リンク情報の利用による Web 検索性能の改善

正田備也[†] 高須淳宏[†] 安達淳[†]

本研究は，リンク情報を利用して Web 検索性能を向上させる効果的な手法に関する研究である．まず，新しいクラスタリング・アルゴリズムを提案する．このアルゴリズムは，同じサイトに属する Web ページを結ぶハイパーリンクだけを利用し，出次数の多い Web ページが異なるクラスタに分散するようなクラスタリングを実現する．これによって，同じクラスタ内でテキスト情報の均一性が適度に確保されることを狙っている．なぜなら，出次数が多い Web ページをたくさん経由するほど，Web ページのテキスト内容が発散しやすいと考えられるからである．本研究では，この仮説を，提案のクラスタリング・アルゴリズムが Web 検索の性能向上に寄与するかどうかを確認することで，検証する．そこで，提案のアルゴリズムによって得られたクラスタを利用し，各 Web ページのテキスト情報をもとに算出された文書ベクトルのエントリを変更する．文書ベクトルは，代表的な単語重み付けスキーマである TF-IDF によって計算され，文書ベクトルのエントリの変更は，金沢らによって提案された RS モデルに基づいて行われる．本研究では，検索性能を客観的に評価するため，NTCIR-3 Web 検索タスクのために準備された文書データと検索質問を，評価実験に用いた．実験の結果によれば，ワン・クリック・ディスタンス文書モデルの下で，クラスタリングの結果を用いない場合に比べて，検索性能を表す重要な指標である平均適合率が 10%以上上昇した．

Improving Web Search Performance with Hyperlink Information

Tomonari Masada,[†] Atsuhiro Takasu[†] and Jun Adachi[†]

This paper concerns an efficient method for improving Web search performance with hyperlink information. We provide a new Web page clustering algorithm. Our algorithm only uses intra-site hyperlinks and constructs clusters so that the Web pages of large out-degree belong to different clusters. We expect our algorithm to provide clusters such that the Web pages in the same clusters are similar to each other by their textual contents. This algorithm is based on a hypothesis that the textual contents of Web pages tend to drift further after passing through more Web pages of larger out-degree. In this paper, we test this hypothesis by checking if our clustering algorithm can improve the performance of Web search. We use clustering results our algorithm gives and modify entries of document vectors. Document vectors are computed with a well-known term weighting scheme, TF-IDF. The vector entry modification is based on RS (relevance superimposition) model invented by Kanazawa et al. We conducted evaluative experiments by using document sets and query sets prepared for NTCIR-3 Web retrieval task and realized an objective evaluation. The results show that when we use one-click-distance document model, we can improve the average precision, an important measure for Web search performance, on the order of more than 10% in comparison with the case where we use no clustering results.

1. はじめに

WWW（World Wide Web）は，膨大な情報の貯蔵庫であり，サーチ・エンジンで，欲しい Web ページを見つけ出すことは難しい．そのため，検索結果の中には欲しかった Web ページそのものが含まれず，検索結果として与えられた Web ページ上のリンクをクリックしてはじめて見たかったページにたどりつく，といったことも起こる．しかし，WWW の巨大さを考えれば，検索結果に現れる Web ページの上にあるリンクをクリックすることで欲しい Web ページにたどりつけるならば，実用上問題ない，と考えられる．検索エンジンは，ユーザの質問に適合する Web ページを直接に提示できなくてもよく，適合するページへと通じるハイパーリンクを含む Web ページを提示できれば，十分に効果的ともいえる．欲しいページが直接与えられなくても，リンクをクリックすることでユーザが能動的に Web ページを探せることは，WWW という文書集合に特有の新しい検索のあり方ではないだ

† 国立情報学研究所
National Institute of Informatics

ろうか．本研究の独自性は，ユーザが探している Web ページそのものではなく，それへのリンクを持つ Web ページを，従来のサーチ・エンジンのように偶然的ではなく，意図的に検索結果に含ませるような，Web 検索の新しい手法の提案にある．

