Webページの更新時間間隔の測定と統計的性質*

4X−05

松岡高治†　森川 綾‡　水谷正大§
東京情報大学情報システム学科¶

1 Webページ空間の調査

Webページ空間は膨大で複雑なリンク関係を持ち、巨大な空間構造の存在が予感されている。例えば、Webページのリンク数とその頻度にはベキ法則が成立するとか、任意のWebページ間の平均リンク数がSmall World性 [D.J. Watts and S.H. Strogatz(1998)] を有するなどの報告 [A.-L. Barabasi and R. Albert(1999)] がある。

本稿ではこのようなWebの空間的性質ではなくWebの動的性質に着目し、その研究方法の1つとしてWebページ群の更新間隔の測定実験を行い、Web空間の時間構造を探るための統計分析法を提案する。

2 URLの更新時間列

URL *url* で指定されるWebページに対して次のようなHTTP HEAD要求

```
HEAD url HTTP/1.1
If-Modified-Since: Thu, 01 Jan 1970 00:00:00 GMT
```

を送り、その応答を記録するという測定を継続的に行うことを考える。このHEAD要求に対してLast-Modified時間を返すURL集合を$\mathcal{U}=\{u_i\}$、その要素数を$N=|\mathcal{U}|$とする。あるURL u_i を一定の測定間隔 Δt でLast-Modefied時間を L 回測定し続けて得られるLast-Modified時間の列を

$$(\ell_1(i),\ \ell_2(i),\ldots,\ell_j(i),\ldots,\ \ell_L(i))$$

とする。このとき、差分列 $(\ell_2(i)-\ell_1(i),\ \ell_3(i)-\ell_2(i),\ldots,\ell_j(i)-\ell_{j-1}(i),\ldots,\ell_L-\ell_{L-1})$ において零でない項はページ更新があったことを示すが、これを順番に並べた列

$$I(u_i)=(I_1(i),\ I_2(i),\ \ldots,\ I_{n_i}(i))$$

を測定間隔 Δt によって得られたURL u_i の更新時間間隔列と定義する。ただし、周期間隔 Δt で測定しているために（さらに実際の測定では正確に周期的とはならない）、ここで得られる更新時間間隔は真の値ではないことを注意しておく。こうして、対象としているURL集合全体に対して同様の測定を行って更新時間間隔列の集合

$$I(\mathcal{U})=\{I(u_1),\ldots,\ I(u_i),\ldots I(u_N)\}$$

が得られる。

3 測定の方法

ページ更新間隔測定の対象であるURL集合をWeb空間から抽出するための合理的方法を確立することは容易ではない。そこで、本稿では

- Google(`http://www.google.co.jp/`)
- Yahoo(`http://www.yahoo.co.jp/`)
- Infoseek(`http://www.infoseek.co.jp/`)

のWeb検索システムのディレクトリサービスから提供されているURL群（それぞれ、約19,000、27,000、19,000個）を測定対象の候補とした。実際の測定に際しては、CGIやSSIページなどのためにHEAD要求に対してLast-Modifiedを返さなかったり、Webサーバ側の時間設定不良などのために更新時間に矛盾をきたすURL群が存在する。これらを取り除き、測定期間にわたって常に矛盾のないLast-Modified時刻を取得できたURL群を測定対象としてのURL集合$\mathcal{U}$とした。測定間隔$\Delta t=1$日として、2001年11月21日から測定実験を行った。

4 着目する統計量

いま、十分に長期間に渡って測定を繰り返した結果、十分長い更新時間列が得られたとする。このとき、次の統計量に着目する。

Measurement of Renewal Intervals of Webpages and their Statistical Properties
†Takaharu Matsuoka
‡Aya Morikawa
§Masahiro Mizutani
¶Dept. of Information Systems, Tokyo University of Information Sciences

4.1 URLごとの更新時間間隔の平均分布

URL u_i の更新時間間隔列 $I(u_i) = (I_1(i), I_2(i), \ldots, I_{n_i}(i))$ において、時間区間 $[T_k, T_{k+1})$ に属する更新時間間隔の個数 $m_k(u_i)$

$$m_k(u_i) = {}^{\#}\{I_j(i) \in [T_k, T_{k+1}) \mid I_j(i) \in I(u_i)\}$$

を求め、時間区間 $\{[T_k, T_{k+1})\}_k$ 上の頻度 $\frac{m_k(u_i)}{n_i}$ を考える。URL集合 $\mathcal{U}$ 内のすべてのURLについてこれを求めて、各時間区間 $[T_K, T_{k+1})$ 上でのURLについての平均値

$$\langle m_k \rangle_{\mathcal{U}} = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} \frac{m_k(u_i)}{n_i}$$

