統合検索のための検索結果件数推定方法

2V−6

古賀 康則† 酒井 美由紀† 廣川 佐千男‡
†九州大学システム情報科学府 ‡九州大学情報基盤センター

1 はじめに

Web上の莫大な情報は必要な情報の発見を困難にし、検索結果の品質が問題となっている。Yahoo!やGoogle等の検索エンジンがWeb全体を対象とした検索機能を提供するのに対して、自サイト内など限られた領域を検索対象にしたサイトがある。本論文ではこれらを検索サイトと呼ぶ。対象が限定されているため、検索結果は品質はよい。検索サイトの数は増えており、10万以上の検索サイトがあるという報告もある[1]。

検索サイト群を効率よく利用するためには、それらの統合が必要であり、統合のためには個別のサイトに特化したラッパーが必要になる。しかし、増え続ける検索サイトに対し、個別のラッパーを人手で実装することは困難で、ラッパー生成の自動化が必須である。我々は自動的なラッパー生成による統合検索サイト構築の研究を行っている[4, 3, 1]。

本論文では検索サイトの返す検索結果の件数を、ファイルサイズやリンク数によって推定する方法を提案する。また具体的な23件の検索サイトについて、ファイルサイズやリンク数と検索結果件数の相関を回帰分析により調べ、検索結果件数推定方法の有効性と問題点を示す。

キーワードに対する検索結果件数は、[3]で提案したラッパーの自動生成法の精度評価や改良に利用できる。また、[2]のように検索サイトの特徴情報にも利用可能であるが、[2]では結果件数の抽出方法は述べられていない。

Automatic evaluation of the number of search results
Y. Koga†, M. Sakai†,S. Hirokawa‡
† The Graduate School of Information Science and Electrical Engineering, Kyushu University
‡ The Computing and Communications Center of Kyushu University
hirokawa@cc.kyushu-u.ac.jp

[1]http://www.completeplanet.com/

2 検索結果件数の推定方法

検索結果の1件1件は検索結果のタイトルや、日付情報、要約などを含んでいるが、サイトが同じ場合はほとんど同じフォーマットをしている。また結果のページへリンクを張っていることが多い。さまざまなキーワードで検索を行い、十分多く検索結果のファイルを準備できれば、検索結果に含まれる結果の件数を推定できる。

ファイルサイズ方式

検索結果1件分のデータサイズをa, 検索結果0件の時のファイルサイズをbとすると、x件の結果を含むファイルサイズyは$y = ax + b$と表される。従って、a, bが求まればファイルサイズyから結果件数xを求めることができる。

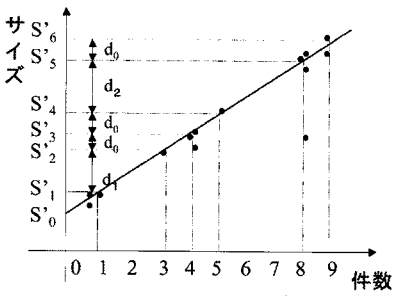

図 1: 検索結果に現れる1件分のサイズの推定

ここではn個の個別のキーワードに対して、n個の検索結果のファイルが既に得られていると仮定する。簡単のために、n個の検索結果の件数が全て異なっていると仮定し、各ファイルのサイズを$s_i(1 \leq i \leq n)$とする。これらを昇順にソートした後のサイズをs'_iとすると、$s'_i \leq s'_{i+1}$が成立している。$d_i = s'_{i+1} - s'_i(1 \leq i < n)$とした差分$d_i$の最小値dが結果1件分のデータサイズである可能性がある。これが正しいかどうかは任意のd_iがdで割り切れるかどうかで判断できる。割り切れない場合はdがk件分の差分であると考え、1件分のデータサイズをd/k(k=2,3,..)として割り切れる場合を

探索する。データの種類 n を十分大きく取れば、結果件数0件のものがあり、差分の最小値は1件分となると仮定できる。このとき、$a = d, b = min(s_i)$ と推定できる。