2. 関 連 研 究

テスト・コレクションを用いた Web 検索のコンペティションでも，これまでの検索性能評価のように，検索アルゴリズムの与える検索結果そのものを評価するだけでなく，検索結果として与えられた Web ページ上のリンクを経由してたどりつける Web ページもまた，評価に含ませるような評価モデルが，提案されている．そのよい例が，NTCIR-3 の Web 検索タスクで用いられている，ワン・クリック・ディスタンス文書モデル（one-click-distance document model）[3),4)] である．この評価モデルによれば，検索結果に含まれる Web ページのうち，そこからリンクを 1 回クリックすることで適合ページにたどりつける場合は，そのページ自身も適合すると見なされることがある（**図 1**）．もちろん，適合ページへのリンクを持っていさえすればつねに適合と判定されるわけではなく，そこには，検索者が欲しがっている Web ページへの案内役として適切な Web ページかどうかという判断が入っている．ワン・クリック・ディスタンス文書モデルに基づく正解集合も，このような判断をふまえて作成されている．したがって，既存の検索手法をワン・クリック・ディスタンス文書モデルに対応させようとして，単にその手法が検索結果として出力する Web ページへのリンクを持つ Web ページを，取捨選択なしに検索結果に混入させるという素朴な方法では不十分である．ここに，新しい検索手法を探る余地がある．なお，NTCIR Web 検索タスクでは，従来の文書モデル，すなわち，各 Web ページ単独でその適合性を判定するモデルは，ページ単位文書モデル（page-unit document model）と呼ばれている．

ところが，NTCIR-3 に参加した研究のうち，ワン・クリック・ディスタンス文書モデルとページ単位文書モデルとで，評価上大きな差を出すことに成功しているものはない．その一方，本研究では，検索質問に適合する Web ページへのリンクを持つ Web ページを，意図的に検索結果に含ませるような検索手法を提案する．そして，NTCIR-3 のために準備された文書データおよび検索質問を用いた実験によれば，提案手法の与える検索結果は，実際に，ワン・クリック・ディスタンス文書モデルによる評価と，ページ単位文書モデルによる評価とで，大きな差を示す．この結果から，従来のページ単位文書モデルによる評価では的確に評価できないが，現実の Web 検索では効果を発揮することができるまったく新たな種類の検索手法が，本研究によって考え出されているといえる．

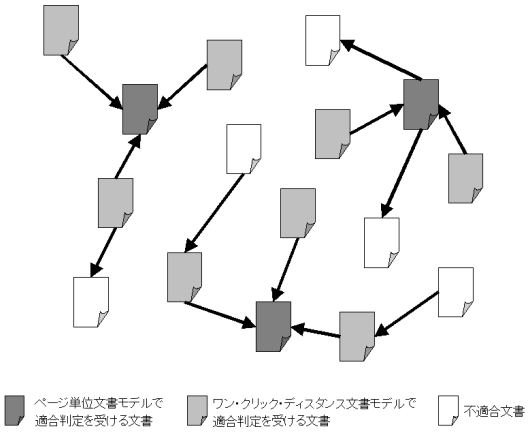

図 1 ワン・クリック・ディスタンス文書モデル
Fig. 1 One-click-distance document model.

