データベースシステム 128－40
（２００２．７．１８）

パラメータ化された連結成分分解による Webページのグループ化

正田 備也†, 高須 淳宏‡, 安達 淳‡
†東京大学 情報理工学系研究科, ‡国立情報学研究所

WWW 上の情報の急速な増大は、テキスト情報のみに基づく Web 検索手法をますます非現実的なものとしている。そこで近年、リンク情報に基づく優れた検索手法が多くの研究によって提供されている。本論文は、リンク情報に基づいて Web ページをグループ化する手法を提案する。そのねらいは、検索の単位を大きくすることで、テキスト情報に基づく後続の検索処理の負担を軽減することにある。さらに、この手法は、一つの閾値パラメータを調整することで、グループの粒度を制御することを可能にする。本論文は、予備的実験の結果を含む。これによって、提案されたグループ化手法の特徴が明らかにされる。

Grouping Web pages based on parameterized connectivity

Tomonari Masada†, Atsuhiro Takasu‡, Jun Adachi‡
†Graduate School of Information Science and Technology, The University of Tokyo
‡The National Institute of Informatics

The rapid growth of the amount of information on WWW makes Web search methods based only on textual information more and more unrealistic. In recent years, many researches provide attractive link-based retrieving methods. This paper proposes a method for link-based Web page grouping, which aims to reduce the complexity of following text-based retrievals by enlarging the size of units for those retrievals. This method also makes the granularity of groups controllable by adjusting one threshold parameter. This paper includes the results of preliminary experiments, which clarify the characteristic of proposed grouping method.

1 はじめに

1.1 研究の目的

リンク情報に基づく検索単位の粒度の制御

WWW 上の Web ページ集合のように膨大な数の文書に対する情報検索を実現する場合、いくつかの Web ページをまとめて一つのグループとし、これを最小単位として大まかな検索を実行した上で、次の段階において個々の Web ページに着目するより精緻な検索を行うという方策が考えられる。このように、検索単位の規模を調整しつつ、検索処理を多段的に構成するという着想はすでに見られる [CPKT92]。だが、Web ページの場合、グループ化の作業を [CPKT92] のようにテキスト情報に基づいて行うことは、以下のような理由から望ましくないであろう。第一に、Web ページの数の多さに鑑みれば、すべてのページについて形態素解析のようなテキスト処理の終了していることが前提とされる手法は、非常な計算量を必要としてしまうため。第二に、例えばコンテンツとして画像や動画しか含まないページなど、Web ページの中にはテキスト解析を行うに十分なテキスト情報を含まないものが散見されるため。

そこで、Web ページから得られるテキスト情報以外の情報を利用して予め複数の Web ページを束ね、検索単位の粒度を調整することによって、テキストの内容にまで立ち入る後続の一層精緻な検索処理に必要とされる計算量を抑えるという方策が考えられる。具体的には、複数の Web ページをまとめて単独の文書とみなすことができれば、文書数そのものが少なくなるため、テキスト情報に基づく検索アルゴリズムが入力サイズに依存する計算量を持つ場合に有利となる。また、こうして束ねられた文書に対してメタデータを付与する場合も、各々のページに対

して個別に付与する場合に比べてコストが減る。

1.2 グループ化のための手法

URLに関するヒューリスティクスによるグループ化

Webページを束ねる際の発見的手法として、サイト内の一つのディレクトリに属するページをグループ化することが考えられる。すなわち、URLをグループ化の手がかりとするのである[THA99]。しかし、URLをグループ化の手がかりとして機能させるためには、様々なサイトの内部構造とURLの階層構造との対応関係を経験的に調査する必要があり、その手間は無視できない。また、少数の事例から得られたヒューリスティクスが多数の事例にそのまま適用できるとは限らない。

