分散型検索システムにおける連携型検索手法の提案

澤田 雅人 富田 準二 池田 哲夫 佐藤 哲司
日本電信電話株式会社 NTT サイバースペース研究所
{sawada@, tomita@, ikeda@dq., satoh@}isl.ntt.co.jp

概 要

Gnutella のように、情報源ごとに検索システムを配置し、これらの検索システム同士をネットワークで接続した分散型の情報検索システムが利用されはじめている。このようなシステムでは、個々の情報源の質に大きなばらつきが存在するために、それぞれの情報源ごとの検索結果を単純にマージするだけでは、検索結果の上位に目的の情報が出現せず、検索精度が低下する場合がある。

本稿では、検索者が目的としている情報が存在しそうな情報源を検索の起点とし、その情報源に対応する検索システムに検索要求を送信することと、情報源に対する評価を検索システム間のリンクの重みとし、この重みを用いて検索結果のスコアを再計算することで検索精度を向上させる、連携型検索手法を提案する。また、本手法の評価実験を行い有効性の評価を行った。

1 はじめに

近年、ADSL をはじめとする高速・常時接続のネットワークが普及してきている。また、個人が利用する計算機においても、CPU の高速化やメモリの大容量化といった性能の向上が著しい。

このようなネットワークの高速化と個々の計算機の性能向上により、Gnutella のように、情報源ごとに検索システムを配置し、これらの検索システム同士をネットワークで接続した分散型の情報検索システムが利用され始めている。

このような分散型の情報検索システムにおける個々の情報源としては、個々人のコンピュータに蓄積された情報が考えられる。また、ある個人や団体が作成し、インターネット上で公開している Web ページの集合を Web サイトと呼ぶことがあるが、これらの Web サイトも個々の情報源とみなすことができる。

このように異なる人物や団体が作成する情報には、それぞれの個人や団体の興味や目的の違いによって、情報源に含まれている情報の内容に偏りが生じることがある。例えば、魚釣りを趣味とする人が持っている情報には、過去の釣果や仕掛けの情報など、魚釣りに関する情報が多いと考えられ、また、登山愛好者の団体が公開する Web サイトでは、登山道具のレビューやおすすめの山の紹介など、登山に関する情報が公開されていると考えられる。

また、作成した人物や団体の能力によって情報の質や量に差が生じることがある。例えば、個人が趣味として調査した情報と、企業が調査した情報では、費やした費用や時間などの差によって、情報源に含まれる情報の質や量に差が生じることは十分に考えられる。さらに、似たような興味を持つ人同士や同じような活動をしている団体同士では、関連性のある情報を互いに保有している可能性が高い。

しかしながら、既存の分散型の情報検索システムでは、基本的には全ての情報源を対等に扱っている。そのため、これら個々の情報源が持つ情報の偏りや質や量の差を考慮して、検索者が目的としている情報を保有する情報源からの検索結果を優先して提示することができず、結果として検索精度が低くなってしまう可能性がある。

そこで本稿では、検索者が目的としている情報が存在しそうな情報源を検索の起点とし、その情報源に配置された検索システムに検索要求を送信することと、関連情報が存在する情報源に配置された検索システムの間に検索システム間リンクを設定し、情報源に対する評価を検索システム間リンクの重みとして付加し、この重みを用いて検索結果のスコアを再計算することで検索精度を向上させる、連携型検索手法を提案する。

以下、2 章では既存の分散型検索手法とその問題

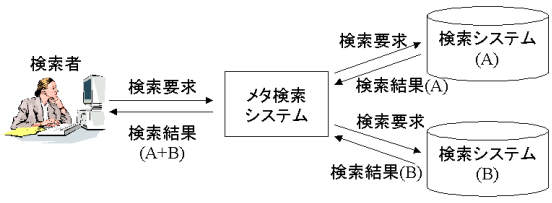

図 1: メタ検索手法

点を述べる。3 章では既存手法の問題点を解決する連携型検索手法について説明し、4 章で提案手法の実装について説明する。5 章で提案手法の評価と考察を行い、6 章でまとめと今後の課題を述べる。

2 既存の分散型検索手法

既存の分散型検索手法として、メタ検索手法 (図 1)[1] や、Gnutella などで利用されているバケツリレー型の検索手法 (図 2)[2] が存在する。以下ではこの 2 つの手法について説明する。