“WWW は巨大であるから，適合ページへのリンクを持つ Web ページを提示することは有用である” という観点に基づく従来研究には，次のようなものがある．

Kleinberg [9)] は，与えられた検索質問について，任意のサーチエンジンで得られた検索結果を体系的な方法でふくらませ，そうして得られた Web ページ集合のリンク情報から，ハブ/オーソリティ・スコアという 2 種類の Web ページの重要度を計算する手法を提案している．これによって，適合するページそのものではなく，リンク構造上それと関連する Web ページの中で性質の良いものを提示することができる．しかし，ハブ/オーソリティ・スコアだけでは，他の検索手法と組み合わせないかぎり，Web 検索において満足な性能が得られないという報告がある[17)]．さらに，たとえ補助的なスコアリング手法としてハブ/オーソリティの枠組みを使うとしても，検索結果をいったん求めた後に，ハブ/オーソリティ・スコアの計算を含めた複雑な処理が必要となる．その一方，本研究の提案する手法は，事前にバッチ処理的に Web ページ集合に対してクラスタリングを行っておき，特定の検索質問を想定しない仕方で，適合ページへのリンクを持つ Web ページの中から良い Web ページを選ぶという課題に対処している．そのため，処理時間のうえで有利である．もちろん，ハブ/オーソリティ・スコアであっても，特定の検索結果に対応するかたちではなく，与えられた Web ページ集合全体を対象に，事前に計

算しておくことができる．しかし，その場合には，ハブ/オーソリティ・スコアは，検索質問が何であるかによらない Web ページの重要度となり，ハブ/オーソリティ・スコアの高いページのうち，個別の検索質問に対してどのページをより適合するものとして選び出すのか，という別の問題が生じる．たとえば，ある 1 つの質問に適合する Web ページへと，複数のハブ・ページがリンクしていたとき，質問に適合するページそのものも含めて，これら多数の Web ページにどのような順位を与えて検索結果として表示すればよいだろうか．この問いに自明な答えはなく，これ自体が 1 つの検討課題となる．その点，提案手法では，Web ページのクラスタリングに基づいて，Web ページの特徴量として求められた文書ベクトルそのものを変化させる．すると，個別の質問に応じて，各 Web ページについて，クラスタリングの情報を使わない場合とは異なる新たなスコアを得ることになる．このように，クラスタ情報を使って直接的に Web ページの順位付けを変化させる点が，本研究の特徴である．

適合ページにリンク構造上で関連している Web ページを提示することをねらった研究には，同じサイト内の検索結果を束ねて表示するもの，関連する Web ページを参考ページとして添えて表示させるものなど，検索結果の提示方法を改良しようとするものもある[15),16)]．しかし，これらの研究は，情報の提示の仕方に着目しているため，たとえば，検索結果の長さがどれだけ短縮されたかや，ユーザがどれだけ大きな利便性を感じたかなど，検索性能の良し悪しとは直接関係しない尺度で研究の評価をすることになる．しかし，本研究では，検索結果の提示手法をいままでの Web 検索とは違うものにするとよい，という意味で，適合ページへのリンクを持つ Web ページを提示することは有用だ，と考えているのではない．むしろ，リンク構造を使ったクラスタリングによって，適合ページへのリンクを持つ Web ページのうち，検索者の役に立つと思われるものを，検索結果のランキングの中で，いままでの Web 検索よりも上位におくことを考えている．つまり，検索結果の外観はそのままに，検索結果の新たなランキングを提示する．そのため，提案手法の評価において，情報検索における標準的な評価方法をそのまま引き継ぐことができる，という利点が出てくる．Web ページのクラスタそのものに対してスコアづけをし，クラスタを 1 つの情報の単位として表示させるという手法を提案した論文もある[19)]．これは，広い意味で，検索結果の新たなランク付け手法といえる．しかし，個別の Web ページを単位として検索する場合に使われる評価方法とは異なる評価方法を考え出さなければならない，という問題点がある．実際，この論文では，個々のクラスタをまとめて 1 つの検索の単位として検索者に提示してしまうのでは，評価対象のデータ母集団が通常の検索と異なるため，「評価の場」が違ってきてしまい，実質的な評価ができない，と論じられている．ほかにも，リンク構造を利用して，個々の Web ページとは異なる検索の単位を構成することで，検索結果の新たな表示手法を提案する研究がある[18)]．だが，この研究でも，検索性能そのものの評価はせず，使用したうえでの感想という主観的な評価を示したり，検索結果の抜粋を論文中に提示して評価を論文の読者にゆだねたりするなどしている．これもまた，従来の情報検索の評価方法を踏襲できないためである．その一方，本研究の大きな利点は，Web ページのクラスタリングによって，検索の提示方法や検索の単位を変えることではなく，検索結果の新たな順位付けを行うことをめざしている，という点にある．このため，従来の Web 検索と共通の評価方法を使えるようになっている．

