ユーザの心理的距離に則したWebページ間の新しい距離の定義

松尾 豊[†1] 大澤幸生[†2,†3] 石塚満[†4]

本論文では，標準クリック数と呼ぶWebページ間の新しい距離を提案する．この距離は，ランダムサーファモデルに基づいた確率を距離に変換したものであり，ページ p の出自リンク数 $OutDegree(p)$ によって $-log_7 1/OutDegree(p)$ と定義される．標準クリック数は，従来のクリック数よりも，ページ作成者の感じる心理的な距離感と相関が高いことを示す．

A New Difinition of Subjective Distance between Web Pages

Yutaka Matsuo,[†1] Yukio Ohsawa[†2,†3] and Mitsuru Ishizuka[†4]

The pages and hyperlinks of the World Wide Web may be viewed as nodes and edges in a directed graph. In this paper, we propose a new definition of the distance between two pages, called *average-clicks*. It is based on the probability to click a link through random surfing. We compare the average-clicks measure to the classical measure of clicks between two pages, and show the average-clicks fits better to the users' intuitions of distance.

1. まえがき

World Wide Webは，ページをノード，リンクをエッジとする有向グラフとして表すことができる．リンクは各ページの作成者の判断に基づいて形成されるので，このグラフはWeb上での社会的な構造を表しているといえる[2)]．多くのページからリンクを張られるページは質の高いページであり，質の高いページがリンクを張るページは多くの場合，質の高いページである．このようなWebのグラフ構造を用いた研究は近年さかんに行われており，なかでも，検索エンジンGoogle☆で用いられているアルゴリズムの1つであるPageRank[5)]や，hubとauthorityをグラフ構造から見つけ出すKleinbergらの研究[10)]が代表的である．

また，グラフ構造を用いて，Webコミュニティを発見する研究も多く行われている．文献12)では，完全2部グラフとして表されるコミュニティを発見する手法が提案されており，文献14)では参照の共起性によりコミュニティを抽出する手法について述べられている．文献8)は，最大流量最小カットの定理を用いて，コミュニティとそれ以外を分ける適切なカットを見つける試みである．これらの研究では，その多くがリンクの長さを単純に1としている．

ほかにも，クリック数をページ間の距離として用いている例は多い．たとえば，文献7)は，ルートノードから数クリックにあるページでクエリに関係のあるページを探す手法であり，文献1)では，任意のWebページ間の平均クリック数は19クリックであると報告されている．

しかしながら，クリック数で測ったページ間の距離がユーザの直観的な距離を適切に表すとは限らないだろう．ページ内に大量のoutlinkを含むページもある一方で，多くのページは数個以下のoutlinkしか持たない[11)]．ユーザにとって，ページ内の多数のリンク☆☆の中から1つを選んでクリックするのと，数個のリンクから1つのリンクをクリックするのでは，前者の方が「より遠くに行っている」感覚があるのではないだろうか．

一例をあげると，ある友人のホームページには数個のリンクしかなく，それらのリンクはその友人の興味や研究，友人関係などである．しかし，別の友人の

†1 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
†2 科学技術振興事業団
Japan Science and Technology Corporation
†3 筑波大学
The University of Tsukuba
†4 東京大学
The University of Tokyo

☆ http://www.google.com/
☆☆ 以下，リンクというときはoutlinkを指す．

ホームページには多くのリンクがあるが，彼の興味や知り合いのほかにも，検索エンジンや企業・オンラインショップへのリンク，交通機関や地図に関するリンク，食事に関するリンクなど，さまざまなリンクを含んでいる．一般に，多くのリンクを持つページでは，リンクの数の少ないページに比べて，リンク先のページまでのコンテクストの変化は大きくなると考えられる．したがって，リンク数が多いページからリンク先までのページの距離は，より長いと考えるのが自然であろう．