- メタ検索手法

  対等な複数の検索システムと、各検索システムを統括するメタ検索システムから構成される。

  検索者がメタ検索システムに対して検索要求を送信すると、メタ検索システムは受け取った検索要求を各検索システムに対して送信し、検索結果を受信する。

  最後に、受信した複数の検索結果を一つの検索結果にまとめて最終的な検索結果として検索者に返す。

- バケツリレー型検索手法

  対等な複数の検索システムのみで構成され、検索システムはいくつかの検索システムと接続されている。この接続を利用して検索要求を転送し、検索結果を返す。

  この手法においては、検索者の送信する検索要求と検索システムが転送する検索要求に基本的な差はない。そこで、一つの検索システムに着目すると、検索システムが検索要求を受け取ると、検索要求の送信元以外の接続されている全ての検索システムに対して検索要求を転送し、転送先から検索結果を受け取る。この検索結果と自分の持つ情報に対する検索結果とを一つにまとめて検索結果として検索要求の送信元に返す。

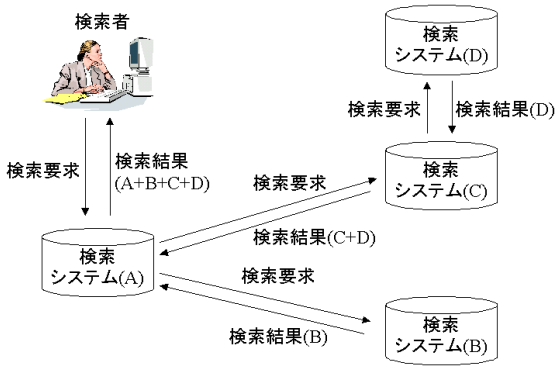

図 2: バケツリレー型検索手法

本稿が対象としている情報源は、個々人のコンピュータに蓄積された情報や、ある個人や団体が作成し、インターネット上で公開している Web サイトである。

このように異なる人物や団体が作成する情報には、それぞれの個人の興味や団体の活動内容などの違いによって、情報源に含まれている情報の内容に偏りが生じることがある。また、情報源を管理する人物や団体の能力によって、情報源に含まれる情報の質や量に差が生じることがある。さらに、似たような興味を持つ人同士や同じような活動をしている団体同士では、関連性のある情報を互いに保有している可能性が高い。

このような情報源に対して既存の分散型検索手法では、情報源に含まれる情報の質や情報源同士の関連性について考慮していないため、これらを利用した検索を行うことは難しい。そのため、全ての情報源を対等に扱うこととなり、検索者が目的としている質の高い情報を持っている情報源の検索結果を優先して出力することができず、結果として検索精度が低くなってしまう。

3 提案手法

本稿では、複数の情報源にまたがって存在する関連情報を一括して検索することができ、検索者が目的としている質の高い情報が検索結果の上位に出力されるようにするために、実世界の情報探索を

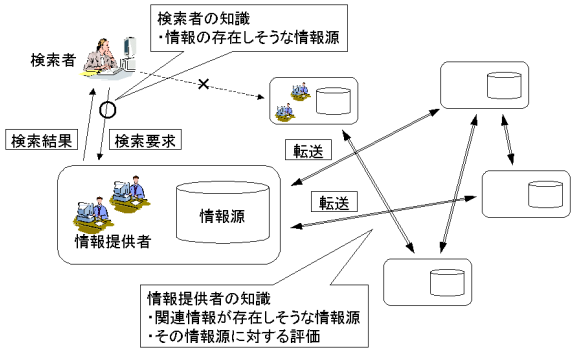

図 3: 手法の構成

モデルとした連携型検索手法を提案する。提案手法では、情報源に含まれる情報に偏りがある点、情報源同士の関連性や情報源の質に差がある点を利用する。

3.1 実世界での情報探索

実世界において情報を探す場面は、主に以下の二つのステップで構成されていると考えられる。

Step 1. 身近にいる「知ってそうな人」に聞く

自分で探し出せない情報については、通常、自分の周りにいる「知ってそうな人」に尋ねる。

Step 2. 知人で「知ってそうな人」に聞いてもらう

Step 1. で十分な情報が得られなかった場合や、さらに情報が必要である場合には、Step 1. で最初に尋ねていった人に、その人の知人で「知ってそうな人」に尋ねてもらい、結果を教えてもらう。