しかし，リンク情報を利用したクラスタリングによって，直接に検索結果のランキングを改善しようとする研究は，すでに杉山らによって行われている[17)]．この研究もまた，本研究のように，クラスタリングの結果を利用して文書ベクトルを変更し，変更後の文書ベクトルによって各 Web ページのスコアを計算することで，検索性能を向上させようとしている．ただし，クラスタリングには文書ベクトルを対象とした K-平均法[14)] を使っている．クラスタリングの対象となる Web ページの絞り込みにリンク情報を用いているとはいえ，実質的にはテキスト情報を使ったクラスタリングである．その一方，本研究では，クラスタリングにあたって，各 Web ページについて求められた文書ベクトルは参照せず，リンク情報だけを使っている．そのため，文書ベクトルの算出に必要な処理とは独立に，Web ページをクラスタリングできる．したがって，杉山らの研究とは，検索性能の改善幅とそれに必要な処理のていねいさとの間のどこでバランスをとるか，という問題について，違う立場をとっている．杉山らの研究では，170 万ページからなる中規模の，しかも英語テキストの Web ページ集合を実験に使っているが，より大きな文書集合で，しかも語彙数の多い日本語のテキストの場合に，提案されている手法が十分なスケーラビリティを発揮できるかという問題が残るように思われる．なぜなら，このとき，ベクトルの個数および次元がともに増加するからである．

3. 提案の Web 検索手法について

3.1 ベクトルとして表現された特徴量

与えられた Web ページの集合を V とする．Web ページの最大の特徴は，お互いがハイパーリンクによってつながり合っていることである．そこで，ハイパーリンクの集合を，2 つの Web ページの順序対の集合 E で表す．Web ページ $v_1 \in V$ から $v_2 \in V$ へのリンクがあることを，$(v_1, v_2) \in E$ と表す．WWW は，Web ページ集合 V とリンク集合 E の二つ組み (V, E) として表される．Web ページはテキスト情報を含んでもいる．本研究では，どのような単語が何回出てくるかを，各 Web ページにおける個々の単語の重みを計算するために使う．そこで，Web ページ集合 V に含まれる語彙の集合を T_V とし，Web ページ $v \in V$ に単語 $t \in T_V$ が含まれる回数を $\mathrm{TF}_v(t)$ と書く．“TF” は term frequency の略である．さらに，各単語 $t \in T$ について，それが含まれる Web ページの個数を，Web ページ集合全体での個々の単語の重要度の指標として使う．この値は，通常 document frequency と呼ばれるので，$\mathrm{DF}(t)$ と書く．

まず，提案手法を適用する際には，各 Web ページについて，その特徴量として文書ベクトル（document vector）が計算されているとする．今回の実験では，形態素解析器 MeCab☆を ipadic-2.5.1 とともに使って，各 Web ページの日本語のテキスト内容を単語の集まりへと分割した後，標準的な単語重み付けスキーマである TF-IDF [1] で文書ベクトルを算出した．ここで，文書ベクトルの次元は，Web ページ集合に現れる語彙の数 $|T_V|$ に一致する．そして，Web ページ $v \in V$ に対応する文書ベクトル $\mathbf{x}_v$ では，単語 $t \in T_V$ に対応するエントリ $\mathbf{x}_v(t)$ が，単語 t の文書 v における重みを示す値をとる．具体的には，本研究では，以下の式を用いて $\mathbf{x}_v(t)$ の値を算出した．

$$\mathbf{x}_v(t) \equiv (1+\log(1+\log \mathrm{TF}_v(t))) \cdot \left(\frac{|V|}{\mathrm{DF}(t)}\right)^{\frac{1}{5}} \quad (1)$$