本論文では，従来のクリック数（clicks）の代わりに，標準クリック数（average-clicks）と呼ぶページ間の新しい距離を提案する．標準クリック数は，ユーザの距離感をより適切に反映しており，Web コニュニティの発見やリンク情報を利用した研究の多くに基礎的な貢献ができるものと考えている．さらに，ユーザのホームページから他のページまでの標準クリック数が，そのページの分かりやすさや情報量に相関があることを示す．ユーザのホームページから離れているページほどユーザはその内容を理解しにくく，またユーザにとっての情報量は多くなる．

以下，2 章では標準クリック数の定義を述べ，3 章では例と評価実験を述べる．4 章で関連研究と議論を行う．

2. 標準クリック数

2.1 標準クリック数の定義

ユーザがあるページ p を閲覧していると仮定しよう．このとき，このページに $OutDegree(p)$ 個のリンクがあるとする．バックボタンで戻る場合や URL を直接入力する場合などを除けば，このユーザは，$OutDegree(p)$ 個のリンクの中から 1 つのリンクを選んでクリックすることになる．どのリンクをクリックするか，事前にまったく分からないと考えると，このユーザがあるリンクをクリックすることによって得られる情報量は，

$$\log(OutDegree(p)) \tag{1}$$

である．つまり，ユーザが多くのリンクの中から 1 つのリンクを選ぶ行為は，少ないリンクの中から 1 つのリンクを選ぶ行為よりも，得られる情報量という点では大きいことになる．ここで対数を用いているのは，a 個のリンクから 1 つ選ぶことを 2 回繰り返すことと，a^2 個のリンクから 1 つを選ぶことは，ユーザが行っているリンクの選択という点では同じであり，対数を用いることで

$$2 \times \log a = \log a^2$$

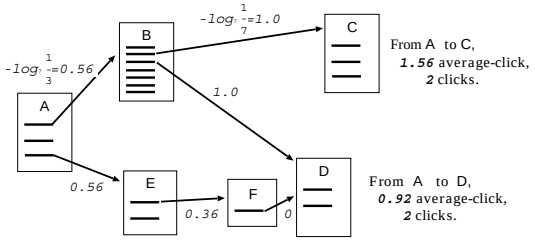

図 1 標準クリック数と従来のクリック数
Fig. 1 Average-dicks and classical clicks.

とすることができるためである[18]．したがって，式 (1) で表される値をリンクの長さと考え，加算することで距離と考えることができる．リンクの長さを次のように定義する．

定義 2.1（リンクの長さ） ページ p における 1 つのリンクの長さを以下で定義する．

$$\log_n(OutDegree(p)) \tag{2}$$

ここでは，$OutDegree(p)$ は，ページ外への html 文書へのリンクであり，画像や音楽ファイルなどへのリンクやページ内リンクなどは除く．平均的なページはページ内に 7 つのリンクを持つと報告されている[3]☆ので，log の底 n を 7 とする．このように定義された距離の単位を**標準クリック**と呼ぶ．

定義 2.2（標準クリック） $n = 7$ のとき式 (2) で定義されるリンクの長さの単位を，標準クリックと呼ぶ．

2 つのページ p と q 間の距離（標準クリック数）はクリック数の場合と同様，最短経路の長さで定義する．

定義 2.3（ページ間の標準クリック数） ページ p からページ q までの標準クリック数は，p から q への最短経路の長さ，すなわち経路中の各リンクの標準クリック数の和の最小値である．

確率的な観点からとらえると，最短経路を求めることは，ランダムサーファが p から q に到達するときの最尤経路を求めていることに相当する．

図 1 に，従来のクリック数と標準クリック数の関係を示す．長方形は 1 つのページを，横棒は 1 つのリンクを表している．ページ A は 3 つのリンクを持つため，各リンクの長さは $-\log_7(1/3) \approx 0.56$ 標準クリックである．ページ B は 7 つのリンクを持つため，1 標準クリックであり，ページ A から C まで 1.56 標準クリックとなる．一方，ページ A から D までは 2

