データベースシステム 112－2
（1997. 5. 13）

分散環境における情報検索を支援するデータベース選択方式

須藤 昌徳　横山 和俊　井上 潮　木谷 強

NTT データ通信株式会社　情報科学研究所

E-mail：{sudo,yokoyama,uinoue,tkitani}@lit.rd.nttdata.co.jp

概要

ネットワーク上に多数のデータベースが分散して存在する環境では、全てのデータベースの内容を把握することができないため、利用者が適切なデータベースを選択し、検索することは困難である。本稿では、このような環境において、利用者の検索条件に応じて適切なデータベースをシステムが自動的に選択する方式を提案する。この方式は、検索条件を満たす記事件数の上限値と下限値を算出し、これらの値と検索語間の相関を表わす係数値を用いて期待値を算出する。実際のデータベースを用いた評価を行った結果、従来方式では考慮されていなかったOR条件への対応が可能になり、かつ検索語数の増加に対する期待値精度の向上が得られた。

キーワード: 情報検索、分散データベース、データベース選択

Database Selection for Information Retrieval under Heterogeneous Distributed Database Environments

Masanori Sudo, Kazutoshi Yokoyama, Ushio Inoue, and Tsuyoshi Kitani

Laboratory for Information Technology, NTT Data Corporation

E-mail：{sudo,yokoyama,uinoue,tkitani}@lit.rd.nttdata.co.jp

Abstract

In heterogeneous distributed environments consisting of many databases connected to netowrk,it is difficult for users to select appropriate databases and to retrieve information efficiently because they cannot understand contents of all databases. In this paper, we propose a new method of selecting databases suited for user requests under such environments.Evaluation results using real databases show that the proposed method enables the estimation for OR conditions and improves the precision for conditions with three keywords.

Keywords: information retrieval, distributed database, database selection

1 はじめに

近年、インターネットを中心としたネットワークインフラが発達し、新聞記事データベースや文献データベースなどに代表される、電子化された情報を大量に蓄積したデータベースが多数ネットワーク上に存在するようになった。その結果、利用者はそれらのデータベースに対してオンラインでアクセスし検索することにより、様々な情報を容易に取得することが可能になった。しかしこれらのデータベースに対してオンラインで検索を行なう場合、以下のような問題点があり、利用者は必要とする情報を効率よく取得することが困難になっている。

(1) データベースに含まれる内容が詳しく分からないため、利用者が必要とする情報を得るために、どのデータベースを検索すればよいか分からない。

(2) データベースごとにコマンド等の検索方法が異なるため、対象とするデータベースの検索手順を調べる必要がある。

我々は、このようなネットワーク上に分散してデータベースが存在する環境において、利用者が必要とする情報を効率よく取得できるシステムの実現を目指して、DistAIRS(Distributed Agent-oriented Information Retrieval System)[1][2]の研究を進めている。DistAIRSは新聞記事等の文書データベースに対して、利用者からの検索要求をもとに適切なデータベースを選択し(データベース選択支援機能)、検索を代行する(データベース検索実行支援機能)システムである。

本稿では、データベース選択支援機能に着目し、その精度を向上させる方式について検討する。まずデータベース選択支援に求められる要件を明らかにし、従来のデータベース選択方式の問題点を示す。その問題点を解決するための新しい選択方式を提案し、その効果について実際のデータベースを用いた評価を行なう。

2 データベース選択支援

2.1 分散環境における情報検索の問題点

ネットワーク上に多数のデータベースが分散した環境で利用者が検索を行なう場合の問題点について述べる。

最も単純な検索方法は、検索可能なすべてのデータベースに対して同じ条件で検索を行なうことである。しかしこの方法は、検索結果を得るまでに時間がかかったり、ネットワークやデータベースの負荷が高くなるため、望ましくない。