なお，この式は，準備段階の実験において，できるだけ良い検索性能が出るようにチューニングされたものである．たとえば，式の前半の $1+\log(1+\log \mathrm{TF}_v(t))$ という項では，準備段階の実験において，TF の値の増大にともなう単語重みの増大はできるだけ緩やかにしたほうが検索性能が良くなると分かったので，対数関数を二重に用いた．また，$|V|/\mathrm{DF}(t)$ を 1/5 乗するという項も，1/2 乗，1/3 乗，1/4 乗，1/6 乗など，いくつかの値を試したうえで 1/5 乗に決めた．このように，ベースラインとなる検索性能をできるだけ上げておいてはじめて，提案手法を適切に評価できる．

☆ http://chasen.org/~taku/software/mecab/

3.2 クラスタリング結果を利用した特徴量の変更

次に，各 Web ページについて求められた文書ベクトルのエントリを，その Web ページが他の Web ページとともにどのようなクラスタを形成しているかに応じて，変更する．本研究では，独自のクラスタリング・アルゴリズムを提案し，ページ単位文書モデルではなく，Web 検索の特殊性を反映したワン・クリック・ディスタンス文書モデルの下での検索性能評価において良い成績をおさめられるように工夫している．このアルゴリズムは，Web ページ間のハイパーリンクを介した参照関係だけを使う．よって，各 Web ページからリンクを抽出する作業が終わっていさえすれば，3.1 節の文書ベクトルの計算のような各 Web ページのテキスト内容に関わる処理とは，まったく独立にクラスタリング処理を実行できる．ところで，Web ページからのリンク抽出は，クローリング時にすでに実行されており，検索システムを構築するために事後的に行う処理ではない．したがって，クラスタリングにリンク情報しか用いないということは，検索システムのパフォーマンス向上に寄与する特徴である．さらに，このアルゴリズムは，サイト内部のリンク情報だけを用い，異なるサイトにある Web ページへつながるリンクは用いない．よって，異なるサイト上でのクラスタリングは独立に実行でき，クラスタリング処理を並列化しやすい利点が生じる．本論文で “サイト” とは，URL において “`http://`” 以降初めて現れる “`/`” までが一致する Web ページの集合を意味する．

本研究が提案する新しいクラスタリング手法を紹介する前に，まず，クラスタリングの結果を用いてどのように文書ベクトルを変更するか，を説明する．この手法は，Kanazawa ら[6)~8)] の提案した RS モデルの一種であり，どのような手法でクラスタリングを行ったかによらず，適用可能な操作である．

クラスタリング結果によって文書ベクトルのエントリを変更する際には，まず，各クラスタの特徴量として，**代表ベクトル**というベクトルを計算する．具体的には，クラスタ $C \subseteq V$ の代表ベクトルを $\mathbf{y}_C$ と書くことにすると，このベクトルにおける単語 $t \in T_V$ に対応するエントリ $\mathbf{y}_C(t)$ は，次の式で計算される．

$$\mathbf{y}_C(t) \equiv \max_{v \in C} \mathbf{x}_v(t) \quad (2)$$

そして，クラスタに属する Web ページの文書ベクトルに，この代表ベクトルを線形混合する．文書 $v \in V$

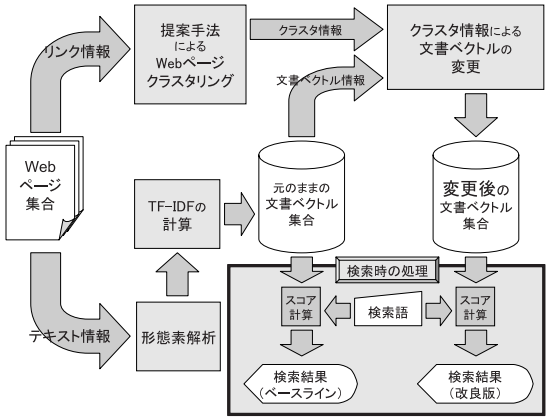

図 2 提案のクラスタリング手法を使った Web 検索システム
Fig. 2 Web search system using our clustering method.