☆ 最近の調査では，平均的なページは 1 つの外部リンクと 4 つの内部リンクを持つという報告もされている[15]．この 7 という数字は便宜的なものであり，実際には e や 2，10 などの値をとってもかまわない．本論文では，「標準的なクリック」いくつ分に相当するかという意味を持たせるために，7 としている．

つの経路があり，クリック数では上側の経路の方が短いが，標準クリック数では下側の経路の方が短い．なお，ページ F のようにリンク数が 1 のページはリンクの長さは 0 である．

標準クリック数を用いることで，ユーザの直観的な距離感をより正確に表すことができる．たとえば Yahoo! のトップページは現在 180 以上のリンクがあるが，直観的には，たとえ同じクリック数であったとしても，Yahoo! のトップページを経由する経路より知り合いのページをたどる経路の方が短く感じるであろう．標準クリック数では，Yahoo! のトップページからサブページへの経路は，図 1 の上側の経路同様，距離が長く，友人や趣味のページへのリンクなどは，図 1 の下側の経路のように距離が短くなる．

2.2 一定の距離内にあるページ数に関する考察

標準クリック数を用いると，あるページの与えられた距離内に存在するページ数の上限を求めることができる．クリック数の場合には，与えられた距離内に存在するページの数は実際に探索したりするまで分からないのに対し，標準クリック数のこの性質は，探索ロボットの数や探索の深さを決める際に有用である．

定理 2.1（距離 r 以内のページ数の上限） ページ a から距離 r 標準クリック以内には，たかだか $(2 \cdot 7^r - 1)$ 個のページしか存在しない．ただし，p 内のリンク数が 1 のページはカウントしない．

証明 2.1 a から r 標準クリック以内のページで構成される有向循環グラフを考える．このグラフの各ノードに対して，a からの最短パスを求め，a を根とし各ノードまでの最短パスを示すツリーを抜き出す（**図 2**）．この変形したツリーのノード数は，もとのグラフのノード数と同じである．

次に，このツリーにおいて，最も遠い葉ノードまでの距離は r 標準クリックであるから，ページ a を閲覧していたユーザが各ノードに達する確率は $1/7^r$ 以上である．したがって，葉ノードの数は 7^r 個以下である．各ノードが 2 個以上の子ノードを持つとした場合，このツリーにおけるノード数はたかだか $2 \cdot 7^r - 1$ 個である． □

たとえば，2 標準クリック以内のページは 97 個以下，3 標準クリック以内のページは 685 個以下，4 標準クリック以内のページは 4,801 個以下である．ある query に関連するページを中心に探索を行う focused crawling [7] と呼ばれる研究もさかんに行われているが，標準クリック数を用いる際には，この定理が実際のインプリメンテーションに役に立つだろう．

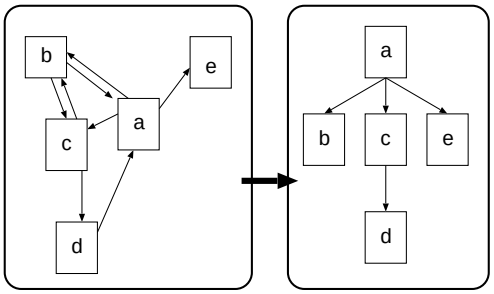

図 2 ツリーへの変換
Fig. 2 Transformation into a tree.