を'URLごとの更新時間間隔の平均分布'と定義する。

4.2 全URLに関する更新時間間隔分布

一方、URL全体にわたる更新時間間隔列の集合 $I(\mathcal{U}) = \bigcup_i I(u_i)$ において、時間区間 $[T_k, T_{k+1})$ 上の更新時間間隔のURL集合 $\mathcal{U}$ についての個数 $M_k(\mathcal{U})$

$$M_k(\mathcal{U}) = {}^{\#}\{I_j \in [T_k, T_{k+1}) \mid I_j \in I(\mathcal{U})\}$$

を求め、時間区間 $[T_k, T_{k+1})$ 上の頻度 $\langle M_k \rangle_{\mathcal{U}} = \frac{M_k(\mathcal{U})}{\sum_i n_i}$ を'全URLに関する更新時間間隔分布'と定義する。$M_k(\mathcal{U}) = \sum_{i=1}^{|\mathcal{U}|} m_k(u_i)$ である。

4.3 平均更新時間間隔分布

URL $u_i \in \mathcal{U}$ の更新時間間隔の平均 $\langle I(i) \rangle = \frac{1}{n_i} \sum_{k=1}^{n_i} I_k(i)$ が $\mathcal{U}$ において時間区間 $[T_k, T_{k+1})$ に属する個数 $m_k(\langle \mathcal{U} \rangle)$

$$m_k(\langle \mathcal{U} \rangle) = {}^{\#}\{\langle I(i) \rangle \in [T_k, T_{k+1}) \mid u_i \in \mathcal{U}\}$$

の頻度 $\frac{m_k(\langle \mathcal{U} \rangle)}{|\mathcal{U}|}$ を'平均更新時間間隔分布'と定義する。

4.4 更新時間間隔列のメンバ数分布

URL u_i の更新時間間隔列 $I(u_i) = (I_1(i), I_2(i), \ldots, I_{n_i}(i))$ に関する時間区間 $[T_k, T_{k+1})$ 上のメンバ関数 $\mathcal{C}_{[T_k, T_{k+1})}(u_i)$ を

$$\mathcal{C}_{[T_k, T_{k+1})}(u_i) = \begin{cases} 1 & {}^{\#}\{I_j \in [T_k, T_{k+1}) \mid I_j \in I(u_i)\} \geq 1 \\ 0 & \text{それ以外} \end{cases}$$

とする。このとき、時間区間 $[T_k, T_{k+1})$ 上の全URL集合にわたるメンバ数頻度

$$\frac{\sum_{i=1}^{|\mathcal{U}|} \mathcal{C}_{[T_k, T_{k+1})}(u_i)}{\sum_{k=0}^{\infty} \sum_{i=1}^{|\mathcal{U}|} \mathcal{C}_{[T_k, T_{k+1})}(u_i)}$$

を'更新時間間隔列のメンバ数分布'と定義する。

4.5 更新時間間隔列のリターン写像の和

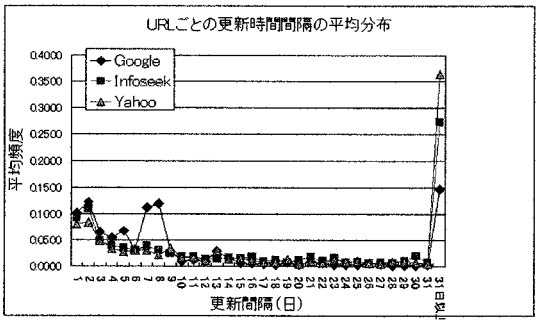
更新時間間隔列 $(I_1, \ldots, I_{n-1}, I_n, I_{n+1}, \ldots)$ について、時間間隔 I_n の次の更新時間間隔 I_{n+1} がリターン写像 $T : [0, \infty) \to [0, \infty)$ によって

$$I_{n+1} = T(I_n)$$

と与えられると仮定する。すなわち、座標点 $(I_1, I_2), (I_2, I_3), \ldots, (I_{n-1}, I_n), (I_n, I_{n+1}), \ldots$ という隣接対は写像 T の関数グラフの上に乗ることになる。URL u の更新時間間隔列 $I(u)$ における隣接対をプロットして得られるリターン写像 T_u のグラフを全URL $\mathcal{U}$ にわたって重ね合わせたものを'更新時間間隔列のリターン写像の和'と定義する。

5 測定結果

URLごとの更新時間間隔の平均分布の例を次に示す。詳細は発表で述べる。



本研究は文部科学省助成の東京情報大学学術フロンティア推進事業から一部補助を受けた。