の変更後の文書ベクトルを $\mathbf{x}'_v$ とすると，ベクトル $\mathbf{x}'_v$ における単語 $t \in T$ に対応するエントリ $\mathbf{x}'_v(t)$ は，

$$\mathbf{x}'_v(t) \equiv (1-\alpha)\mathbf{x}_v(t) + \alpha \mathbf{y}_C(t) \qquad (3)$$

と計算される．ここで，α は，$0 \leq \alpha \leq 1$ を満たす実数であり，この値が 1 に近づくほど，クラスタリング結果による文書ベクトルの変更が強くはたらく．そこで，α を**クラスタ情報混合率**と呼ぶ．なお，複数のクラスタに属する Web ページの場合は，以下のようにして文書ベクトルを変更する．問題の Web ページを $v \in V$ とし，与えられた 1 つのクラスタリング結果 $\mathcal{C} \subseteq 2^V$ において v が属するクラスタの集合を $\mathcal{C}_v \equiv \{C \in \mathcal{C} : v \in C\} \subseteq \mathcal{C}$ とすると，

$$\mathbf{x}'_v(t) \equiv (1-\alpha)\mathbf{x}_v(t) + \alpha \max_{C \in \mathcal{C}_v} \mathbf{y}_C(t) \qquad (4)$$

と，文書が属するクラスタすべてにわたって，各エントリの最大値をとり，それを α の割合で，元の文書ベクトル $\mathbf{x}_v$ の各エントリに線形混合する．

最後に，以上のようにして変更を加えられた文書ベクトルを使って，検索質問に対する各 Web ページのスコアを計算する．今回の実験では，TF-IDF スキーマに基づいて文書ベクトルを求めているので，検索質問についても，同じ式 (1) を使ってベクトル表現を求め，変更後の文書ベクトルとの内積の値を，対応する Web ページのスコアとする．以上のような，提案のクラスタリング手法を使った Web 検索システムの構成を，**図 2** に示した．こうして各 Web ページについて計算されたスコアの高低が，その Web ページの検索質問に対する適合・不適合によく合致するならば，提案の検索手法は優れているといえる．そこで，本研究では，NTCIR-3 Web 検索タスクで実際に用いられた文書データ，検索質問，および適合判定のための正解データで実験を行い，提案手法を評価した．

3.3 新しいクラスタリング手法

本研究では，リンク情報のみを利用した新しい Web ページ・クラスタリング手法を提案する．そこでは，一定の順序で Web ページを選び，選ばれた Web ページを，これから構成するクラスタの中心ページとする．そして，この中心ページから他の Web ページへの最短パス長を計算し，その最短パス長が，与えられたパラメータ τ 以下の Web ページだけを，中心ページと同じクラスタに属すると決める．パラメータ τ は，以下，**閾値パラメータ**（threshold parameter）と呼ぶ．

3.3.1 クラスタ中心ページ選出の順序

中心ページは，THP（Two Hop Return Probability）[12] という値が大きい順に選ばれる．Web ページ v の THP は，次の式で定義される．

$$\mathrm{THP}_v \equiv \sum_{u \in V \text{ s.t. } (v,u) \in E} \frac{1}{d_v^+} \cdot \frac{1}{d_u^+} \qquad (5)$$

d_v^+ は，Web ページ v から出て行くハイパーリンクの数で，Web ページ v の**出次数**と呼ばれる．THP は，WWW のリンク構造の中で，リンクが密に張られている部分の中心的な位置にある Web ページにおいて大きな値をとる．ただし，どの Web ページも，中心ページとしてであれ，別の Web ページを中心ページとするクラスタのメンバとしてであれ，いずれかのクラスタに属すると定められた時点で，いくら THP の値が大きくても，それ以後，中心ページとして選ばれることがないようにする．もちろん，中心ページとして選ばれることがないだけで，後から他の Web ページを中心ページとして行われたクラスタ構成において，そのメンバとして認定されることはある．いい換えれば，1 つの Web ページが，複数のクラスタに属することはある．こうして，すべての Web ページが，少なくとも 1 つのクラスタに属するようになるまで，処理を続ける．以下，アルゴリズムについて，さらに詳しく説明する．