つまり、情報を探している人 (以下、質問者とする) は、その情報を知ってそうな人が誰であるかという知識を利用して、最初に尋ねにいく人を決めている。また、その人が尋ねた知人も、自分以外で知ってそうな人が誰であるかという知識を利用して、誰に尋ねるのかを決めている。

3.2 連携型検索手法

3.2.1 実世界をモデルにした検索手法

実世界における情報探索の構成要素としては、質問者、質問者の知人、その知人が尋ねる知人が存在する。これを分散型検索システムに当てはめると、実世界での質問者というのは検索者であり、質問者が尋ねた知人や知人が尋ねた知人というのは、情報源の情報提供者ということになる。

これらを基に、実世界での情報探索をモデルにし、「検索者の知識」と「情報源の情報提供者の知識」を利用した検索手法の構成を図 3 に示す。

検索者の知識として利用できるものとしては、実世界における「目的の情報を知ってそうな人が誰であるか」という知識である。これを分散型検索システムに当てはめ、「情報が存在しそうな情報源はどこか」という知識と考える。本稿では、この知識を検索起点として表現し利用する。

情報源の情報提供者の知識として利用できるものとしては、実世界における「自分以外で知ってそうな人が誰であるか」という知識である。これを分散型検索システムに当てはめ、「関連情報を持っている情報源はどこか」という知識と、「その情報源に対する評価」という知識と考える。本稿では、この知識を検索システム間リンクとして表現し利用する。

3.2.2 検索起点

「情報が存在しそうな情報源はどこか」という知識を利用するために、検索者は通常の検索キーワードに加え、検索起点を指定する。

検索起点の指定には、以前に行った同じような情報に対する検索の結果から、必要としている情報を持っている可能性が高いとおもわれる情報源を推定するといった手段が利用できる。

この検索起点の指定を行うことで、所望の情報が存在しそうな情報源を中心とした検索を行うことが可能となる。

3.2.3 検索システム間リンク

「関連情報を持っている情報源はどこか」という知識を利用するために、情報源における情報提供者が、自分の管理している情報源と関連情報を持っている情報源の間に検索システム間リンクを張る。この検索システム間リンクを利用することで、ある情報源の持っている情報と関連する情報を持っている情報源を指定することが可能となる。

「情報源に対する評価」という知識を利用するために、同じく、情報源における情報提供者が、リン

ク先の情報源の持つ情報を評価して、リンクの重み付けを行う。このリンクの重みを利用することで、複数存在するリンク先の情報源のうち、どの情報源の検索結果を優先して検索者に提示するかを決定することが可能となる。

3.2.4 検索アルゴリズム

まず、3.2.3 節で述べた検索システム間リンクと検索システム間リンクの重みを事前に決定しておく。

提案手法では、この検索システム間リンクと検索システム間リンクの重みを用いて、以下のアルゴリズムにしたがって検索を行う。

1. 検索者は検索起点とする検索システムを選択し、その検索システムに対して検索要求を送信する
2. 検索要求を受け取った検索システムは、検索システム間リンクのなかから、検索要求が経由してきた検索システム以外の検索システムに対して検索要求を転送する
3. 検索システム自身が持つ情報に対して検索を行う
4. 検索システム自身の検索結果と検索要求の転送先の検索システムからの検索結果、それぞれの優先度を検索システム間リンクの重みを用いて算出する
5. 全ての検索結果を優先度の順に並べ替えて最終的な検索結果として検索者に返す

4. における検索結果の優先度 (L_w) は、検索結果が経由した検索システム間リンクの集合を C、C に含まれる検索システム間リンクの重みをそれぞれ W_i として以下の式で算出する。

$$L_w = \prod_{W_i \in C} W_i \tag{1}$$

なお、検索システム自身の検索結果の優先度は $L_w = 1$ とする。

また、複数の経路を経由して送られてきた同じ検索システムからの検索結果が、異なる優先度で出現する場合が起こる。これに対しては、提案手法では優先度が最も高いものを検索結果として採用することとする。

