WWW検索のための情報収集技術の開発

5L－2

○谷田望，石川浩通，須賀田裕臣

三菱電機株式会社　情報技術総合研究所

１．はじめに

近年のインターネットでの WWW(World Wide Web)の発展はめざましいものがあり、それに伴い WWW 情報の検索サービスが必要となっている。そこで、我々はこのサービスを実現するための検索システムを開発した。この WWW 情報検索システムは大きく分けて情報収集技術と検索技術[1]により実現されている。本稿では、そのうちの情報収集技術について報告する。

WWW に提供されている情報の量は爆発的に増え、どこに何があるかを知るための検索サービスが多数(Lycos[2],Yahoo[3]等)出現した。これらのサービスは多くのユーザーにとって便利だが、WWW 全体の情報量が増えてくるにつれ、以下のような問題がでてきた。

①検索に工夫が必要

検索に対するヒット数が多すぎる。陳腐語では何の情報も得られない位だが、逆に絞込みに多項間論理積をとると、何もヒットしなくなることも多い。

②データ更新を捕捉できない

検索した結果みつかった URL が既に移動・消滅しており、情報がなくなっている場合が多々ある。

これらの問題は収録情報のボリュームと雑多性に起因するといえ、解決するには検索対象を特定のものに絞り収集情報の量を適切なレベルに抑制、WWW サーバの頻繁なデータ更新に機敏に対応する必要がある。そこで収集情報を特定内容のものに限定し、更新収集を効率的に行う等の特長を持つ情報収集 S/W(ロボット)を開発した。

２．システムの概要

このロボットは WWW 上の情報を収集した際にその情報からのハイパーリンクを解析し、収集情報にリンクされた他の情報を再帰的に収集するものである。現在この種のロボットの基本的な技術は既に確立されているといってよい。しかし、以下の問題を解決する技術についてはまだ確立されたとはいえない状況にある。

①特定分野の情報のみを必要量だけ収集するには人間が情報を選別する以外には良策はなく、これを自動的に行うことは難しい。

②WWW ではよく通信リクエスト拒絶が発生するが、それが一時的なものか、また多数のマシンのどこが異常かはよく分からない。このため異常な収集情報をユーザに通知したり、自動的に削除するのは難しい。

③サーバの情報更新状況を監視し、一度収集した情報を効率よく更新することは、いろいろなサーバがあって更新状況が確実に分からないことや、通信の異常などが有り得るので、難しい。

今回のロボットでは、このような問題を解決するために、次のような特長をもつ技術を開発した。

２．１収集範囲制限

特定分野の情報だけを効率よく収集するには、収集開始 URL からたどる範囲を何らかの方法で制限するのがよい。今回は URL アドレスで制限することとした。

２．１．１ドメイン・パス制限

ロボットは収集開始 URL(Root URL)を幾つか指定すると、そこからリンクされた URL を収集する。但し収集範囲は、Root URL と同じドメイン名・パス名をもつものに制限する。また、ファイルタイプによる収集制限も行う。

ここでいうドメインとは、あるマシンの IP アドレス又はそれにつけられた名称と、通信の際のプロトコル名を意味する。例えば、http://A/B/C/D.html なるアドレスがあったとすると、http://A/までがドメイン名である。

またパスとは、そのマシンでの WWW サービスのルートディレクトリからのパスのことであり、この例では http://A/以下の/B/C/がパス名となる。

収集する対象を Root URL と同じドメイン・パスを持つ URL のみに制限するのは強力な収集範囲指定方法であり、これによってユーザが意図しない URL を収集することを防ぐことが容易に行える。

ファイルタイプによる収集制限は、URL アドレスの拡張子をみて行われる。今回のロボットでは、html,htm,txt 及び拡張子なしを収集対象とした。

２．１．２非収集パス

ユーザが収集した範囲の情報を見て、不要と思えば収集範囲から外すことができる。これは Root URL ごとに非収集パスというものを設け、そのパスに指定された文字列を URL アドレス中にもつものは収集対象外とすることで、任意 URL を収集範囲から外せるようにして実現した。

２．２異常収集情報排除

HTTP で通信するとサーバの送るデータにレスポンスヘッダと呼ばれる情報があり、これに含まれるコード番号より通信異常を認識できる。しかし異常はサーバのファイルの一時的な移動や収集側のプロクシサーバの異常等によっても起こり得るため、コードの判断には注意が必要である。今回はこのコード情報を次のように判断した。

Developmetnt of an information gathering technology for retrieval systems of WWW

Nozomu Tanida, Hiroyuki Ishikawa and Hiro'omi Sugata Mitsubishi Electric Corp., Information Technology R&D Center

①ヘッダのコード番号 20x, 30x → 正常
②ヘッダのコード番号 40x, 50x → 異常

そして、異常な場合には情報を収集せず、さらに異常が連続するような場合には収集対象からも除外するようにした。但し、異常の程度に応じてその評価値を変えてある。これにより異常な情報を検索インデクスとしないようにすることを可能とした。