グラフ理論上の概念の利用

そこで、リンク情報、すなわち、Webページに含まれるハイパーリンクから得られる情報を利用して、件のグループ化を実現することが考えられる。複数のWebページの織りなすリンク構造は、一つの有向グラフとみなすことができる。そして、有向グラフ上で頂点のグループ化を引き起こす概念としては、強連結性の概念がある。だが、Webページの集合について強連結成分分解を行うと粗野なグループ化しか得られないことが知られている。

例えば、比較的大規模なWebページの集合に対して強連結成分分解を行う場合、与えられたページ集合の規模に匹敵するような巨大な成分が得られてしまう[BKM+00]。また、一つのサイト内を強連結成分に分解する場合も、サイト全体の規模に匹敵する規模のページ集合が得られてしまうことが明らかにされている[小島02]。つまり、処理の対象となる全Webページの集合との対比で考えたとき、テキスト内容上も一つにまとまっていると期待するにはあまりにも多くのページを含むようなグループが構成されてしまう。

パラメータ化された連結成分分解

そこで、本研究では、強連結成分を細分化するかたちでWebページのグループ化を実現するアルゴリズムを提案する。このアルゴリズムには、閾値としてはたらく一つのパラメータが増減するのに応じて、Webページが同じグループに属するか否かの判定の厳しさを変化させる仕組みが備わっている。これによってグループの粒度の調整が可能となっている。

また、Webページが同じグループに属するか否かの判定は、ページ間の距離の大小によって判定される。本研究では、この距離を、リンク構造のみに基づいて定義する。したがって、リンク構造上でのWebページ間の近さの情報のみにもとづいて、Webページの可変的なグループ化が可能となっている。

すなわち、本研究の特徴は、次の3点にまとめられる。

1. 提案されているアルゴリズムによるグループ化が、強連結成分分解の細分化になっている点。
2. グループの大きさを、一つのパラメータの増減によって制御できるようになっている点。
3. グループの粒度の制御を可能にするための理論上の道具として、リンク構造上でのページ間の近さを定量的に表す概念を提案している点。

これら三点を兼ね備えた研究は、今までになかったと思われる。

1.3 論文の概要

第2章では、リンク構造上でWebページ間の近さを表すために用いるドリフトという概念を定義する。ドリフトとは、あるWebページから別のWebページへの移行のしやすさを示す値である。このドリフトは、基本的には[Kle99]に紹介されているKatzのstandingの概念[Kat53]に倣ったものである。だが、その利用の仕方が全く異なる。Katzがこれを文書のランク付けに利用しているのに対し、本研究はWebページ間のリンク構造上での近さを定量的に評価するために利用する。実際、第2章では、ドリフトに基づいて、リンク構造上での2つのWebページの間の距離を定義している。この距離は**相互リンク距離**と呼ばれる。

相互リンク距離を利用したグループ化のためのアルゴリズムは、第3章で示される。このアルゴリズムによれば、特定の値以上の相互リンク距離で隔てられたWebページは、同じグループに属することがない。そして、得られるグループの大きさは、閾値パラメータτによって制御される。$\tau = \infty$のときに最もグループの粒度が粗くなり、グループ化は強連結成分分解に一致する。τを減少させるにつれ、強連結成分をますます細分化するグループが得られる。

第4章は、予備実験の結果を含んでいる。まず、実

際に得られたグループの例がいくつか示される。次に、閾値パラメータを調整することによってグループの大きさの実際に変化する様が、実験データにより確かめられる。さらには、同じグループに属する文書が、テキスト情報の上から見ても互いに類似していることが明らかにされる。

第5章では、本研究の提案するドリフト、および相互リンク距離という概念が本研究にもたらしている発展性に言及し、今後の課題を述べる。

2 概念の定義

2.1 ドリフト

本研究は、あるWebページから別のWebページへの移行のしやすさを表わす尺度として、ドリフトという概念を提案する。そして、このドリフトに基づいて、リンク構造上でのWebページ間の近さである相互リンク距離を定義する。