上記の式を用いることにより、検索システム間リンクの重みが高く、検索起点に近い情報源の検索結果の優先度を高くすることができる。

次に、5. における検索結果の並び替えについて説明する。一般には、それぞれの情報源に配置される検索システムが異なると、検索結果のスコアリング方式が異なることや、そもそもスコアリングを行わないことがある。そのため、各検索システムの検索結果の要素ごとにスコアを比較して並べ替えを行うことができない。そこで提案方式では、検索結果を検索システム毎にまとめ、それぞれの検索システムの優先度の順に並べ替えを行う。

ところが、各情報源の検索システムが採用しているスコアリング方式が統一されている場合には、検索結果の要素毎のスコアを比較することができる。そこで、このような場合には以下に述べるスコアの再計算と並び替えを行う。

- 各検索システムの検索結果の要素毎のスコアにその検索システムの優先度を掛けたものを最終的なスコアとする
- 検索結果の各要素を最終的なスコアの高い順に並び替える

検索結果のスコアの再計算と並び替えにより、優先度が高い情報源、すなわち質の高い情報を持つ情報源において検索要求に適合している情報を、最終的な検索結果の上位に出力することが可能となる。

4 実装

4.1 システム構成

提案手法を実装したシステムの構成 (図 4) について説明する。各情報源には、既存の検索システムと、連携型検索手法を実現するための連携サーバを配置する。

連携サーバでは、検索者からの検索要求の受信や検索結果の送信、検索システムを利用した情報源に対する検索、他の連携サーバへの検索要求の転送や検索結果の優先度の算出、検索結果の並べ替えを行う。連携型検索手法で利用する検索システム間リンクは設定ファイルに記述し、連携サーバ起動時に読み込む。

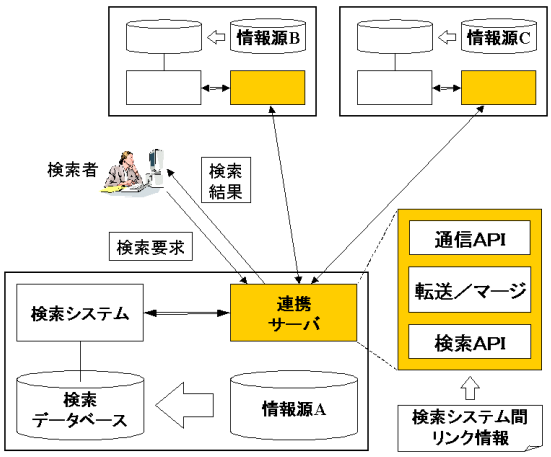

図 4: システム構成

連携サーバの内部は大きく分けて、「通信 API」、「転送／マージ」「検索 API」の 3 つのモジュールから構成されている。

- 通信 API

  検索者や他の連携サーバとの通信処理を行うモジュール

- 転送／マージ

  検索要求の転送および検索結果の優先度の算出、検索結果の並べ替え処理を行うモジュール

- 検索 API

  検索システムと通信し、情報源に対する検索処理を行うモジュール

このようなモジュール構成をとることによって、通信プロトコルの変更に対しては、通信 API モジュールの交換のみで対応できる。また、検索システムの変更に対しても検索 API モジュールの交換を行うことで対応できる。同様に、連携型検索手法のアルゴリズムの変更に対しても、転送／マージモジュールの交換のみで対応できる。

今回試作したシステムでは、検索システムには、検索システムごとのスコアリング方式の違いによる検索精度への影響を避け、提案手法の有用性を正確に検証するために、全ての検索システムに我々の研究グループで開発している XML 文書検索エンジン LISTA [3] を用いた。

また、連携サーバでは、連携サーバ間の通信プロトコルとして XML をベースにした独自プロトコルを開発し、このプロトコルを用いて通信を行っている。通信プロトコルを XML ベースとしたことにより、プロトコルの拡張や変更を柔軟に行うことが可能となっている。