そこで、利用者は少数のデータベースのみを検索して必要な情報を取得できるようにデータベースを選択する必要がある。そのためには、データベースの内容を把握しておく必要があるが、多数のデータベースすべてに対してその内容を知ることは困難である。内容が類似したデータベースが複数存在した場合、利用者が必要とする情報を得るために、どのデータベースを検索するのが最適かを判断することはさらに困難である。加えて、各データベースの内容が随時更新されるという点も、データベースの内容を把握することを困難にしている。すなわち検索データベースの選択は、利用者にとって大きな負担となっている。

このような利用者の負担を軽減するため、多数のデータベースの中から検索条件に応じた適切なデータベースを自動的に提示し、利用者の情報検索を支援するシステムが必要である。

2.2 関連研究

データベース選択支援機能を持つシステムとして、Lycos[3]、WAIS[4]、GlOSS[5]、gGlOSS[6]が提案されている。

Lycosは WWW における URL 形式の情報タグとタグに関連したキーワードを蓄積したシステムであり、利用者からの問い合わせに対して適切と思われる URL を通知する。WAIS は、複数データベースに対してデータベースの内容を管理するサーバを構築し、利用者からの要求に対してサーバが適切な情報を持つデータベースの所在を通知する。しかし、Lycos は URL のみを検索対象としており、WAIS は独自のデータベースシステムであるため、どちらも既存のデータベースに蓄積された情報の検索には適用できない。

一方、GlOSS および gGlOSS は既存のデータベースを対象としており、利用者の要求に適合するレコードを最も多く持つと推定されるデータベースを選択する。詳細については3章で議論する。

2.3 DistAIRS

我々が研究を進めているエージェント指向情報検索システム DistAIRS は、新聞記事や文献など文書データベースの選択・検索実行を支援するシステムである。DistAIRS の構成を図1に示す。インタフェースエージェントは利用者の要求を解釈し、検索コマ

ンドに変換する。検索プランエージェントは、検索コマンドにもとづいて、データベースを選択し、結果を統合する。データベースエージェントはデータベース検索を実行する役割を持つ。

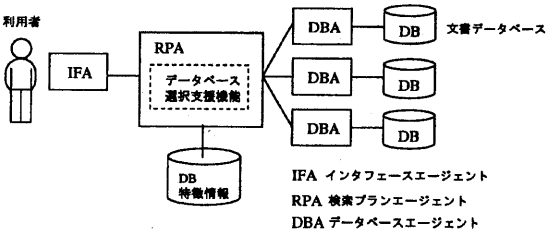

図 1: DistAIRS 構成図

本稿で議論する検索プランエージェントのデータベース選択機能は、具体的に以下のモデルでデータベースの選択を行なう。

(1) 入力 利用者から与えられる検索条件は、複数の検索語を AND 条件と OR 条件で組合せた論理式とする。この検索条件は利用者による入力が容易であり、文書データベースの検索方法として一般的である。

(2) データベース特徴情報 データベースの内容を表わす情報として、あらかじめ用意されたキーワードとそれを含む記事の出現頻度を格納する。これをデータベース特徴情報と呼ぶ。表 1にデータベース特徴情報の構成イメージを示す。

表 1: データベース特徴情報の構成例

キーワード	データベース A	データベース B
雇用	1,144	947
人事	1,847	1,290
総記事件数	101,058	91,774

(3) データベース選択方式 検索条件に合った適切なデータベースを選択するにあたり、その検索条件を満たす記事がどれだけ多く含まれているかを基準とする。すなわち、検索条件を満たす記事を多く含んでいるほど、優先順位の高いデータベースと判断する。そのため、利用者からの検索条件とデータベース特徴情報を用いて、各データベースの記事件数を推定し、その値をもとにデータベースの選択を行なう。データベース選択方式に要求される機能として、次の 2 つが挙げられる。

(a) 利用者の検索効率を高めるため、検索条件に適合したデータベースの優先順序を精度良く算出する。

(b) 利用者にとってデータベースを検索した結果得られる記事件数は重要である。そのため各データベースごとに取得できる予想記事件数を精度良く算出する。

3 GlOSS 方式によるデータベース選択

2章で述べたデータベース選択支援機能を実現するにあたり、 GlOSS 方式及び gGlOSS 方式が提案されている。しかし gGlOSS 方式では検索優先順序を各データベース間の相対値で決定しており、取得記事の予想件数を得ることができない。そこで本章では、GlOSS 方式による期待値算出アルゴリズムの詳細と、その問題点について述べる。