3. 具体例と評価

Web コミュニティを抽出したり，あるクエリに対して目的のページを反復的に探索したりするアルゴリズムでは，ページ間の距離をどのように計るかが重要であると考えられる．たとえば，ある人の研究のページとそこからリンクが張られている研究会のページでは，用いられている単語は異なるかもしれないが，心理的な距離は近いであろう．このような心理的な距離は，ページの内容が似ているかどうかではなく，ページの作成者が他のページをどのくらい距離があると感じるかによって検証することができると考えられる．

本章では，まず，標準クリック数によるページ間の距離の計算例を示し，標準クリック数がクリック数に比べ，ページ作成者にとっての距離感をより適切に表していることを示す．

3.1 標準クリック数の具体例

2 ページ間の距離を求めるには，最短経路を探索する必要がある．ページ s からページ t への最短経路を探索する最良優先探索アルゴリズムを**図 3** に示す☆．

本論文の第 1 筆者のホームページ☆☆から，各サイトへの距離は**表 1** のようになる．この結果から，以下のことが分かる．

- 経路がリンク集を含んでいない．
- 標準クリック数は，直観的な距離に適合している．つまり，第 1 著者が親しみのあるページは距離が小さく，そうでないページは距離が大きい．
- 最短経路は，2 ページ間の間接的な関係を明らかにしている．

研究室の後輩や Yahoo!への距離は短いが，情報処理学会や IJCAI のホームページはやや遠い．これは著者

☆ なお，あるクエリに対して（距離としてクリック数を用いた場合には）幅優先探索による探索が良い結果をもたらすことが報告されている [16]．

☆☆ http://www.miv.t.u-tokyo.ac.jp/~matsuo/

```
function SEARCH_SHORTEST_PATH (s, t, d_thre)
  α ← 1.0,   n ← 7.
  list ← ADD_LIST(s, empty),  d(s) ← 0.
  p ← s
  while p ≠ t
    Fetch page p and extract links which points to
page p_k (k = 1, ..., n_p)
    for k ← 1 to n_p
      d(p_k) ← d(p) − log_n(α/n_p)
      if d(p_k) > d_thre then next
      list ← ADD_LIST(p_k, list)
    end
    if list is empty return failure
    p ← CHOOSE_MINIMAL(list, d)
  end
  return d(t)
```

ADD_LIST$(a, list)$ は $list$ に a を加える関数.
CHOOSE_MINIMAL$(list, d)$ は，$d(a)$ を最小化するような $a \in list$ を見つける関数.
d_{thre} は探索空間の半径（標準クリック数）.

図 3 最短経路を見つける最良優先探索
Fig. 3 Best-first search for the shortest path.

の距離感に合っている．著者は，相撲が好きなのであるが，残念ながら大相撲協会への距離は，J リーグの公式サイトよりも距離が遠い．このように，著者がより身近に感じているにもかかわらず，距離が遠くなってしまうことも当然ある．

3.2 ユーザの感じる距離感による評価

標準クリック数がユーザの距離感を適切に表しているかどうかを確かめるため，アンケートによる評価実験を行った．実験の方法は以下である．まず，被験者（自分のホームページを持っている人に限る）のホームページをスタートページとして，数クリックの範囲のあるページをすべて探索する．次に，その中からランダムに 30 個，URL を取り出す．ホームページからの距離やページの内容は見せずに，取り出した URL だけを被験者に提示し，どのくらい身近に感じるかを 5 段階で評価してもらった．1 が「とても身近に感じる」，5 が「とても遠くに感じる」である．各被験者は，提示された URL を自分の分かる範囲で解釈し，その心理的な距離を答えることになる．

被験者 1 の回答とホームページからの距離の散布図を**図 4**，**図 5** に示す．被験者の回答と標準クリック数は強い相関が読み取れるのに対し，クリック数とはあまり相関はみられない．13 人の被験者に対しての回答と標準クリック数の相関係数を**表 2** に示す☆．比較

☆ 被験者 13 人は，大学生/大学院生/教員であり，そのうち人工知能の分野が 6 人，言語学などそれ以外が 7 人である．また男性が 10 人，女性が 3 人である．

表 1 第 1 著者のホームページからの標準クリック数.
Table 1 Average-clicks from the author's homepage.