4.2 動作手順

図 4 を用いて本システムの動作について説明する。以下では、情報源 x の連携サーバを 連携サーバ x 、検索システムを 検索システム x と表記する。

1. 検索者 : 検索起点とする検索システム A を選択
2. 検索者 : 連携サーバ A に検索要求を送信
3. 連携サーバ A : 通信 API を通じて検索要求を受信
4. 連携サーバ A : 転送／マージモジュールで転送経路を確認
5. 連携サーバ A : 通信 API を通じて連携サーバ B, C に検索要求を転送
6. 連携サーバ A : 検索 API を通じて検索システム A に検索要求を送信
7. 連携サーバ A : 検索 API を通じて検索システム A から検索結果 R_A を受信
8. 連携サーバ A : 通信 API を通じて転送した検索要求に対する検索結果 R_B, R_C を受信
9. 連携サーバ A : 転送／マージモジュールで R_B, R_C の優先度を算出
10. 連携サーバ A : 転送／マージモジュールで R_A, R_B, R_C を一つのリストに並び替え、最終的な検索結果 R を作成 (重複する結果があれば優先度の高いものを選択する)
11. 連携サーバ A : 通信 API を通じて、検索者に R を送信

本システムでは、検索要求の転送範囲については制限しない。すなわち、検索システム間リンクを末端までたどることにしている。また、同一の検索要求が複数送られてきた場合には、全ての検索要求に対して検索結果を返す。これは、検索要求を受信した連携サーバには、最終的にどの経路から来た検索要求に対する検索結果が採用されるのかを知る手段がなく、また、最初に到着した検索要求に対する検索結果が採用される保障もないためである。

このように全ての検索要求に対して検索結果を返すことで、システムの処理速度などの性能が悪化

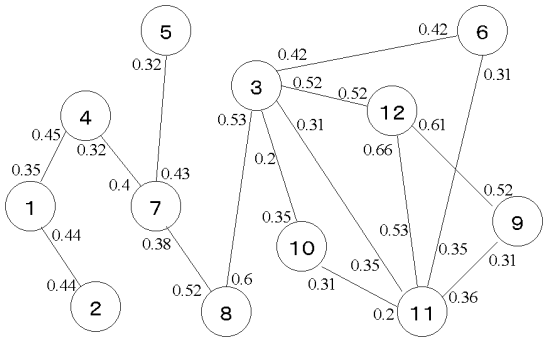

図 5: 検索システム間リンク

する可能性がある。しかし本稿の目的は、提案手法における検索精度の評価であるため、この点について、現時点では考慮していない。

・目的情報
映画「サウンド・オブ・ミュージック」の挿入歌である「エーデルワイス」に関する情報
・検索要求
エーデルワイス
・正解ページセット
○「サウンド・オブ・ミュージック」に関する情報のトップページ
○「サウンド・オブ・ミュージック」の挿入歌一覧が記述されたページ
○「エーデルワイス」の歌詞が記述されたページ
○「サウンド・オブ・ミュージック」に関する情報と感想が記述されたページ

図 6: 検索要求と正解ページセットの一例

5 評価実験

5.1 評価データ

ある個人や団体がインターネット上で公開している Web ページの集合を Web サイトと呼ぶ。Web サイトでは、Web サイトを作成している人の興味に沿った情報や、Web サイトを管理している団体の活動に関連する情報が公開されていることが多い。つまり、Web サイト内には個人の興味や団体の活動内容に応じた情報が存在しているといえる。そこで、提案手法の評価には、1 つの Web サイトを 1 つの情報源と見立て、Web サイト内の Web ページを個々の情報として実験を行った。

5.1.1 検索対象情報の作成

評価実験で利用する情報と情報源を以下の手順で作成した。まず、映画関連の内容を扱っている Web サイト を

1. 検索エンジンを用いて映画関連のリンク集のページを探す
2. リンク集のページからハイパーリンクされていて、トピックが明確な Web サイトを無作為に抽出する
3. 抽出した Web サイト間にハイパーリンクが 1 本でも存在する Web サイトのみを選択する

の手順で 12 サイト選択した。選択した Web サイトに含まれる Web ページ数は、それぞれ約 30 ページ から 約 7200 ページであった。また、全 Web サイトの Web ページ数の合計は約 10600 ページであった。

つぎに、検索システム間リンクと検索システム間のリンクの重みを設定した。検索システム間リンクの有無については、ある Web サイト内の Web ページから別の Web サイトの Web ページへのハイパーリンクが存在すれば、その二つの Web サイトの検索システム間にリンクを設定するという方法で決定した。

