3T－4

大規模なWWW全文検索エンジンにおけるIPアドレスによる索引

黒田 洋介　村岡 洋一

早稲田大学　理工学部

1. はじめに

本稿では、WWW検索エンジンにおける計算機負荷およびネットワーク負荷を減らすためのアルゴリズムの一手法について述べる。

本アルゴリズムはWWWロボットが収集したすべてのURLに対し、

- ホスト名をIPアドレスのドット表記に変換する
- ポート名を明記形で記述する。

といった変形を加えることにより、一意性を持つURLの標準形を導入する。これによりWWWロボットが収集すべきURLの重複を除去することができ、ネットワーク負荷および検索にかかる計算負荷を削減できる。また、このURLの標準形を利用した索引でもURL限定検索が可能である。

2. URLにおけるホスト名、ポート番号

現状のURLにおいては、あるWWWサーバに対するホスト名は、IPアドレスのドット表記, DNSによる本名の絶対ドメイン名, 本名の相対ドメイン名, DNSによる別名の絶対ドメイン名, 別名の相対ドメイン名の5種類が存在する[表1]。

別名の絶対ドメイン名	http://www.waseda.ac.jp./path
別名の相対ドメイン名	http://www.waseda.ac.jp/path
本名の絶対ドメイン名	http://sun.waseda.ac.jp./path
本名の相対ドメイン名	http://sun.waseda.ac.jp/path
IPアドレスのドット表記	http://133.9.68.42/path

表1: 5種類のホスト名

さらにポート番号表記についても、省略形と明記形の2種類が存在する．これはポート番号が80の際にはURL中で省略可能であることによる．

[表2]。

省略形	http://www.waseda.ac.jp./path
明記形	http://www.waseda.ac.jp.:80/path

表2: 2種類のポート番号表記

以上のように同一のページに対するURLは多数存在してしまうが、それぞれのURLに対して、ホスト名をIPアドレスのドット表記に変換し、ポート名を明記するように変換すれば、同一ページを指すURLが一意にさだまる。本稿ではこれをURLの標準形とよぶ。

[0]A WWW search engine index represented by IP address
Yosuke Kuroda, Yoichi Muraoka
School of Science and Engineering,Waseda University

WWWロボットが収集したURLをすべてこの標準形に変換することで、URLの重複を削除でき、WWWロボットが収集すべきページの数および検索エンジンがつかう索引の大きさを小さくできる。

3. URL限定検索

URL限定検索とは特定のWWWサーバのページのみを対象にしたいときなどに用いる検索であり、多くのWWW検索エンジンで「HTMLのURLがURL限定式と中間一致するページのみを対象とする検索」と定義されている。

ホームページを公開する多くの企業や団体では、検索対象を自らが作成したページに限定した独自の検索エンジンを作成することが多い。しかし、検索エンジンの作成は非常に手間がかかるため、jpドメインなどを対象にする検索エンジンが、それらの業務をURL限定検索機能により代行できるようにすることが望まれている。

従来の検索エンジンでは収集すべきURLの数を減らすため、IPアドレスのURLや本名のURLには、ページ収集をしないことも多い。しかしこの方法では、URL限定検索においてIPアドレスや本名を条件にした場合に、正しい検索ができないなどの問題が発生する。

4. IPアドレス表記索引とURL限定検索

IPアドレス表記を用いた索引に対するURL限定検索では、索引中のURLのIPアドレス表記の本名表記, 別名表記への展開、およびポート番号の省略表現への展開[表3]をおこなったのち、展開後のURLそれぞれについて、URL限定式に中間一致で適合するかを判断するのが望ましい。

索引中のURL	展開後のURL
http://133.9.68.42:80/path	http://133.9.68.42:80/path http://133.9.68.42/path http://sun.waseda.ac.jp.:80/path http://sun.waseda.ac.jp./path http://www.waseda.ac.jp.:80/path http://www.waseda.ac.jp./path

表3: URLの展開

しかし、URL限定検索の定義を変更することで展開処理を省略することができる。本稿では、「下記の例外を除いて、HTMLのURLがURL限定式と前方一致するページのみを対象とする検索」と定義する。

例外として、URL 限定式において、*http://* の後の一文字がアルファベットでかつ *http://* の後にスラッシュもコロンもないときは、「*http://* の後をホスト名と考え、ホストが完全一致するページのみを対象とする検索」と定義する。