WWWのリンク構造は、Webページを頂点(vertex)、ハイパーリンクを有向枝(arc)と見なすことによって、一つの有向グラフと解されうる。この有向グラフGの頂点集合、つまりWebページの集合をV、有向枝の集合、つまりハイパーリンクの集合をEとする。なお、頂点から自分自身に対して張られている有向枝は無視することにする。また、同じ2つの頂点間に複数の同じ向きの有向枝がある場合、これらを一つの有向枝とみなすことにする。

有向グラフ$G=(V,E)$の隣接行列Aとは、$(i,j)\in E$のとき、またそのときにかぎり第(i,j)エントリが1、それ以外のエントリは0であるような正方行列である。Gにおいて、頂点$i\in V$から出て行く有向枝の数をd_i^+、頂点$i\in V$へと入って行く有向枝の数をd_i^-と書くことにする。有向グラフGの頂点iから頂点jへの歩道(walk)とは、頂点の列$i,i_1,...,i_p=j$および有向枝の列$(i,i_1),(i_1,i_2),...,(i_{p-1},j)$で、頂点や有向枝が必ずしも相異なっていないもののことをいう。有向路(path)とは、相異なる頂点からなる歩道のことである。グラフGは、任意の$i,j\in V$について、iからjへの有向路と、jからiへの有向路が存在するとき、強連結(strongly connected)と呼ばれる。

グラフGの歩道の数とGの隣接行列Aの巾乗との間には、以下のような関係がある。

観察 1 A^lの第(i,j)エントリ$a_{ij}^{(k)}$は、頂点iから頂点jへの長さkの歩道の総数に等しい。

mを、下式を満たす非負の整数とする。

$$m=\min\left\{\begin{array}{l}\max_{i\in V} d_i^+,\\ \max_{i\in V} d_i^-,\\ \max_{i\in V}\sqrt{d_i^+ d_i^-}\}\end{array}\right\} \quad (1)$$

なお、3行目にある値は、[Kwa96]において与えられている有向グラフの隣接行列のスペクトル半径の上界である。そこで、実数rを$r=\frac{1}{m}$と定め、このrを使って行列Bを次のように定義する。

$$B\equiv\sum_{l=1}^{\infty}(rA)^l \quad (2)$$

下記は線形代数からの周知の結果である[BR97]。

命題 1 rが上のように定義されるとき、Bの定義式である和は収束する。

行列Bは、定義式(2)より、$n\times n$の単位行列をIとして$B=(I-rA)^{-1}-I$という式によって求めることができる。

行列Bの第(i,j)エントリb_{ij}は、あらゆる長さの歩道の本数を、歩道の長さが増大するにともなって指数関数的に減少する重み付けによって、加え合わせたものになっている。そこで、本研究ではこの値を、ドリフト **(drift)** と呼ぶことにし、ある頂点から別の頂点への移行のしやすさの定量的評価に用いる。実際、ある頂点を始点とする全ての歩道の集合に対して、その長さが1だけ増えるごとに生起頻度が$r=\frac{1}{m}$倍となるような頻度分布を考える。すると、この頂点から別の頂点へのドリフトは、この頂点からその別の頂点へと達する頻度に等しくなる。この意味において、ドリフトという概念は、ランダムウォークと同様、ネットサーフィンの一つのモデル化をもたらす概念である。

定義 1 頂点iからjへのドリフト$Dr(i,j)$とは、行列Bの第(i,j)エントリのことを言う。

2.2 相互リンク距離

上述のドリフトに基づき、Webページ間の近さdを下記のように定義する。

定義 2

$$d(i,j)\equiv-\log_m Dr(i,j)-\log_m Dr(j,i)$$

なお、頂点 i,j がいわゆる相互リンクだけで結ばれている場合、$d(i,j)=2$ となる。つまり、上に定義された近さは、相互リンクを典型例とする2頂点間の様々な相互関係について、その“親密さ”の度合いを表している。さらに、上に定義された近さは、三角不等式を満たす。