3.1 期待値算出方式

(1) データベース特徴情報として、あらかじめ用意されたキーワードとそれを含む記事の出現頻度、及びデータベース全体の総記事件数を用いる。

(2) 次に示す式 (1) を用いて検索条件を満たす記事件数の期待値 GDK_{DB_n} を算出する。

$$GDK_{DB_n} = T_{DB_n} \prod_{k \in K} \frac{D_{(k,DB_n)}}{T_{DB_n}} \quad (1)$$

GDK_{DB_n} : データベース (DB_n) において、利用者が与えた検索条件を満たす記事件数の期待値
T_{DB_n} : データベース (DB_n) に格納されている総記事件数
K : 利用者が与えた検索条件
k : キーワード
$D_{(k,DB_n)}$: データベース (DB_n) において、キーワード k を含む記事件数

(3) (2) で算出した各データベースの記事件数期待値の大きい順にランク付を行ない、データベースの検索優先順序とする。

3.2 GlOSS 方式の問題点

GlOSS 方式によるデータベースの選択の問題点を、以下に示す。

(1) GlOSS 方式で用いる期待値を算出する式は、検索語の AND 条件による検索条件にしか対応していない。そのため OR 条件や、AND 条件と OR 条件の組合せで与えられた検索条件に対して、記事件数の期待値を算出することができず、データベースの選択が行なえない。

(2) GlOSS 方式によって算出される記事件数の期待値と、実際の検索によって得られる記事件数との誤差が大きい。

(3) 検索条件の中に含まれる検索語が増加すると期待値が極端に小さくなり、期待値の精度がさらに低下する。

4 相関関係を考慮した期待値算出方式

前章で述べた問題点を解決するために、新しい期待値算出方式を提案する。

以下、提案する期待値算出方式について述べる。

4.1 基本方式

4.1.1 着眼点

GlOSS 方式では検索語間が独立であると仮定しているが、実際には検索語間には相関がある場合が多い。そこで提案方式ではデータベース特徴情報から検索条件を満たす記事件数の上限値と下限値を算出し、この上下限値と検索語間の相関関係を表わす係数値 α を用いて期待値を算出する。

上下限値は検索条件中に AND 条件や OR 条件が混在しても算出することができるため、GlOSS 方式と比べて期待値を算出することができる検索条件の幅が広がる。また相関を表わす係数値 α を適切な値に設定することにより、期待値精度を高めることができる。さらに検索語の個数の増加に対しては、上下限値をベースにした算出方式と検索語の個数を考慮して係数値 α を設定することにより極端な精度の悪化を抑えることができる。

4.1.2 期待値算出手順

以下、2 検索語からなる検索条件の期待値算出手順について述べる。

(1) 検索語の記事件数の取得 検索条件 K 中の検索語 k_A, k_B において、データベース DB_n の中で検索語を含む記事件数 $D_{(k_A,DB_n)}$、$D_{(k_B,DB_n)}$ をそれぞれ取得する。

(2) 上下限値の決定 (1) で抽出した記事件数と検索条件の論理式 (AND 条件、OR 条件) から、検索条件 K を満たすデータベース DB_n の記事件数の上限値と下限値を求める。式 (2)(3) に上限値 $Max(K, DB_n)$、下限値 $Min(K, DB_n)$ を算出する式を示す。

$$Max(K, DB_n) = \begin{cases} min\{D_{(k_A,DB_n)}, D_{(k_B,DB_n)}\} \\ \quad (K \in \text{AND 条件}) \\ D_{(k_A,DB_n)} + D_{(k_B,DB_n)} \\ \quad (K \in \text{OR 条件}) \end{cases} \quad (2)$$

$$Min(K, DB_n) = \begin{cases} 0 \\ \quad (K \in \text{AND 条件}) \\ max\{D_{(k_A,DB_n)}, D_{(k_B,DB_n)}\} \\ \quad (K \in \text{OR 条件}) \end{cases} \quad (3)$$

