WWW情報の重要度に基づく自動収集の試み*

4Aa－9

松岡 淳子　鈴木 悦子　来住 伸子　小川 貴英†
津田塾大学 数学科‡

1　はじめに

WWW(World Wide Web) とは、インターネットの広域情報検索システムの一つであり、WWWの情報量は急速に増加している。そのため、全ての情報を把握することや、必要な情報を探すことが困難になってきた。

WWWで必要な情報を検索するには、検索ツールが必要である。最近の検索ツールは、ロボットで情報を自動収集し、キーワード検索を行なっている。既存のロボットでは、大規模な計算機環境・高速ネットワークのもとでは、網羅的に情報を収集している。

しかしながら、限られた環境下で情報を収集する場合は、情報の選択を行なわなければならない。ここでは、情報の重要度・関心度を動的に計算しながら、情報の重要度を自動的に評価し、重要度の高い情報を効率的に収集することに試みた。

2　重要度・関心度

HTML(Hypertext Markup Language) ドキュメントAの中に、ドキュメントBのURL(Uniform Resorce Location) の記述があるとき、AがBを参照していると言う。また、Bを参照しているドキュメントの個数をBの被参照回数と言い、Aが参照するドキュメントの個数をAの参照回数と言う。

重要度とは公共性や関心の高さであり、被参照回数の多いものを重要度が高いとする。しかし、全体からの被参照回数を求めることは、莫大な数のドキュメントから相互参照の統計をとることなので、不可能である。そこで、参照回数の多いもの関心度が高いとし、その関心度を用いることによって、被参照回数を推定し、重要度を定義することにした。

関心度とは、ドキュメントに書かれた参照タグの個数をもとに計算する。関心度は、

- 同一 Server 内のドキュメントへの参照回数
- 外部 Server 上のドキュメントへの参照回数

を用いて計算する。

参照元ドキュメント i の関心度 $f(i)$ を、

$$\begin{aligned} f(i) \;=\; & log_2(a*(\text{同一}Server\text{内のドキュメントへの参照回数})+ \\ & b*(\text{外部}Server\text{内のドキュメントへの参照回数})+1) \end{aligned}$$

$$(0 \le a, b,\quad a, b\text{は実数})$$

とする。

重要度は、参照元ドキュメントの関心度に重み付をしたものの和である。被参照ドキュメント j の重要度 $X(j)$ を、

$$X(j) = \sum_i g_{ij} * f(i)$$

(i は既に内容を入手したドキュメントの全て)

$$g_{ij} = \begin{cases} 0 & (i\text{から}j\text{への参照が無い場合}) \\ c & (i\text{から}j\text{への参照が有り、}i\text{と}j\text{が同一サーバの場合}) \\ d & (i\text{から}j\text{への参照が有り、}i\text{と}j\text{が異なるサーバの場合}) \end{cases}$$

$$(0 \le c, d,\quad c, d\text{は実数})$$

とする。

3　アルゴリズム

ドキュメント収集の順番は、以下のアルゴリズムで行なう。

1. サーバAのドキュメントA1を入手(GET)する。
2. ドキュメントA1の関心度を計算する。
3. ドキュメントA1から参照しているドキュメントの重要度を計算する。
4. まだGETされていないドキュメントのうち、最も重要度の高いドキュメントをGETする。
5. 新しくGETしたドキュメントについて、2～4を繰り返す。

例として、c＝1, d＝2の場合を図1,2に示す。図1の状態から、最も重要度の高いdocument-B1をGETすると図2の状態になる。

*Automatic retrieval of WWW resources based on importance estimation

†Junko Matsuoka, Etsuko Suzuki, Nobuko Kishi, Takahide Ogawa
‡Tsuda College, Department of Mathematics