定理 1

$$d(i,j) \leq d(i,k)+d(k,j) \text{ for all } i,j,k \in V$$

証明 略。 □

つまり、この近さは距離と呼んでさしつかえない。以下、これを**相互リンク距離**と呼ぶ。なお、頂点 i から頂点 j への有向路が存在しないか、頂点 j から頂点 i への有向路が存在しない場合は、$Dr(i,j)=0$ または $Dr(j,i)=0$ となるため、相互リンク距離 $d(i,j)=\infty$ と定めることにする。また任意の $i \in V$ について $d(i,i)=0$ とする。

3 アルゴリズム

有向グラフの理論には、頂点のグループ化に利用できる概念として強連結成分分解がある。これは、与えられた有向グラフの、強連結な部分グラフへの分解のである。しかし、第1章でも述べたように、強連結成分分解はWebページ集合の粗野な分割しか与えないことが知られている。そこで、本研究では、強連結成分をさらに細分化する分解を与えるアルゴリズムを提案する。さらに、このアルゴリズムは、一つのパラメータを増減させることで、結果的に得られるWebページのグループの大きさを調整することを可能にする。すなわち、単に強連結成分分解を細分化するだけでなく、その細分化の程度を制御することができるようになっている。そこで、こうして得られるグラフの分解を、**パラメータ化された連結成分分解**と呼ぶことにする。

今回の実験では、次に示すアルゴリズムによって、パラメータ化された連結成分分解を得た。

for each $i \in V$ **do**
  Do breadth first search from i
  and Compute drift to every other page;
$C := \{1\}$;
for each $i \in V \setminus C$ **do**
  if $d(i,j) \geq \tau$ holds for all $j \in C$ **then**
    $C := C \cup \{i\}$;
for each $i \in V \setminus C$ **do**
begin
  Find $j \in C$ nearest to i;
  $PCC(j) := PCC(j) \cup \{i\}$;
end.

ドリフトの計算

最初の for ループでは、任意の Web ページから、他のすべての Web ページへのドリフトを求めている。ドリフトの算出には、二通りの方法がある。

1. 行列 B の定義式(2)を利用し、$B=(I-rA)^{-1}-I$ によって全ドリフトを直接求める方法。
2. 各 Web ページを始点とする歩道を、当の頂点からの幅優先探索によって一つずつ枚挙しながら、ドリフトを累積的に算出する方法。

第一の方法については、逆行列の計算には $O(n^3)$ よりも本質的に少ない計算量を持つアルゴリズムが存在する [PFTV88, Ch. 2.11] し、より実践的には、疎行列に関して数値計算の分野で提案されている洗練された逆行列計算の手法を、直接適用することもできる [BT98]。だが、今回の予備実験では、第二の方法を選択した。一つには実装が容易であるという理由もあるが、次のような別の理由もある。つまり、各 Web ページからの幅優先探索を、一定の深さまで進んだ段階で打ち切れば、相対的に長い歩道の寄与を無視するのと引き替えに、全体の処理時間を短縮できる。さらには、ドリフトを得るための幅優先探索は、各 Web ページ毎に完全に独立に行うことができるため、処理の並列化が比較的容易である。しかしながら、20億の Web ページに対する Hub/Authority スコア [Kle99] の計算に成功している研究 [安村 02] もすでにあるため、第一の数値計算的な手法がより現実的な選択である可能性を否定することはできない。この点についての検討は、今後の課題である。

グループの中心となるページの選定

さて、第二の for ループでは、与えられた頂点集合から、結果として得られるグループの中心となるものだけを枚挙している。グループの中心となる頂点は、以下のようにして選び出している。まず、適当な Web ページ（ここでは通し番号として1を与えられている Web ページ）をアド・ホックに中心と定

め、以下、すでにグループの中心として登録されている頂点のすべてから、閾値として用いられるパラメータτ以上に離れているものだけを、新たな中心として登録する。

$\tau = \infty$のとき、アルゴリズムの与えるグループ化は、強連結成分分解に一致する。そして、τを徐々に減少させることで、強連結成分分解よりも段階的にグループの粒度が細かくなっていくようなグループ化が実現される。