$Max(K, DB_n)$: データベース DB_n に含まれる検索条件 K を満たす記事件数の上限値

$Min(K, DB_n)$: データベース DB_n に含まれる検索条件 K を満たす記事件数の下限値

k_A, k_B : 検索条件 K の検索語

$D_{(k_A,DB_n)}$: データベース DB_n の中の検索語 k_A,k_B を含む記事件数

$D_{(k_B,DB_n)}$

例えば、表 1のデータベース特徴情報が与えられた場合、検索条件「雇用 AND 人事」、「雇用 OR 人事」のデータベース A の上下限値は、表 2の通りである。

表 2: データベース A の上下限値例

	雇用 AND 人事	雇用 OR 人事
上限値	1,144	2,991
下限値	0	1,847

(3) 期待値の算出 期待値 $E(K, DB_n)$ を次に示す式 (4) を用いて算出する。係数 α の決定方法については、4.3節で述べる。

$$E(K, DB_n) = (Max(K, DB_n) + Min(K, DB_n)) \times \alpha(K, DB_n) \quad (4)$$

$E(K, DB_n)$: 検索条件 K に対するデータベース DB_n の記事件数期待値

$\alpha(K, DB_n)$: 検索条件 K に対する相関度を表す係数

以上の方法で算出した期待値をもとに、各データベースの検索優先順序を決定する。

4.2 3検索語以上への拡張

3語以上の検索語からなる検索条件の場合、検索条件を2語の検索語からなるAND条件とOR条件の組合せとみなして2語ごとに分割し、2検索語の期待値算出方式をを利用する分割方式によって求める。

例えば検索語 k_A、k_B、k_C に対して、検索条件「(k_A AND k_B) OR k_C」が与えられたとき、「k_A AND k_B」の期待値を求める。その期待値を検索条件「k_A AND k_B」の記事件数とみなし、k_C の記事件数とOR条件の係数値 α を用いて検索条件「(k_A AND k_B) OR k_C」の期待値を求める。

しかし、連続する3検索語以上のAND条件またはOR条件のみの組み合わせに対し、この提案方式では分割の仕方によって期待値が異なる。このような検索条件に対し、以下2通りの分割方式による期待値算出方法を示す。

(1) 検索語の並びに従い、左から順番に分割する。例えば検索条件「k_A AND k_B AND k_C」の場合、はじめに「k_A AND k_B」の期待値を求め、その期待値と k_C の記事件数から検索条件の期待値を算出する。この方式を検索順方式と呼ぶ。

(2) 与えられた検索条件の記事件数が大きい順に検索語を組み合わせて、期待値を算出する。例えば検索条件「k_A AND k_B AND k_C」で k_A、k_B、k_C の記事件数がそれぞれ40、20、80件の場合、はじめに「k_A AND k_C」の期待値を求め、その期待値と k_B の記事件数から検索条件の期待値を算出する。この方式を件数順方式と呼ぶ。

4.3 係数αの設定方法

検索語間の相関度は、それぞれの検索語に依存するため、一般的に係数 α は検索条件によって変化する。しかし全ての検索語間の相関度をデータベース特徴情報として持つのは困難である。

提案方式では係数 α を定数とし、複数の検索条件による予備実験からAND条件とOR条件の設定値をデータベースごとに決定する。

2検索語の係数 α の決定方法を、以下に述べる。

1. ランダムに選択した2語の検索語からなるAND条件の検索条件を生成し、α 設定用検索条件とする。
2. 各データベース毎に α 設定用検索条件に対する期待値と実際の記事件数とを算出し、その誤差が最小になるように α を決定する。