また、映画を趣味とする被験者が各 Web サイト内の Web ページと検索システム間リンクを見て、検索システム間リンクの重みを設定した。

その結果得られた検索システム間リンクは図 5 のようなものとなった。

5.1.2 検索要求と正解ページセット

評価用の検索要求と正解ページセットについても、検索システム間リンクのパラメータを決定した被験者が作成した。

まず、映画関連の情報で目的とする情報というものを決め、それを端的に表す検索キーワードで構成した検索要求を作成する。次に、評価用の情報に含

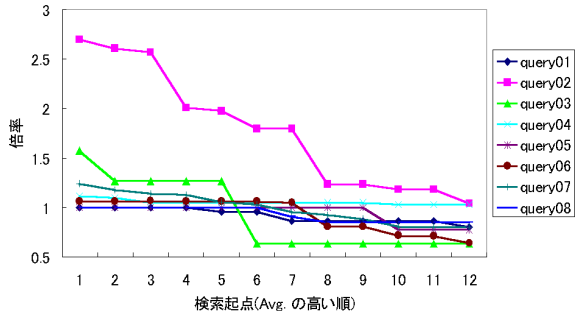

図 7: Avg. の倍率

まれる Web ページ内から目的に合う Web ページを正解ページセットとして手作業で抽出する。

このようにして検索要求と対応する正解ページセットの組を 8 組作成した。作成した検索要求と正解ページセットの一例を図 6 に示す。

5.2 実験結果と考察

5.2.1 連携型検索手法の効果

検索起点の指定と検索システム間リンクの重みの効果を調べるために、バケツリレー式の検索手法で、スコアの再計算と並び替えを行う場合 (提案手法) と行わない場合 (既存手法) の検索精度の比較を行った。

検索精度の比較には、検索要求毎の Interporated Recall-Precision [4] の 11 点平均 (以下 Avg. と呼ぶ) を用いた。Interporated Recall-Precision の 11 点平均とは、例えば再現率[1] が 0.10 での適合率[2] として 再現率 ≥ 0.1 の範囲での最大の適合率を用いるという方法で 11 点の適合率を求め、その平均を取ったものである。

提案手法では、全ての検索起点での検索結果から検索精度を算出した。既存手法では、どの検索起点を指定した場合も検索結果は同じであるので、一つの検索起点からの検索結果から検索精度を算出した。

[1] 全正解のうち検索結果に含まれている正解の割合を示し、

$$\text{再現率} = \frac{\text{検索結果中の正解件数}}{\text{全正解の件数}}$$ で算出する。

[2] 全検索結果のうち正解が含まれる割合を示し、

$$\text{適合率} = \frac{\text{検索結果中の正解件数}}{\text{検索結果の件数}}$$ で算出する。

図 7 は検索要求毎に、横軸に検索起点を Avg. の高い順に並べ、検索起点ごとの Avg. が既存方式の Avg. を 1 とした場合の何倍になっているかをグラフにしたものである。グラフから、最善の検索起点を指定することにより、既存手法と同等かそれ以上の検索精度が得られることが分かる。また、最善の検索起点以外でもいくつかの検索起点では既存手法と同等か上回る検索精度が得られている。しかし、検索起点の指定が適切でない場合には、検索精度が既存手法を下回ってしまう。

次に、検索起点によって検索精度が既存手法を上回る場合と下回る場合が生じる原因を調べるために、検索結果の件数が比較的多く、既存手法を上回る場合と下回る場合が起こっている query07 における、既存手法と提案手法の検索起点毎の Interporated Recall-Precision を比較する。図 8 は query07 における Interporated Recall-Precision のグラフである。グラフから、的確な検索起点を指定した場合には、再現率が低い部分での適合率が高くなっており、逆に、検索起点の指定が適切でない場合には、再現率が低い部分での適合率が低くなっている。これについては、他の検索要求でも同様の傾向が見られた。

これらの結果から、提案手法について以下のことが言える。提案手法では最善の検索起点を指定することで既存手法よりも検索精度を向上させることができる。また、サイト内検索のように最善のサイトを必ずしも指定する必要がなく、検索システム間リンクのネットワーク上で最善の検索起点の近くに存在する検索システムを検索起点に指定することで、既存手法と同等か上回る検索精度が得られる。また、適切な検索起点の指定を行うことで低再現率における適合率が高くなるという結果から、本手法は適切な検索起点を指定することで検索結果の上位における検索精度を向上させることができる。