グループの構成

そして、第三のforループにおいて、残った頂点を、それに最も近い中心の配下にあるグループの構成員として登録していく。なお、$PCC(i)$とは頂点iを中心とするWebページのグループである。

τの値を変更すれば、結果として得られるグループの大きさが変化しうる。なぜなら、τを増大させると、グループの中心として登録される頂点が減り、頂点集合はより少ないグループへと分割されることになるからである。また、$\tau = \infty$という極限においては、以下に示すように、強連結成分分解が得られる。

定理 2 上記アルゴリズムは、$\tau = \infty$のとき、グラフGの強連結成分分解を与える。

証明 $\tau = \infty$とし、2頂点i, jが、$d(i,j) < \infty$を満たし、かつ、上のアルゴリズムによって得られた相異なるグループに属したと仮定する。iの属するグループの中心である頂点をi_C、jの属するグループの中心である頂点をj_Cとすると、中心となる頂点の選び方より$d(i_C, j_C) = \infty$である。ところが、定理1より、$d(i_C, j_C) \leq d(i, i_C) + d(i,j) + d(j, j_C)$が成り立たねばならない。よって、$d(i, i_C) = \infty$、$d(j, j_C) = \infty$の少なくとも一方が成立する。

$d(i, i_C) = \infty$の場合、頂点iがどのグループの中心頂点でもないことから、$d(i,k) < \infty$を満たし、かつ、いずれかのグループの中心頂点として登録されている頂点kの存在が帰結する。しかし、この帰結は、頂点iが頂点i_Cを中心とするグループに属しており、かつ$d(i, i_C) = \infty$が成立していることに矛盾する。$d(j, j_C) = \infty$の場合も同様に議論できる。□

さらに、上記アルゴリズムによって得られるグループについて、以下の事実を証明することができる。

定理 3 上のアルゴリズムによって得られる任意のグループについて、それに属する任意の2頂点i, jの距離$d(i,j)$は、

$$d(i,j) \leq 2\tau$$

を満たす。

証明 i, jの属するグループの中心として登録されている頂点をkとすると、定理1より、$d(i,j) \leq d(i,k) + d(j,k)$が成り立つ。$d(i,k) > \tau$と仮定すると、$i$が頂点$k$を中心とするグループに属していることより、他の任意のグループの中心である頂点k'について、$d(i,k') > \tau$が成り立つ。ところが、このことは、iがいずれのグループの中心としても登録されていないという事実に反する。したがって、$d(i,k) \leq \tau$が成立しなければならない。$d(j,k) \leq \tau$についても同様に議論すれば、所望の結論を得る。□

4 予備実験

今回、特定のWebページからのクローリングによって得られたちょうど3万のWebページを予備実験の対象とした。実験に際しては、一台のワークステーション（Sun Blade 1000。CPUはUltraSPARC-IIIの750MHzと900MHz。メモリ8192Mバイト。）上ですべての処理を行っている。ドリフトを求める際に使う値$r = \frac{1}{m}$は式1によって求めた。今回のデータでは$m = 567$となった。

```
www.kirihara.co.jp/index.html
www.kirihara.co.jp/tm21/inquiry.html
www.kirihara.co.jp/tm21/q_and_a.html
www.kirihara.co.jp/tm21/index.html
www.kirihara.co.jp/toeic/menu.html
www.kirihara.co.jp/toeic/index.html
www.kirihara.co.jp/menu.html

www.kirihara.co.jp/textbook/index.html
www.kirihara.co.jp/textbook/eigo/eigo_index.html

www.kirihara.co.jp/scope/APR2001/apr2001.html
www.kirihara.co.jp/scope/KIKO/kiko.html
www.kirihara.co.jp/scope/KYOYO/kyoyo.html
www.kirihara.co.jp/scope/TANBO/tanbo.html
www.kirihara.co.jp/scope/TEMALIST/temalist.html
```

