出力形式が分からなければ、データを解釈することができない。

しかしながら、これらの手順を実行するにあたっては、さまざまな問題がある。もし、欲している情報が漠然としたものであり、明確にキーワード等で表せなければ、検索を行うことができない。また、予め、どのデータベースにどんな情報があるのかを知っていなければ、適切なデータベースの選択を行うことはできない。さらに、選択したデータベースへのアクセス方法、そこでの検索の実行方法についても知っている必要がある。そのデータベースの検索結果の出力形式についても、当然知っていなければならない。結果として、必要な情報を収集するために、その情報内容とは直接関係のない思考を強いられ、本来の業務の生産性低下につながる。

2.データベース仮想化システムの構成

本研究では、文書データベースを対象として、上記の手順のどれかが不完全であったとしても、文書の検索が可能となるようなシステムを構築し、ユーザーの負担を軽減することを目指す。具体的には、ユーザーが文章を入力してだけで、それに関連する文書が自動的に検索され、表示されるシステムの実現を目指す。図 1に、システム全体の構成を示す。フロントエンド部では、キーワードの自動的抽出等により、前記手順１のユーザーの負担を軽減する。バックエンド部では、各種のデータベース毎に、接続・検索のためのプロトコル、出力される文書のフォーマット形式を記憶しておき、抽出されたキーワードに応じて、接続するデータベースをスイッチしながら自動的に検索を行うことで、前記手順2, 3のユーザーの負担を軽減する。また、検索結果を統合して表示することで、前記手順４のユーザーの負担を軽減する。本稿では、上記システムのうち、バックエンド部のデータベース仮想化について述べる。

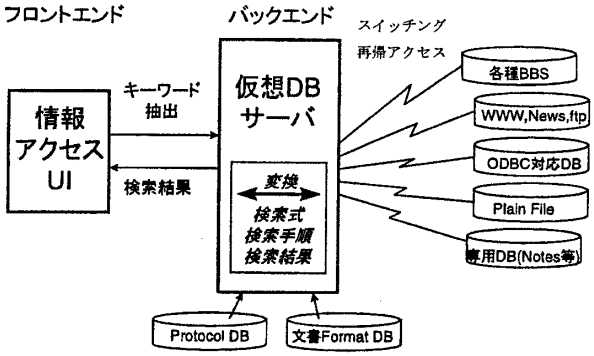

図1.データベース仮想化システム

本稿で提案する仮想化の手法では、応答時間およびデータベース利用に必要な金銭的コストを適切に抑えるため、仮想データベースへの検索要求に応じる際、すべてのデータベースに対して検索を行うのではなく、必要なデータベースだけに絞って検索を行う。次節では、検索要求に応じて、必要なデータベースを選択する手法について述べる。

3.データベース選択の手法

3.1.重み付きキーワードの利用

予め、検索の際に用いられる可能性の高いキーワードを列挙し、標準キーワードセットとして定義しておく。

仮想データベースとして１つに統合するデータベースすべてについて、予め、そのデータベース中のデータに標準キーワードセット中の各キーワードが

An Architecture for Unified Access to Multi-protocol Multiple Databases
Takahiro Ikeda and Naoyuki Nomura
NEC Information Technology Labs.

出現する頻度をキーワード毎に計算しておく。この頻度を、それぞれのキーワードに対する重みと呼ぶことにする。

仮想データベースへの検索要求があったとき、検索時に指定されたキーワードが標準キーワードセットに含まれている場合には、そのキーワードに対応する重みをデータベース間で比較し、この値がある定数値を超えているデータベースに限り、検索対象のデータベースとして選択する。逆に、ある別な定数値を下回るデータベースについては、検索対象のデータベースとしては選択しないことにする。検索時に指定されたキーワードが標準キーワードセットに含まれていない場合や、キーワードに対応する重みによる判定で検索対象とするかどうかを決定できない場合には、次の処理に進む。

仮想データベースへの検索要求時に複数のキーワードが指定された場合には、それぞれのキーワードに対応する重みの和を比較することで同様の処理を行う。

3.2.仮接続により収集した情報の利用

前節で検索対象として選択されなかったデータベースについては、そのデータベースに接続し、データよりも上位レベルの情報のみを取り寄せる。ここで上位レベルの情報とは、データベースにおけるディレクトリ構成に関する情報などを指している。これは、WWW (World Wide Web)の場合は、各 URL (Uniform Resource Location)のリンク先を調べることなく、その文書だけを取り寄せることに、パソコン通信サービスの場合は、サービスのメニューだけを取り寄せることに、ファイルシステムをデータベースとして扱う場合には、上位のディレクトリ情報だけを取り寄せることに相当する。

取り寄せた上位レベルの情報の中に検索時に指定されたキーワードがある定数値以上の割合で含まれていれば、そのデータベースを検索の対象として選択する。逆に、ある別な定数値以下の割合でしか含まれていなければ、そのデータベースを検索の対象にしないことにする。

すべてのデータベースについて、検索の対象にするかしないか分類できていない場合には、取り寄せる情報のレベルを 1 段階下げて、同様のことを行う。これは、WWW の場合は、各 URL の次のリンク先の文書を取り寄せることに、パソコン通信サービスの場合は、サービスの 1 階層下のメニューを取り寄せることに、ファイルシステムをデータベースとして取り扱う場合には、1 階層下のディレクトリ情報を取り寄せることに相当する。すべてのデータベースについて、それが検索の対象になるかならないかを決定できるまで、この処理を繰り返す。

4.議論

従来から、複数データベースを統合する、仮想的なデータベースの実現として、1 つの検索要求に対して、複数のデータベースにアクセスし、フォーマットをあわせて出力する研究が行われている[1]。しかしながら、必要とする情報は必ずしもすべてのデータベースに含まれているとは限らないため、毎回の検索毎に統合するデータベースすべてに対して検索処理を実行することは、時間的・金銭的コストの点で無駄が大きい。

WWW の単一プロトコルを対象とする検索では、インターネットロボット[2]により WWW 上のデータを予め収集しておき、収集したデータに対して検索を行うという手法が取られてきた。しかし、この場合でも、収集できるデータには量的・時間的な限界があることから、収集対象をどのように絞るかということが議論されている[3][4]。

本稿のデータベース仮想化は、1 節で述べたユーザーの負担軽減のために複数のプロトコルの差異、および、検索コストの差異を吸収し、コストの絶対値を抑えることを目標とするものである。

5.まとめ

本論文では、複数のデータベースを仮想的に 1 つのデータベースとして扱えるようにするための一手法を提案した。この手法は、統合化するデータベース数が多くなっても、検索に必要なコストが不必要に大きくならないようにするため、重み付きキーワードによるデータベースのランキングと、データベースへの仮接続による予備調査により、検索処置の対象とするデータベースを絞り込むことを特徴とするものである。

参考文献

[1] 岡他,"複数のデータベースに対する検索処理の実現方式,"電子情報通信学会第 7 回データ工学ワークショップ,pp.157-162,1996.

[2] http://info.webcrawler.com/mak/projects/robots/norobots.html.

[3] 松岡他,"WWW 情報の重要度に基づく自動収集の試み,"情報処理学会第 52 回全国大会,1-165,1996.

[4] 嶋田他,"WWW におけるユーザ主導型転送データ量制御の一手法,"情報処理学会第 52 回全国大会,3-247,1996.