To URL 最短経路	標準クリック数 / クリック数 そこまでの標準クリック数
研究室の後輩のホームページ	1.62 / 2
http://www.miv.t.u-tokyo.ac.jp/~matumura/	1.62
http://www.miv.t.u-tokyo.ac.jp/JAICO/	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
Yahoo!	3.02 / 3
http://www.yahoo.co.jp/	3.02
http://www.geocities.co.jp/.../6353/whatsnew.html	2.67
http://www.geocities.co.jp/Athlete-Athene/6353/	1.13
http://www.miv.t.u-tokyo.ac.jp/ matsuo	0.0
情報処理学会	5.15 / 4
http://www.ipsj.or.jp/	5.15
http://wwwsoc.nacsis.ac.jp/jsai/links.html	2.75
http://wwwsoc.../jsai/whatsai/navigation.html	1.75
http://wwwsoc.nacsis.ac.jp/jsai/whatsai/	1.18
http://www.miv.t.u-tokyo.ac.jp/ matsuo/	0.0
IJCAI ホームページ	5.39 / 5
http://ijcai.org/	5.39
http://w3.sys.es.osaka-u.ac.jp/~osawa/AIlinks.html	3.33
http://www.gssm.otsuka.tsukuba.ac.jp/staff/osawa	1.97
http://www.miv.../~matumura/research.html	1.62
http://www.miv.t.u-tokyo.ac.jp/JAICO/	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
J リーグ公式サイト	4.03 / 3
http://www.j-league.or.jp/index.html	4.03
http://www.miv.t.u-tokyo.ac.jp/~tomobe/	2.50
http://www.miv.../member/present-mem.htm	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0
大相撲協会ホームページ	9.08 / 5
http://www.wnn.or.jp/wnn-t/	9.08
http://www.ntt.co.jp/SQUARE/www-in-JP.html	6.07
http://www.ic.u-tokyo.ac.jp/index.html	4.30
http://www.u-tokyo.ac.jp/	2.67
http://www.geocities.co.jp/Athlete-Athene/6353/	1.13
http://www.miv.t.u-tokyo.ac.jp/~matsuo	0.0

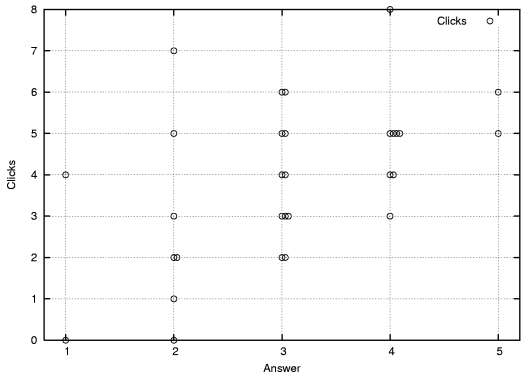

図 4 被験者 1 の回答とクリック数の散布図
Fig. 4 Scatter plot of participant 1's answer vs. clicks.

のため，クリック数のほかに，重みつきクリック数との相関係数も示す．重みつきクリック数とは，文献 7) のヒューリスティックとして用いられている「同じサイト内のリンクは重みを減らす」という考えに基づくものであり，本論文の場合には，同じサイト内のリンクは長さが短く，異なったサイト間のリンクは長いと

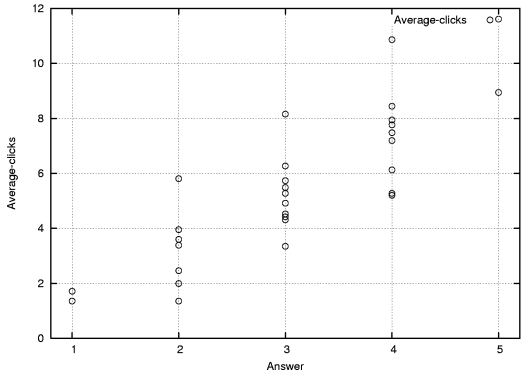

図 5 被験者 1 の回答と標準クリック数の散布図
Fig. 5 Scatter plot of participant 1's answer vs. average-clicks.