図1: グループの例1 順に7、2、5個のWebページからなる。

まず、具体的にどのようなWebページからなるグループが得られたかを示す。ここに紹介する結果は、$\tau = 10$とした場合のものである。図1を見ると、一つのサイト内が、あたかもURLに関するヒューリスティクスを利用したかのようにグループ化されていることが分かる。しかし、単純にサイト名の直後の最初のディレクトリ構造にしたがってグループ化されているわけではない。それは、同じ条件の下で

相互リンク距離が三角不等式を満たすという事実は、本研究に以下のような発展性をもたらす。つまり、全く異なる観点からグループ化を行う可能性が開かれる。

例えば、[DMR+00] において提案されている、無向グラフの頂点をクラスタリングするためのアルゴリズムは、事前にクラスタの個数を指定するかたちで、クラスタリングを実現する。また、得られるクラスタの直径（同じクラスタ内に属する頂点を結ぶ枝の重みの最大値）の和をできるだけ小さくするという意味において性質の良いクラスタリングを与える。だが、このアルゴリズムを適用するためには、枝の重みが三角不等式を満たす必要がある。

そこで、本研究の提案する相互リンク距離を利用すれば、リンク情報のみに基づくWebページのグループ化という問題が、無向グラフ上での頂点のクラスタリング問題へと変換される。つまり、Webページを無向グラフの頂点とみなし、相互リンク距離 d を、頂点間に張られた枝の重みとみなすのである。

もちろん、これでは、今回提案したアルゴリズムのように、強連結成分分解の細分化としてのグループ化は得られない。だが、本研究の提案する相互リンク距離は本稿で提案したアルゴリズムにしか使えないわけではないという事実が、Webページのグループ化の問題を、様々なタイプのグラフ上のクラスタリング問題へと変換する可能性を開いている。実際、上掲論文にあるアルゴリズムを使えば、予めグループの個数を指定するというかたちでWebページのグループ化を行うことができる。

しかし、WWW上での効果的な情報検索実現のためのWebページのグループ化という目的には、いずれのグラフ・クラスタリングの手法が適しているのかについては、下記のスケーラビリティなど、実装上の問題とも絡めて今後さらに調査されねばならない。

5.2 今後の課題

WWW上での情報検索を実現するためには、第一に、スケーラビリティを確保することが重要である。本論文において提示された枠組みでは、行列 B、すなわち、有向グラフの隣接行列の母関数をいかに高速に計算するかという点が、スケーラビリティの確保に最も大きく関わる。この問題をクリアするには、各ノードから始まる幅優先探索を並列化することが有効なのか、あるいは、逆行列を数値計算的に洗練された方法で求めることが有効なのかを明らかにすることが、今後の課題として残されている。

また、Webページをグループ化するためのアルゴリズムは、理想的には、クローリングと同時に実行できるものが良い。残念ながら、今回提案したアルゴリズムにおいては、処理の対象となるWebページがあらかじめすべて与えられているのでなければならない。クローリングとグループ化とを同時進行させるためには、グループ化のアルゴリズムがどのようなものでなければならないか。この点についての考察もまた、重要な課題である。

謝辞

本研究は、文部科学省科学研究費補助金特定領域研究「情報学」（課題番号 13224087）の助成のもとに行われた。

参考文献

[BKM+00] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of 9th WWW Conference*, pp. 309–320, 2000.

[BR97] R. B. Bapat and T. E. S. Raghavan. *Nonnegative Matrices and Applications*, Vol. 64 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, 1997.

[BT98] Michele Benzi and Miroslav Tůma. A comparative study of sparse approximate inverse preconditioners. Technical Report LA-UR-98-0024, Los Alamos National Laboratory, 1998.

[CPKT92] Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318–329, 1992.

