# LM-5　WWWページ検索結果の選択における利用者支援

## User assistance of choosing results of retrieving WWW pages

柊 和佑†　Wasuke Hiiragi
阪口 哲男†　Tetsuo Sakaguchi

## 1．背景

インターネットという言葉が聞かれない日はない現在、ネットワーク上に存在する情報も、量・種類ともに増加してきている。個人による情報発信に加え、様々な企業が行う情報サービスまで、膨大な量の World Wide Web(WWW) ページが存在する。そのような状況では、目的の情報を得るだけでも大変な作業であり、WWW ページの検索サービスの重要性は情報量に比例して大きくなっている。しかしながら、検索サービスを利用してもインターネットから目的の情報を引き出すことは容易ではない。なぜなら、検索サービスを利用しても目的の情報を含む WWW ページにたどり着けないことがあるからである。

対象数が膨大なため、最新の WWW ページを検索サービスで使われているデータベースに即時に反映させることは困難である。そのため検索結果には、すでに無関係な内容に更新されていたり、存在していない WWW ページも少なからず混ざっている。このような状態はいわゆるリンク切れと呼ばれ、目的の WWW ページを得にくくする原因となっている。

本研究の目的は、検索サービス利用時に表示される WWW ページ検索結果について、それらを選択する利用者の支援を行うシステムの構築である。具体的には、検索サービスにおいて検索効率を低下させる要因となるリンク切れに注目し、システムを介した履歴データの共有によってその選択を回避することである。

## 2．既存のリンク切れ対処方式

「Yahoo!」[1]等にみられる人手による登録型検索サービスでは、リンク切れの報告用フォームを用意しておくのが一般的である。検索結果一覧からすぐに報告用フォームへアクセスできるようになっていることが多い。しかし、強制力はなく利用者の善意に頼ったシステムとなっている。そのため、検索結果の中にリンク切れを発見した利用者が、それを必ず検索サービス運営側に報告するとは限らない。さらに、報告に沿って修正や更新を行うのも、検索サービス運営側の仕事となっている。また、検索サービスによっては検索結果一覧からは該当サービスに直接リンクせず、検索サービス内の proxy サーバ等を経由することでリンク先の状態等をチェックできるように作られているところもある。しかし、その方法では検索結果一覧以外のブラウジングでは、リンク切れ WWW ページを検出することはできない。

ロボット型検索サービスである「Google」[2]ではキャッシュシステムを採用している。キャッシュシステムとは、該当 URL をロボットによって記録した時点の HTML データを保存し、キャッシュとして公開しているものである。これにより、すでに WWW ページの状態が変わっていたとしても、確実に検索結果の含まれる WWW ページを参照できる。しかしながら、キャッシュシステムは、検索結果と同じ時点の WWW ページを表示するためのシステムであり、最新の内容を閲覧する目的には合致しない。

リンク切れを発見することを目的とした「りんくの道しるべ」[3] や「孤島発見器」[4]のようなツール群も存在する。これは、WWW ブラウザ上に表示しているリンクを、リアルタイムに検査するというものである。一般的な WWW ページのリンク先を検査するという目的の他に、個人で作成した WWW ページのリンク切れを検査し修正する等の目的に使用されるものである。リンク先の WWW ページが存在している場合と、存在していない場合それぞれのマークを付加する、というシステムが標準的である。逐一リンク先に問い合わせ、その結果を元に判別しているため応答時間がかかる。そのため、本研究の目的のような対話型検索における利用者支援には不向きと考えられる。

## 3．WWWページの状態検出と利用者の支援

本研究は以下の考えに基づいてシステムを構築し、検索サービスを利用することにより提示される WWW ページの検索結果を選択する際の利用者支援を行う。

利用者個人のブラウジングによって検出されたリンク切れを含む履歴データを、自動的にデータベースに蓄積し、その後の利用者が検索サービスを利用する際に、そのデータを参照することでリンク切れであることを明示する。それにより、利用者は無駄なリンクをたどる頻度が低くなり、検索作業の効率をあげることができると考えられる。

本システムは、利用者のブラウジング作業の結果を利用するものであるが、ブラウザが自動的にリンク切れを処理するため、利用者への負担にはならない。また、WWW 検索システムは既存の検索サービスを利用し、本システム自体は履歴データの蓄積・利用を行うだけである。ページの有無についての判断は、HTTP 規格のステータスライン中のステータスコードで行っている。[5][6]

利用者側に対しては検索サービスを利用した際に、図1のように「リンク切れマーク」を検索結果一覧に付与する。

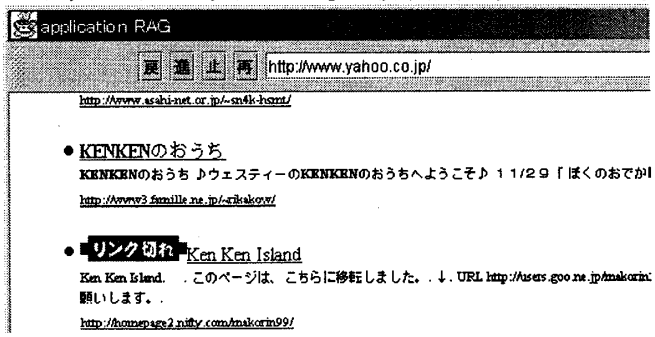


図1　本システムの動作画面
（リンク切れの提示）

† 図書館情報大学

これにより事前に、過去にリンク切れになっていたWWWページを示すことができ、利用者は無駄にリンク切れを辿らなくてすむようになる。もちろん、リンク切れマークのついたURLにアクセスすることも可能であるが、それは利用者の判断に任される。