考える☆．被験者によって程度に差はあるものの，すべての被験者においてクリック数よりも標準クリック数の方が高い相関が得られている．また，重みつきクリック数の方が相関係数が高い場合もあるが，平均では標準クリック数が大きく上回っている．

3.3 分かりやすさ，情報量による評価

次に，実際にページの内容をユーザに提示し，その内容に関して次の項目にそれぞれ 5 段階評価で答えてもらった☆☆．

- 理解しやすさ：「このページはあなたにとってどのくらい分かりやすいですか?」
- 情報量：「あなたの知らないことがどのくらい書かれていますか?」
- 面白さ：「あなたにとってどのくらい面白いページでしたか?」

各項目とホームページからの距離との相関を**表 3** に示す．理解しやすさの項目は，距離に対して負の相関がある．つまり，ユーザのホームページからの距離が離れれば離れるほどページの内容を理解しにくくなっている．情報量の項目は，距離と正の相関がある．つまり，距離が離れれば離れるほどユーザにとっての情報量は多くなる．いずれの項目に対しても，クリック数，重みつきクリック数と比べて標準クリック数の方が相関が強く，より的確にユーザのコンテクストからの距離を表しているものと考えられる．

しかしながら，面白さとホームページからの距離

☆ 文献 7) には具体的な値は示されていないが，本論文では相関係数が最大になるように，同サイトのリンクは長さ 1，異なったサイト間のリンクは長さ 4 とした．

☆☆ なお，理解しやすさや情報量といった概念は非常に定義しにくいが，ここでは主に，各質問に対する被験者の答えという意味で用いている．

表 2 被験者の距離感との相関係数
Table 2 Correlation with participant's intuitive distance.

被験者	クリック数	重みつきクリック数	標準クリック数
1	0.524	0.623	0.836
2	0.325	0.435	0.804
3	0.696	0.599	0.715
4	0.517	0.626	0.699
5	0.471	0.483	0.685
6	0.641	0.662	0.674
7	0.268	0.502	0.572
8	0.490	0.482	0.569
9	0.499	0.517	0.528
10	0.435	0.568	0.512
11	0.358	0.477	0.436
12	0.116	0.108	0.400
13	0.302	0.358	0.367
平均相関係数	0.434	0.495	0.600

表 3 理解しやすさ，情報量，面白さとの平均相関係数
Table 3 Average correlation with understandability, information and interestingness.

	クリック数	重みつきクリック数	標準クリック数
理解しやすさ	−0.320	−0.359	−0.404
情報量	0.279	4.37	0.462
面白さ	−0.174	−0.269	−0.174

にははっきりとした相関はみられなかった．これは，あるページがユーザにとって面白いかどうかは，そのページを理解できるかどうか，情報量が多いかどうか，ページの質が高いかどうか，ユーザの好みに一致するかどうかなど，さまざまな要因があるためであろう．

4. 関連研究と議論

標準クリック数の定義のもとになっているのは，ランダムサーファモデルである．ランダムサーファモデル[5] は，ページの閲覧者がページ内の各リンクを等確率でランダムにクリックしていくという以下のモデルである．

> ページ p が $OutDegree(p)$ 個のリンクを持つとき，ランダムサーファは $\alpha/OutDegree(p)$ の確率でページ内の各リンクをクリックし，$1-\alpha$ の確率で（ブックマークや URL を直接入力することで）Web 上の任意のページにジャンプする．ただし，α は 0 以上 1 以下の定数である．