同様な方法で、OR条件及び複数検索語に関する係数 α を決定する。

5 評価

前節で述べた提案方式による効果が期待通り得られるかどうかを検証するため、実際のデータベースを用いて評価を行った。

5.1 評価項目

(1) 各方式の期待値精度の評価 各方式で算出した期待値の精度 $EP(DB_n)$ を式 (5) を用いて算出し、評価する。これは検索条件によって得られる期待値と実記事件数との平均誤差と、各検索条件によって得られる平均記事件数との割合である。すなわち、精度が0に近いほど誤差が小さく、その期待値は実際の記事件数に近いと言える。

$$EP(DB_n) = \frac{\text{期待値と実記事件数との平均誤差}}{\text{平均取得記事件数}} \quad (5)$$

(2) 各方式の選択正答率の評価 各方式で決定した検索優先順序の精度を選択正答率で評価する。

検索条件に対して最も記事件数が多いデータベースを選択正答データベースと呼ぶ。このとき検索優先順序に従って選択したn個のデータベースの中に、選択正答データベースが含まれている場合の検索条件の個数と、評価に用いる検索条件の総数との割合を選択データベース数n個の選択正答率 $DSCR(n)$ と呼ぶ。(式 (6))。すなわち選択データベース数が少なくて選択正

答率が高ければ、検索優先順序の精度が高いと言える。

$$DSCR(n) = \frac{RCN(n)}{TRC} \qquad (6)$$

$DSCR(n)$: 選択データベース数が n 個のときの選択正答率

$RCN(n)$: 検索優先順序に従って選択したデータベース n 個の中に、最も記事件数が多いデータベースが含まれている検索条件の個数

TRC : 評価に用いる検索条件の総数

以上2つの評価項目を用いて、提案方式を評価する。評価対象として、2検索語のAND条件ではGlOSS方式と提案方式とを比較し、OR条件では提案方式による結果のみを示す。また3検索語のAND条件に関してはGlOSS方式と検索順方式、件数順方式に加えて、検索条件を分割しない一括方式との比較も行なう。一括方式とは、検索条件「k_A AND k_B AND k_C」が与えられたとき、「k_A AND k_B AND k_C」の上下限値と3検索語AND条件用のαを用いて期待値を算出する方式である。OR条件は検索順方式、件数順方式、一括方式を比較する。

5.2 実験環境

評価実験の環境と条件について、以下に述べる。

(1) 検索対象データベース 評価に用いた12種の検索対象データベースと実験に使用した記事件数を表3に示す。

(2) 検索条件の生成 評価に用いた検索条件を以下の手順で生成する。

1. ランダムに抽出した記事の中で、出現頻度の高い名詞を検索語候補群とする。
2. 抽出した検索語候補群の中からランダムに検索語を選択し、検索式を生成する。今回は2語の検索式に関しては、AND条件、OR条件それぞれについて1000組の検索式を用いた。また、3語の検索語式に関しては、AND条件、OR条件それぞれについて3000組の検索式を用いた。

(3) αの設定方式 係数αは4.3節で述べた設定方法にもとづき、2検索語、3検索語のAND条件、OR条件の値をそれぞれ決定する。予備実験で用いるα設定用検索条件の数は、実記事件数が10件以上である1000組とする。

表3: 検索対象データベース

	データベース名	記事件数
1	日経金融新聞94年版	28,162
2	日経流通新聞94年版	18,897
3	日経産業新聞94年版	67,371
4	日経新聞(本紙)94年版	184,930
5	毎日新聞94年版	101,058
6	毎日新聞93年版	91,774
7	毎日新聞92年版	86,094
8	毎日新聞91年版	78,549
9	ネットニュース(fj.compグループ)	13,159
10	ネットニュース(fj.socグループ)	44,593
11	ネットニュース(fj.sciグループ)	16,388
12	ネットニュース(fj.recグループ)	134,384

(4) 有効検索条件 評価に用いている検索条件は、ランダムに生成するため、実記事件数が小さな検索条件が多く、期待値の精度を評価するには適切ではない。そこで実際の検索で用いられる検索条件に近いもので評価するため、実記事件数が10件以上の検索条件を、期待値精度の評価対象とする。

5.3 評価結果

5.3.1 期待値精度の評価結果

各検索条件における提案方式とGlOSS方式の期待値精度を図2、図3、図4、図5にそれぞれ示す。ただし、OR条件は提案方式による評価結果のみを示す。また横軸の番号は、表3のデータベースの番号に対応している。