5.2.2 提案方式の適用領域

次に、提案手法がどのような状況で有用なのかを検証するために、評価に用いた検索要求と正解ページセットについての分析を行った。

図 9 のグラフは、正解ページを最も多く含む Web サイトを検索起点とした場合の、ホップ数と正解ページ数の関係を、検索起点とした Web サイトのホップ数を 1 として示したものである。

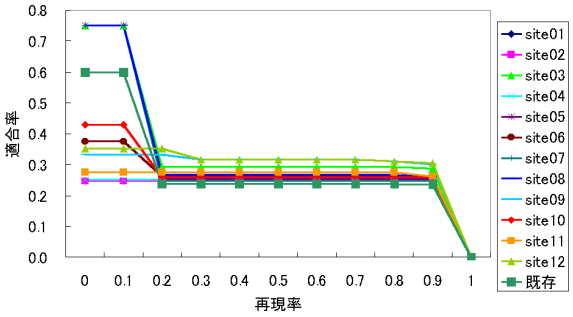

図 8: query07 の Interporated Recall-Precision

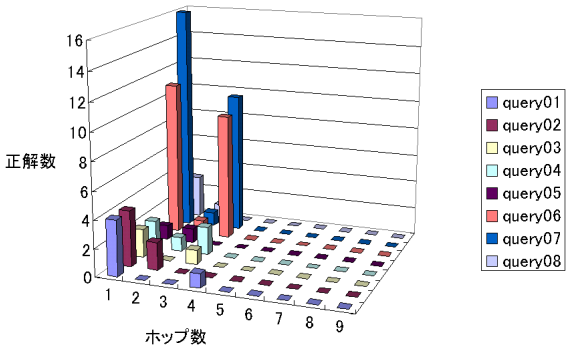

図 9: 正解ページの分布

このグラフから、全ての検索要求において正解ページが一定のホップ数内の Web サイトに存在していることが分かる。つまり、正解ページが全ての Web サイトに存在しているわけではなく、ある特定の Web サイトの周辺に集中しているといえる。

提案手法では、検索結果の優先度として経由した検索システム間リンクの重みの積を用いている。そのため、多くの場合で経由した検索システム間リンクの数が少ないほど、すなわち、検索起点の Web サイトに近い Web サイトの検索結果ほど最終的な検索結果では上位にランクされる。

このことから、提案手法では検索者が必要としている情報がいくつかの情報源にのみ偏って存在していて、これらの情報源が検索システム間リンクのネットワーク上で近くに存在している状況で有用である。

6 まとめと今後の課題

6.1 まとめ

情報源に含まれる情報の内容の偏りと、検索者や情報源の情報提供者が持つ所望の情報や関連情報がどの情報源に存在するかという知識を利用することで、関連情報を一括検索することができる連携型検索手法を提案し、評価実験を行った。評価実験の結果、検索起点を適切に指定することで、既存手法よりも高い検索精度を得ることが可能となることが確認できた。さらに、検索要求に対する正解がいくつかの情報源に集中しており、それらの情報源が検索システム間リンクのネットワーク上で近い位置に存在している状況において、提案手法が有用であることを示した。

6.2 今後の課題

リンク重みの決定方法について、今回は被験者に一存する形で決定したが、この決定方法についての検討を行っていく必要がある。また、検索起点の指定が検索精度に大きく影響することから、検索者の検索起点の指定をサポートする方法についても検討を行っていく必要がある。提案した検索方式の適用範囲が実世界においてどの程度存在しているのかについても、さらに大規模なデータセットを利用した評価実験を行い、明確化を行っていこうと考えている。

参考文献

[1] Recardo Baeza-Yates, Berthier Ribeiro-Neto, “Modern Information Retrieval”, Addison-Wesey, 1999

[2] Jnutella, “Jnutella.org”, http://www.jnutella.org/

[3] 富田 準二, “グラフモデルと XML 検索エンジンを用いた高度情報アクセス”, ACM SIGMOD 日本支部 第 19 回大会, 2001

[4] IREX 実行委員会, “IR 評価結果”, IREX ワークショップ予稿集 p.244, 1999