検索エンジンの Google では，このモデルを用いた PageRank というアルゴリズムにより Web ページのランキングを行っている[17]．標準クリック数は，この確率を距離に直したものである．このモデルに従えば，リンクの長さは

$$-\log_n(\alpha/OutDegree(p))$$

によって定義されることになる．α は，ランダムサーファがどの Web ページ上にいるかという確率的な分布を求める際に，リンクのないページの確率が大きくなってしまうのをを防ぐ意味もある．PageRank では，$\alpha = 0.85$ としているが，本論文では $\alpha = 1$ としている（$\alpha < 1$ と設定することは，リンク数が 1 のページでもそのリンクの距離が 0 より大であることに相当する．しかし，実験では $\alpha < 1$ とする必要性は認められなかった）．確率の対数をとることでコストとし，最小コストの解（つまり最尤経路）を求めることはコストに基づく仮説推論[9] ではよく知られた考え方である．

リンク情報から，ページ間の構造を見出す研究には，文献 4) や 19) などがある．これらは，ページの距離を定義することで見やすいページを作ったり，リンク間の関係がサイト内か外であるかを判定したりするヒューリスティックについて論じているが，ページ作成者の心理的な距離感に着目したものではない．また，リンクに重みをつける研究として文献 6) がある．ここでは，ページ内の “href” のそばのテキストが，求めるトピックの表現を含んでいればリンクの重みを重くするという処理を行っている．この処理はテキストの分析を必要とするが，標準クリック数はページ内のリンクの数だけで定めることができる．

1 章でも述べたとおり，さまざまな研究でクリック数を用いてページ間の距離を計っているが，標準クリック数を用いるのが適切な場合も多いだろう．たとえば，Web 上でコニュニティを発見する際に，一般的な話題が含まれやすいことが報告されている[3] が，標準クリック数を用いれば，一般的な話題のページは，多くのリンクを持つため遠くにあると考えることができ，除去することができる．また，リンク先のページを反復的に探索するアルゴリズムは多く用いられるが，クリック数で閾値を決めるのではなく，標準クリック数で閾値を決めるべきであろう．数クリック以内という基準では，非常に多くのページを取ってきてしまうかもしれないし，内容を何段にもわたって書いてあるページには対応できない．標準クリック数を用いれば，「もしリンク数が少ないのなら探索を続け，リンク数が多ければやめる」という実用的な方策の根拠となる．

従来のクリック数は，すべてのインターネットユーザに直観的に分かりやすいが，確率に基づく距離は比較的理解するのが難しい．そのために，我々は対数の底を標準的なページのリンク数と定め，「標準的なクリック」でいくつにあたるかというセマンティックを導入し，距離の測度とした．

Web ページには，

- 個人のページで，比較的匿名性が低く，個人の社会的なネットワークを反映するもの，
- 個人のページで，比較的匿名性が高く，趣味や 1 つの事柄についてだけ書いてあるもの，
- 企業・大学など組織や団体のページ，
- ポータルサイト，ニュースサイト，検索エンジン，オンラインショップなど機能的な側面が強いページ，

などがあるが，本論文で対象とし，検証を行っているのは 1 番目のものである．それ以外については，検証が難しいため本論文では対象としていないが，できるものから検証を試みたいと考えている．また，本論文では，ページ作成者からの距離感を検証しており，Web コミュニティなどをとらえるうえでは適切だと考えられるが，そのほかにもページの閲覧者からみた距離感についても考える必要があるだろう．これについても，今後，研究を進めていきたいと考えている．

5. ま と め

本論文では，非常に簡単な定義により，ユーザの距離感をより的確に反映するページ間の新しい距離測度として標準クリック数を提案した．さらに，ユーザのホームページからあるページまでの標準クリック数が，ページの理解しやすさ，得られる情報量などと相関があることを示した．文献 13) では，検索エンジンがユーザのコンテクストを利用すべきだと主張されているが，ユーザのホームページの周りの構造もユーザのコンテクストに関する 1 つの情報源となると考えている．