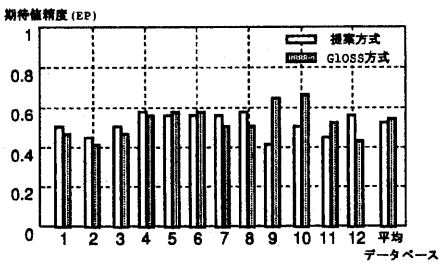

図2: 2検索語のAND条件の期待値精度

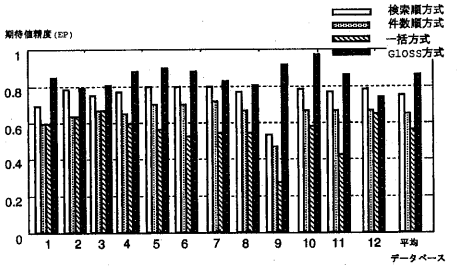

図 3: 3 検索語の AND 条件の期待値精度

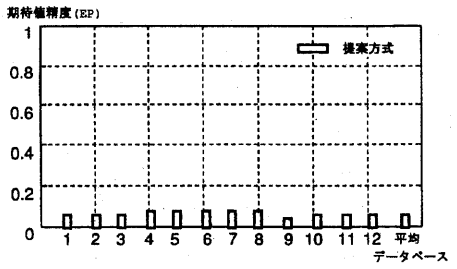

図 4: 2 検索語の OR 条件の期待値精度

5.3.2 選択正答率の評価結果

各検索条件における提案方式と GlOSS 方式の選択正答率の推移を図 6、図 7、図 8、図 9にそれぞれ示す。ただし OR 条件は提案方式による評価結果のみを示す。

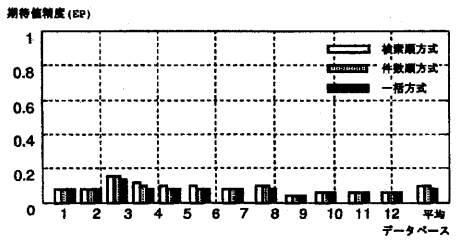

図 5: 3 検索語の OR 条件の期待値精度

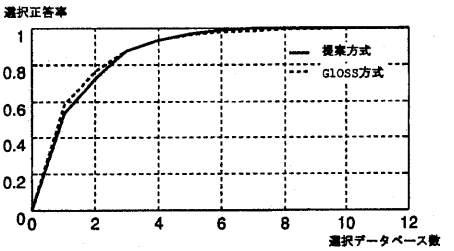

図 6: 2 検索語の AND 条件による選択正答率

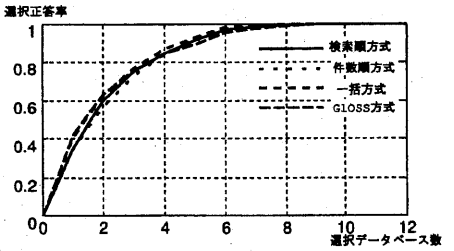

図 7: 3 検索語の AND 条件による選択正答率

5.4 考察

OR 条件による期待値精度は、すべてのデータベースで 0.1 以下であり、良好な結果を得ることができた。また選択正答率も選択データベース数が 1 個の場合で 9 割を超えており、提案方式による OR 条件のデータベース選択が有効であることを示している。