## 4．実現環境

本システムはクライアントである専用ブラウザと、蓄積用サーバプログラムと、検索用サーバプログラム、データを蓄積するためのデータベースで構成されている。

クライアントはOSにWindows2000、開発はJAVA2 SDK 1.3.1を使用した。サーバプログラムはOSにUNIXのFreeBSD4.4、開発はJDK 1.1.8を使用し、DBMSにはPostgreSQL[7]を使用した。今回、本システムは「Yahoo! Japan」を対象にして構築した。

## 5．システムの実現

データベース内に蓄積する履歴データは図2の通りである。URLをキーとしてステータスコードと日時と専用ブラウザの動作している計算機のIPアドレスを記録している。

| URL | ステータスコード | 日時 | IPアドレス |
|---|---|---|---|
| http://www.ulis.ac.jp/jindex.html | 200 | 2002-1-30 20:21:12 | 133.***.***.*** |
| http://www.abcd.com/ | 404 | 2002-1-30 19:56:51 | 133.***.***.*** |
| …… | …… | …… | …… |

図2　履歴データの例

実際に専用ブラウザによってリンク切れが検出されると、その時点までの専用ブラウザが表示し、記録しておいたURLの履歴データを一括して蓄積用サーバに転送する。

サーバ側ではそれぞれのURLを比べ、蓄積されていないデータは新たに加え、すでに蓄積されているものに関しては日時を比べてより新しいのものみを更新・蓄積する。その後クライアント側はそれまでの履歴データを削除する。

これにより、検出されたページ状態が蓄積されていく。また、全ての履歴データを比べ、時間によって蓄積するデータを選別しているため、データベース中のリンク切れが解消されていた場合にも対応することができる。その場合は、データベース中でリンク切れになっているURLについて、ステータスコード、日時、IPアドレスを上書きする。しかしながら、リンク切れの解消については即座にデータベースに反映させることができてはいない。

次に図3にシステムの構成を示す。なお図3はリンク切れ発見時の動作を中心に記述してある。複数の利用者が専用ブラウザを使用してWWWページを閲覧する。その際、リンク切れを発見するとブラウザは自動的に、URL等の履歴データを蓄積用サーバに送り、サーバはそれを蓄積する。

蓄積用サーバに蓄積された履歴データは、その後の利用者が専用ブラウザで既存の検索サービスを利用したときに参照される。実際には、検索サービスから送り返されてきた検索結果一覧からURLを抽出し、検索用サーバプログラムに送信する。検索用サーバプログラムは、送信されたURLと蓄積した履歴データとを照合する。照合した結果、対応したURLの履歴データが蓄積されていれば、そのステータスコードだけが専用ブラウザに送られる。

専用ブラウザはサーバから返信されてきたステータスコードを基に、検索結果一覧にリンク切れマークを付加して利用者に提示する。

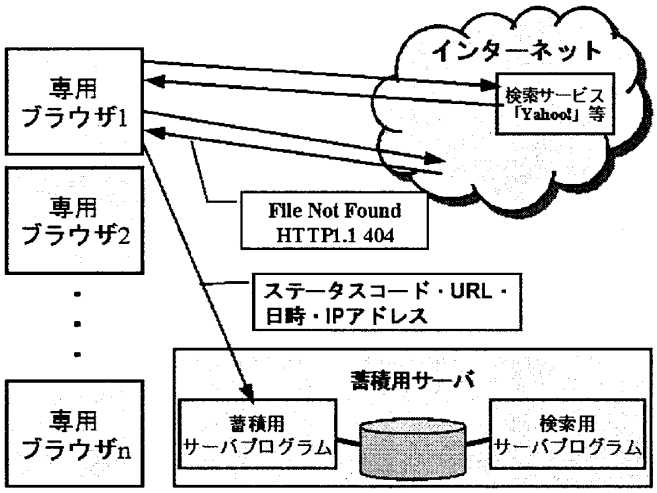


図3　システムの構成

## 6．おわりに

今回の研究で構築したシステムを使用することにより、検索サービスにあったリンク切れを前もって確認することができた。これにより、このシステムはリンク切れを回避する利用者支援の手法のシステム化について、成果を上げたと思われる。また、本システムでは検索結果一覧のうち、上位のものであるほどリンク切れが即座にデータベースに反映されやすく、下位のものほど反映されずらくなっている。しかし、一般的に検索結果の下位は目的の情報とは関係が薄く、利用者も参照することが少ないため、問題はないと考える。また、手法上の問題として必ず一回は利用者がリンク切れに遭遇する必要があり、それは回避することができない。しかしながら、本システムはそのような利用者が増えることにより網羅性が上ることになり、後の多くの利用者の利便性向上に繋がるのである。

## 参考文献

[1]. Yahoo Japan Corporation. Yahoo! JAPAN. <http://www.yahoo.co.jp/>
[2]. Google. Google. <http://www.google.com/>
[3]. Yuma. りんくの道しるべ ver0.0.4. <http://i.am/mapper64>
[4]. 桜井 博志. 孤島発見器 ver2.755 <http://hp.vector.co.jp/authors/VA014575/>
[5]. H-Hash@StudyingHTTP.NET. Hypertext Transfer Protocol -- HTTP/1.1. < http://www.studyinghttp.net/rfc_ja/2616/rfc2616_ja.html >
[6]. The Internet Society. Hypertext Transfer Protocol -- HTTP/1.1. <http://WWW.ietf.org/rfc/rfc2616.txt>
[7]. PostgreSQL. <http://WWW.postgresql.org/>