参 考 文 献

1) Adamic, L.A.: The Small World Web, *Proc. ECDL'99*, pp.443–452 (1999).
2) Adamic, L.A. and Adar, E.: Friends and Neighbors on the Web (to appear). URL:// www.hpl.hp.com/shl/papers/web10/fnn.pdf.
3) Bharat, K. and Broder, A.: A technique for measuring the relative size and overlap of public Web search engines, *Proc. 7th WWW Conf.* (1998).
4) Botafogo, R.A., Rivlin, E. and Schneiderman, B.: Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics, *ACM Trans. Inf. Syst.*, Vol.10, No.2 (1992).
5) Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proc. 7th WWW Conf.* (1998).
6) Chakrabarti, S., Dom, B., Raghavan, P., Ra-

jagopalan, S., Gibson, D. and Kleinberg, J.: Automatic resource compilation by analyzing hyperlink structure and associated text, *Proc. 7th WWW Conf.* (1998).
7) Chakrabarti, S., van den Berg, M. and Dom, B.: Focused crawling: a new approach to topic-specific Web resource discovery, *Proc. 8th WWW Conf.* (1999).
8) Flake, G.W., Lawrence, S. and Giles, C. L.: Efficient Identification of Web Communities, *Proc. ACM SIGKDD-2000*, pp.150–160 (2000).
9) 石塚 満：知識の表現と高速推論，chapter 6, 丸善 (1996).
10) Kleinberg, J.: The Small-World Phenomenon: An Algorithmic Perspective, Technical Report TR 99-1776, Cornell University (1999).
11) Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.S.: The Web as a graph: measurements, models, and methods, *Proc. International Conf. on Combinatorics and Computing* (1999).
12) Kumar, S.R., Raghavan, P., Rajagopalan, S. and Tokins, A.: Trawling the web for emerging cyber communities, *Proc. 8th WWW Conf.* (1999).
13) Lawrence, S.: Context in Web Search, *IEEE Data Engineering Bulletin*, Vol.23, No.3, pp.25–32 (2000).
14) 村田剛志：参照の共起性に基づく Web コミュニティの発見，人工知能学会誌，Vol.16, No.3, pp.316–323 (2001).
15) Murray, B. and Moore, A.: Sizing the Internet, White paper, Cyveillance, Inc. (2000). http://www.cyveillance.com.
16) Najork, M. and Wiener, J.L.: Breadth-first search crawling yields high-quality pages, *Proc. 10th WWW Conf.* (2001).
17) Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing order to the Web, Technical Report, Stanford University (1998).
18) Shannon, C.E.: A mathematical theory of communication, *Bell System Technical Journal*, Vol.27, pp.379–423 and pp.623–656 (1948).
19) Spertus, E.: ParaSite: Mining Structural Information on the Web, *Proc. 6th WWW Conf.* (1997).

(平成 14 年 3 月 20 日受付)
(平成 14 年 11 月 5 日採録)

松尾 豊（正会員）

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士課程修了．博士（工学）．2002 年より産業技術総合研究所サイバーアシスト研究センター勤務．推論，数理計画法，探索アルゴリズムだけでなく，その結果を人間にとってどう役に立てるかにも興味があり，価値の高い情報の提示を目指している．人工知能学会，電気学会，AAAI 各会員．

大澤 幸生（正会員）

1990 年東京大学工学部電子工学科卒業．1995 年同大学院博士課程修了．博士（工学）．大阪大学助手を経て，現在，筑波大学大学院経営システム科学専攻助教授．AAAI，IEEE 各会員．人工知能学会において，1994 年・1998 年全国大会優秀論文賞，1998 年人工知能学会誌の優秀論文賞受賞．

石塚 満（正会員）

1971 年東京大学工学部電子工学科卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所．1978 年東京大学生産技術研究所助教授，1992 年同大学工学部電子情報工学科教授．2001 年より情報理工学研究科電子情報学専攻．研究分野は人工知能，知識処理，マルチモーダル擬人化エージェント，ネットワーク化知的情報環境．IEEE，AAAI，人工知能学会，映像情報メディア学会，画像電子学会等の会員．