2．3更新収集

更新収集を効率的に行うには過去の収集結果をできる限り利用する必要がある。今回は以下の方法をとった。

①前回収集時刻から設定した収集更新間隔を過ぎていない URL は更新有無等を調べず、収集もしない。
②リクエストメソッドとして HEAD を用いて情報の更新時刻取得を行い、サーバ更新時刻の変更されていないものは収集しない。
③サーバ更新時刻が比較できない場合には GET メソッドで収集を行うが、前回収集時に対してファイルサイズのかわっていないものは更新されていないと判断し、検索インデクスを作り直さない。

更新収集間隔はユーザが任意に決めることができ、サーバによるデータの更新率を考慮しながら決定することにより、効率的な収集が可能となる。また更新収集は検索インデクス作成にも関係がある。収集後に更新されていない情報を判断することにより、インデクス作成のための処理を抑制させることができる。

3．収集結果と考察

3．1 収集制限

開発ロボットを用いて収集実験を行った。医療関係の6つのサーバのトップアドレスを Root URL としたが、ドメイン・パス等の収集範囲制限機能により、収集情報の内容は殆ど全て医療関係となった。初期収集量は 30,000URL 程度で容量総計は約 80Mbyte である。初期収集時のダウンロード速度は平均 1.27URL/sec(3.00kbyte/sec)で、総収集時間は 6.6 時間という結果であった。

ここでの問題としては、収集範囲の設定をよりきめ細かく行おうとした場合、非収集パスの設定などに手間がかかるということがあげられる。

3．2 収集異常の検出

初期収集の後、この範囲の収集を繰り返した。この時ロボットが検出した収集異常 URL の全体に占める割合が実験開始からどう変化したかを各サーバ別にまとめた結果を図 1 に示す。図 1 では収集異常を出したサーバは A,B,D,E である。ここで A と B はサーバの全情報の 10% 程度が途中から異常になっており、一部情報のアドレスが途中で消滅したと判断できる。また D と E は異常率が極端に変動することより不安定な通信状態が続くサーバであると判断できる。

ここでは、異常な収集結果を出したサーバが最初予想していた以上に多いことが分かった。収集異常の量が極端に多く、さらにそれが激しく変動するようなサーバは安定的に収集できないものであり、できれば収集対象から外した方がよいと考えられる。また、途中で一定量の異常が定常的に発生するサーバはその部分の情報の更新がなされている可能性が高く、このサーバには非収集パスを設けた方がよいと考えられる。

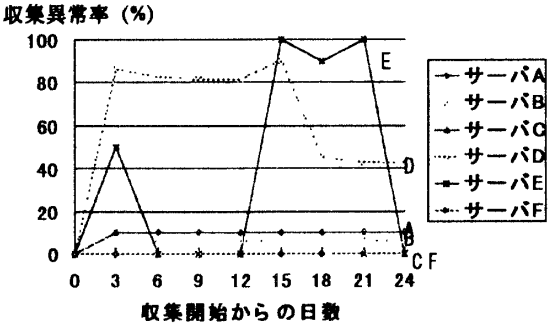

図1: 各サーバの収集異常率推移

3．3 更新収集

更新収集時に 10URL 程度のサーバだけを随時収集にし、残りの収集間隔を 10 日間に設定した。するとこの収集間隔の間での収集では新規収集で 6.6 時間要した収集範囲を 600 秒で更新した。しかしこの収集間隔が明けると、更新有無を確かめて収集し直すため 14,000 秒を要した。

一方これらの更新収集において、実際に意味のある収集 URL 数は、一貫して数十～数百程度であった。このことより、サーバごとに更新収集間隔を適正に設定することで、提供される情報の更新に遅延することなく、収集時間を減らすことが可能であることがわかった。

10 日の更新収集期間が明けた時に 14,000 秒も要したのは、gopher サーバの URL が多かった(91%)ためと思われる（gopher サーバの情報は更新の有無が不明なことが多く、この場合必ず収集し直しが必要である）。http のサーバを主とすれば状況は変わると考えられる。

4．まとめ

WWW 検索のための情報収集技術を開発した。開発ロボットの収集選択は完全に自動的とはいえないが、収集ユニットと非収集パスの概念により、収集対象を目的の情報に絞るのが容易である。また、収集異常を起こした URL の情報は収集しないようにするので、通信エラーの頻発するサーバを容易に判定できる。さらに、更新収集時には収集間隔の設定などにより、既収集の範囲に対しては新規収集に比して少量のデータを収集するだけで済む。

今後の課題としては、収集間隔の自動設定や、サーバタイプに応じて If-Modified-Since の条件つき GET を適用するなどの処理を加えることがあげられる。

参考文献

[1]宮井,徳永,"WWW 情報検索システムにおける検索支援技術の開発," 情処学会 第 54 回全国大会, 5L-01, 1997.
[2]Lycos → http://www.lycos.com/
[3]Yahoo → http://www.yahoo.com/