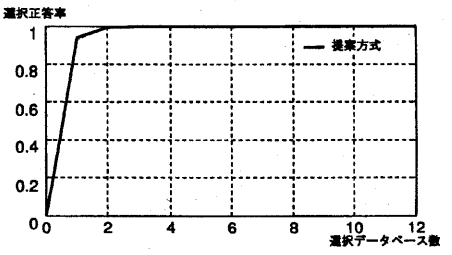

図 8: 2 検索語の OR 条件による選択正答率

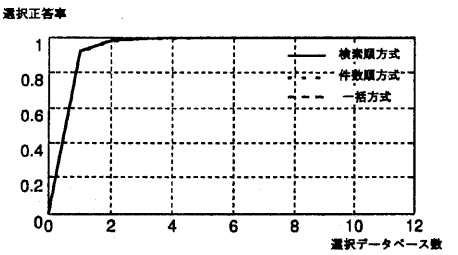

図 9: 3 検索語の OR 条件による選択正答率

また 3 検索語における各方式の期待値精度の差は 0.01 程度であり、方式の違いによる差は見られなかった。

2 検索語の AND 条件による期待値精度は、提案方式により改善されたデータベースもあり、平均で約 0.02 ポイント改善されているが、全般的に GlOSS 方式と同等程度の精度であった。しかし 3 検索語の AND 条件による期待値精度は、分割方式による 2 方式はいずれも GlOSS 方式と比べ良い精度が得られ、検索順方式で 0.1 ポイント、件数順方式で 0.2 ポイント向上している。この結果から、検索語数の増加によって期待値精度がさらに低下する問題点は、改善されている。さらに過去の検索結果を係数 α に動的に反映させる機能を組込むことにより、期待値の精度を高められる可能性がある。また分割方式による 2 方式と一括方式の精度を比較すると、一括方式が最もよい精度が得られた。しかし一括方式の場合、4 語以上の検索条件での評価を行なう必要があり、課題が残されている。

一方選択正答率は、選択データベース数が 1 個のとき、2 検索語の場合は 0.04 ポイント低下し、3 検索語の一括方式の場合は 0.02 ポイント低下している。すなわち 2 検索語、3 検索語の場合ともに選択正答率の向上は見られなかった。

6 まとめ

ネットワーク上に複数のデータベースが存在する分散環境において、情報検索を行なうためのデータベース選択支援について述べた。

分散環境において、効率よく情報検索を行うためには、与えられた検索条件に応じた適切なデータベースを選択する必要がある。そのため本稿では、データベース特徴情報から求められる検索条件を満たす記事件数の上下限値と、予備実験で設定した検索語間の相関係数から期待値を算出する方式を提案し、期待値精度と選択正答率を評価した。

その結果、従来方式では考慮されていなかった OR 条件の検索条件に関するデータベース選択が可能になった点と、検索語数の増加に対する期待値精度の向上という点で、提案方式の有効性が明らかになった。しかしデータベースの選択正答率に関しては向上が見られないことがわかった。

今後の課題として、以下の点が挙げられる。

(1) 一括方式における複数検索語用の係数 α の設定

(2) 過去の検索結果にもとづき、動的に変化させる方法を含めた係数 α の設定方法の検討

(3) AND 条件や OR 条件が混在する検索条件での、提案方式の評価

謝辞

本研究にあたり、新聞記事 CD-ROM の使用を了解いただいた (株) 日本経済新聞社ならびに (株) 毎日新聞社に感謝致します。

参考文献

[1] 箱守聰、横山和俊、田辺雅則、井上潮：異種分散環境におけるエージェント指向型情報検索システム – 設計方針と基本構成 –, 情報処理学会第 52 回全国大会, pp.227-228, 1996.

[2] 田辺雅則、箱守聰、井上潮：異種分散環境における エージェントを用いた情報検索方式, アドバンスト・データベースシステム・シンポジウム, pp.215-222, 1995.

[3] M. L. Mauldin, J. R. R. Leavitt: Web Agent Related Research at the Center for Machine Translation, SIGNIDR meeting, 1994.

[4] B.Kahle,A.Medlar: An Information system for corporate users: Wide Area Information Servers, Techinical Report TMC199, Thinking Machines Corporation, 1991.

[5] L. Gravano, H. Garcia-Molina, A. Tomasic: The Effectiveness of GlOSS for the Text Database Discovery Problem, Proc. of ACM-SIGMOD, pp.126-137, 1994.

[6] L. Gravano, H. Garcia-Molina: Generalizing GlOSS to Vector-Space Databases and Broker Hierarchies, Proc. of 21st VLDB conference, 1995.