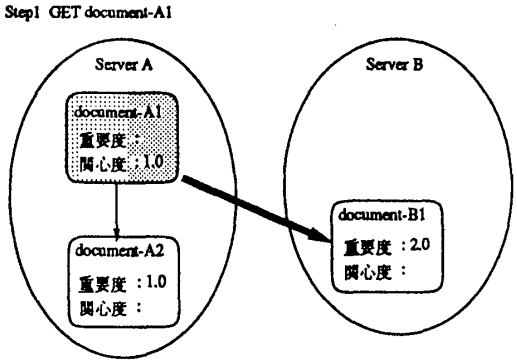

図 1: document-A1 が、3 まで終了した状態

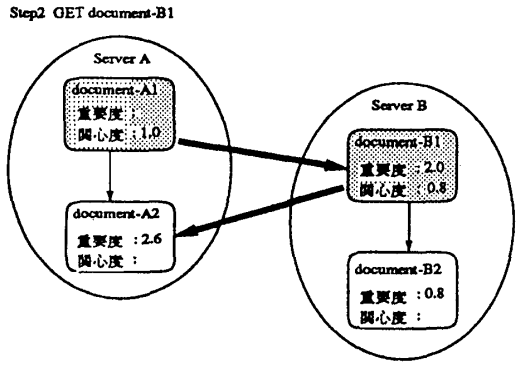

図 2: document-B1 が、3 まで終了した状態

4 実装

収集する対象は、HTML ドキュメントのみとした。プロトコルは、HTTP (Hypertext Transfer Protocol) を用いた。具体的には、以下の 1 または 2 の Request を用いた。

1. `HEAD` *URI* `HTTP/1.0`
   `GET` *URI* `HTTP/1.0`
2. `GET` *URI* `HTTP/1.0`
   `If-Modified-Since:` *HTTP-date*

5 評価

1 つの研究機関内の Server 数 4, ドキュメント数 1000 を対象に実験を行なった。同一 Server への参照を重視し、$a = 0.2$, $b = 0.1$, $c = 1$, $d = 0.5$ と設定し、図 3の結果を得た。太線は重要度にもとづく収集を行なった場合で、細線は幅優先の収集を行なった場合である。重要度にもとづく収集を用いると、前半は被参照回数の多いドキュメントを GET する確率が高いが、後半は確率が低くなっている。しかし、全体からの被参照回数が多いにもかかわらず、GET された時点が遅いドキュメントは、ある特定のユーザが自分のドキュメントの間で相互参照を行ない被参照回数

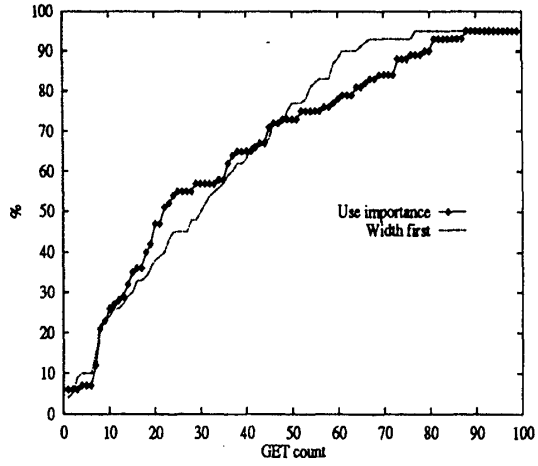

図 3: GET 回数と, 被参照回数の多いドキュメントの GET 率

x 軸 Request GET を行なった回数
y 軸 被参照回数の多いドキュメントを GET した回数 / 被参照回数の多いドキュメントの個数

が多くなったものであり、公共性・他人からの関心を考えると重要性が高いとは言えない。よって、本方式では、公共性や他人からの関心の高いドキュメントは早い時点で GET し、全体からの被参照回数は多くても公共性・他人からの関心の低いドキュメントは遅い時点で GET すると言える。

情報収集空間を狭めた場合には、$0 \leq b < a$, $0 \leq d < c$ と設定し、同一 Server への重みを重くすることによって、効率的な情報収集を行なうことが出来た。現在は、情報収集空間を広げて、$0 \leq a < b$, $0 \leq c < d$ と設定し、実験中である。

6 まとめと今後の展望

以上の評価から、重要度の高い情報を優先的に収集することが、WWW 情報の効率的な収集に有効であると言える。

今後は、重要度・関心度の評価式の見直しと、目的に見あった定数の決定を行ない、実用性を高めていく。

参考文献

[1] T.Berners-Lee, R.Fielding, and H.Frystyk "Hypertext Transfer Protocol–HTTP/1.0", <http://www.ics.uci.edu/pub/ietf/http/draft-ietf-http-v10-spec-04.txt>

[2] "Guidelines for Robot Writers" <http://info.webcrawler.com/mak/projects/robots/guidelines.html>