リンク数方式

検索結果は通常、結果のページへのリンクを含んでいるため、結果の件数によってリンクの数が異なる。従って、リンクの数を数え、ファイル毎の差分をとることによってファイルサイズの時と同様な方法で結果1件あたりのリンク数を求めることができる。

3 実験と考察

検索サイト23件について、CompletePlanet [2] のディレクトリから133の単語をキーワードとして選んで検索を行い、個別のファイルに対して提案した方法の検証を行った。実際の件数を調べ、件数とファイルサイズ、件数とリンク数をプロットし、単回帰分析の寄与率を調査した。なお、実験では同じ件数の検索結果が出てくるため、ファイルサイズの差分が、ファイル自体の大きさに比べ非常に小さな場合は同じ件数とみなし、ファイルサイズの平均をとっている。図2、3はhttp://www.clayzee.com/における実験結果のグラフである。このサイトではファイルサイズ方式ではa=715.213、b=1959.35、寄与率は0.9965となり、リンク方式ではa=1.9090、b=6.6363、寄与率は0.9932であった。ほとんどのサイトについて、ファイルサイズ、リンク数ともに件数の一次式で近似できることが確認できた。2件のサイトについては、検索結果の件数が2通りしかなく自明となった。また、検索結果が複数のページに渡っていて、次のページへのリンクが多数ありそれらの除去の処理で例外が起こり、リンク数による a の推定結果が0となる場合があった。しかし、検索結果部分以外に変動のあるリンクをうまく除けた場合には、寄与率が1となり完全な件数推定ができたものもある。全体としてはデータサイズの方が優れた結果となった。この結果から検索結果件数の推定はファイルサイズから十分可能であり、リンク数による手法には改善によってより高い精度で件数を推定できる可能性があることが分かった。

4 まとめと今後の課題

本論文では検索サイトのラッパーの精度評価に必要な、検索結果推定方法について2つの方法を提案した。また、23のサイトについて回帰分析を行って方法の有効性を検証した。その結果、データサイズの線形性を確認でき、リンク数の場合には寄与率が1となる場合があることがわかった。手法の精度を向上させるためには、今後データサイズの揺れの幅をどのように決めるかを検討し、数に変動のある余分なリンクを排除する方法を改善する必要がある。また、検索結果を得るためのキーワードの選びかたも一つの重要な要素である。

[2]http://www.completeplanet.com/

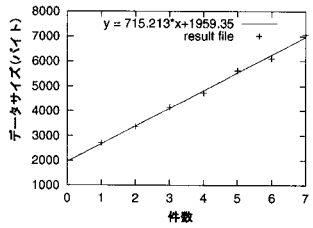

図 2: ファイルサイズの場合の単回帰分析

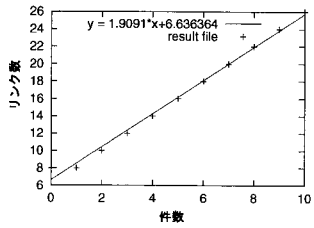

図 3: リンクの場合の単回帰分析

なお、本研究の一部は財団法人ソフトウェア工学研究財団の高度情報化支援ソフトウェアシリーズ育成事業によるものである。

参考文献

[1] S. Hirokawa, S. Watanabe, Y. Koga, T. Taguchi, Automatic Feature extraction of Search sites, Proc.SSGRR2001(to appear).

[2] P. Ipeirotis, L. Gravano, M. Sahami, Automatic Classification of Text Databases through Query Probing, Proc. WebDB'00, 2000.

[3] 古賀 康則, 田口 剛史, 廣川 佐千男：検索サイト統合のためのラッパー生成法, DEWS2001 CD-ROM:6b-1, 2001.

[4] T. Tagchi, Y. Koga and S. Hirokawa, Integration of Search Sites of the World Wide Web, Proc.CUM, vol.2, pp.25-32,2000.