以上の定義によって、URL 限定式の適合を調べた例を表4に示す。

URL 限定式	比較対象 URL	適合する
http://133.9.68.42/	http://133.9.68.42/path	O
http://133.9	http://133.9.68.42/path	O
http://abc.	http://abc./path	O
http://abc	http://abc.jp./path	X

表 4: URL 限定検索の前方一致による定義

このように定義することで、URL 限定式におけるホスト名を IP アドレスに変形しても等価となる。表5 に abc の IP アドレスを 1.2.3.4 としたときの例を示す。

変形前の URL 限定式	IP アドレス表記への変形後
http://abc./path	http://1.2.3.4:80/path
http://abc.:80/path	http://1.2.3.4:80/path
http://abc.:8	http://1.2.3.4:8
http://abc.:	http://1.2.3.4:
http://1.2.3.4	http://1.2.3.4
http://133.9	http://133.9
http://abc.	http://1.2.3.4

表 5: URL 限定式の IP アドレス表記への変形

URL 限定式が IP アドレス表記になれば、IP アドレス URL の本名 URL, 別名 URL への展開処理も不要になり、URL 適合判断処理の計算量も減らすことができる。表3の例では、適合判断処理の対象となる URL を 1/6 に減らすことができる。

5. DNS のラウンドロビンにおける問題点

アクセスの多い WWW サーバでは、負荷分散のためミラーサーバを複数用意し、それらに均等に負荷をかけるため、DNS においてラウンドロビンによるアドレス解決を行なうことがある。

例えば、*www.nifty.ne.jp.* は複数のミラーサーバから構成されており、アクセスするたびに IP アドレスが 192.47.24.74 → 192.47.24.78 → 192.47.24.79 → 202.219.63.91 → 202.219.63.92 → 192.47.24.74 とかわる。

このような WWW サーバの URL に IP アドレスによる URL の標準形を使うと、収集すべき URL がミラーサーバの数に比例して増加してしまう。

しかし、それに比例してユーザに提供できる URL の数が増加するので、全くの無駄というわけではない。また、ユーザはそれぞれのミラーサーバに対しアクセスをすることで、速度や信頼性を選択することができるようになる。た、ミラーページの検索処理への負担を減らす方法はすでに研究されており、この URL の増加はそれほど問題にはならない。

6. バーチャルホストにおける問題

バーチャルホストとは、複数の小規模 WWW サーバを 1 つの IP アドレスで動かすための機能である。

HTTP1.1 では Host メッセージヘッダを追加することで、1 つの IP アドレス上に複数の WWW サーバを設置できるようになった。このため、この機能を利用している WWW サーバ上の URL は、本稿の URL の標準形では同一かどうかの判断ができない。

しかし、この機能を利用するサーバはまだ少数で小規模であるので、これらの WWW サーバを無視しても影響は少ないと考えられる。

7. ユーザのための URL 表示

WWW 検索エンジンのユーザは、ホームページの品質等についてそれを作成した組織の名前から判断することが多い。そのため、検索結果の URL は IP アドレスではなく、ドメイン名によって表現されることが望まれる。

検索結果の URL にたいして DNS の逆引きをおこない、可能ならば IP アドレスをドメイン名に変換ののち検索結果を表示するという方法が簡単である。

8. おわりに

検索エンジン Ringring 上に以上の機能を実装し、正しく動作することを確認した。

Ringring は jp ドメインの検索エンジンとしてだけではなく、本稿の URL 限定検索機能を使って、早稲田大学などの検索エンジンとしても利用されている。

今後は、バーチャルホストにも対応できるよう、URL 標準形の拡張を考えていく予定である。

参考文献

[1] P. Mockapetris: *Domain Names - Concepts and Facilities*, RFC-1034,Nov,1987

[2] P. Mockapetris: *Domain Names - Implementation and Specification*, RFC-1035,Nov,1987

[3] T. Berners-Lee,CERN,L. Masinter,Xerox Corporation,M. McCahill,University of Minnesota: *Uniform Resource Locators (URL)*, RFC-1738,Dec,1994

[4] T. Berners-Lee, MIT/LCS, R.Fielding, UC Irvine, H.Frystyk: *Hypertext Transfer Protocol - HTTP/1.0*, RFC-1945,May,1996

[5] R.Fielding, UC Irvine, J.Gettys, J.Mogul, DEC, H.Frystyk, T.Berners-Lee, MIT/LCS: *Hypertext Transfer Protocol - HTTP/1.1*, RFC-2068,Jan,1997