[DMR+00] Srinivas Doddi, Madhav V. Marathe, S. S. Ravi, David Scot Taylor, and Peter Widmayer. Approximation algorithms for clustering to minimize the sum of diameters. In *Scandinavian Workshop on Algorithm Theory*, pp. 237–250, 2000.

[Kat53] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, Vol. 18, pp. 39–43, 1953.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632, 1999.

[Kwa96] Jaroslaw Kwapisz. On the spectral radius of a directed graph. *Journal of Graph Theory*, Vol. Vol. 23, No. No. 4, pp. 405–411, 1996.

[PFTV88] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C: The art of scientific computing.* Cambridge University Press, Cambridge, UK, 1988.

[THA99] Loren Terveen, Will Hill, and Brian Amento. Constructing, organizing, and visualizing collections of topically related Web resources. *ACM Transactions on Computer-Human Interaction*, Vol. 6, No. 1, pp. 67–94, 1999.

[安村 02] 安村賢英, 川原稔, 岩下武史, 金澤正憲. Web コミュニティ発見のための大規模有向グラフに対するデータ圧縮計算手法の vpp への実装. 京都大学大型計算機センター研究開発部 研究発表報告集, 第 17 号, pp. 71–80, 2002.

[小島 02] 小島秀一, 高須淳宏, 安達淳. Web ページ群の構造解析とグループ化. *NII Journal*, Vol. 4, pp. 23–35, 2002.

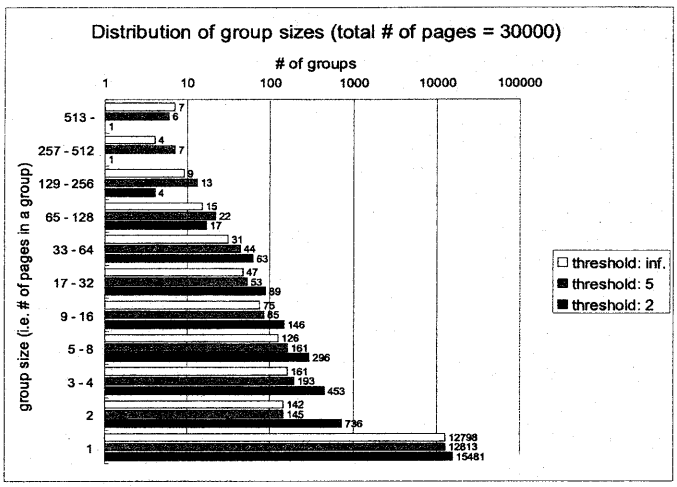

図 3: 閾値が **2, 5,** 無限大それぞれの場合のグループの大きさの分布

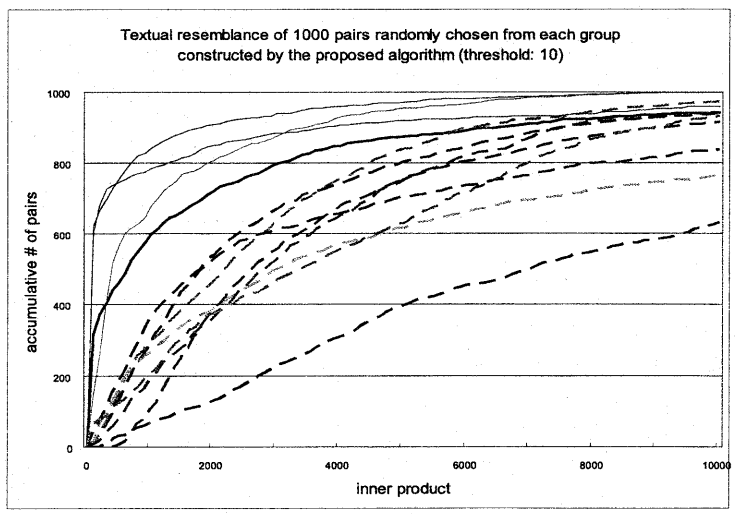

図 4: 構成された **Web** ページグループのテキスト情報による評価