THP の大きい順にピック・アップされた中心ページから，他の Web ページへのパス長は，パスに沿って存在する Web ページの出次数の和として定義される．したがって，大きい出次数を持つページを数多く経由するパスほど，長いパスとされる．この定義によれば，出次数の大きい Web ページは互いに遠く離れ，同じクラスタに属しにくくなる．以上のように，出次数でパス長を定義したのは，

(1) 出次数を，Web ページのテキスト内容の発散の度合いを表す指標と見ることができるため．いい

換えれば，出次数の大きいページを数多く経由するほど，内容的に大きく異なる Web ページへ移行しやすくなる，と考えられるため，

(2) 出次数は，クローリングで Web ページを収集しているときに得られる情報であり，それを求めるのに事後的な計算コストが要らない特徴量であるため，

以上 2 つの理由による．(1) の理由は，クラスタの質に関係する．これは，本研究が立てる仮説である．提案のクラスタリング手法が，どの程度，検索性能の向上に寄与するかを，実験で確かめて，この仮説が正しいかどうかを明らかにする．なお，今回は，サイト内のリンク情報しか使っていない．そのため，サイト間を結ぶリンクも含ませると，(1) の仮説にどのような影響が出るか，という問題も出てくる．実際，本研究の前段階で行った一連の実験では，サイト間のリンクも使ってクラスタリングを行った[20]．だが，その実験と比較して，Web 検索の性能が改良されることはなかった．そのため，3.2 節で述べたように，クラスタリングの並列化が簡単になることを考えれば，サイト内のリンクだけを使うことは良い選択だと考えられる．(2) の理由は，計算量に関係する．出次数よりもさらに複雑な Web グラフの特徴量，たとえばハブ・スコアなど[9] を用いるという選択肢もありうるが，今回の評価実験から，出次数でパス長を定義することは，計算量とのバランスを考えた良い選択だと考えられる．

3.3.2 粒度制御をともなう 3 種類のクラスタリング

次に，中心ページを通るパスとしてどのようなパスを考えるかで，3 つの場合を考える（**図 3** 参照）．

(1) 中心ページを終点とするパス．他の Web ページから発して，中心ページで終わるパス．

(2) 中心ページを始点とするパス．中心ページから発して，他の Web ページで終わるパス．

(3) 中心ページが，始点であると同時に，終点になっているパス．この場合，パスは，中心ページを通過する閉路（cycle）となる．

(1) の場合は，中心ページからリンクを逆にたどるかたちで，他の Web ページへの最短パスを求め，パスに沿って存在する Web ページの出次数の総和が閾値パラメータ τ 以下となるパスだけを列挙し，それらパス上にある Web ページをクラスタにまとめる．この場合，中心ページへのファン・イン（fan-in）としてのクラスタを構成することになっている．このケースは，同じ Web ページに向かうパス上には，互いに内容的に関連する Web ページが多いだろう，という仮説に基づいて考え出されている．(2) の場合は，中心ページからリンクを順方向にたどって長さ τ 以下の最短パスを列挙，それらのパスの上にある Web ページを同じクラスタにまとめる．この場合，中心ページからのファン・アウト（fan-out）としてのクラスタを構成している．このケースは，同じ Web ページから出てくるパス上には，内容的に関連する Web ページが多いだろう，という仮説に基づいて考え出されている．(3) の場合は，そこから中心ページへと至る最短パスの長さと，中心ページからそこへと至る最短パスの長さとの和が，τ 以下になるような Web ページを列挙し，同じクラスタにまとめる．つまり，中心ページを通過する一定の長さ以下の閉路（cycle）の束として，クラスタを構成することになる．このケースは，同じ Web ページから出て，そしてそこに戻っていくパス上には，内容的に関連する Web ページが多いだろう，という仮説に基づいて考え出されている．要す

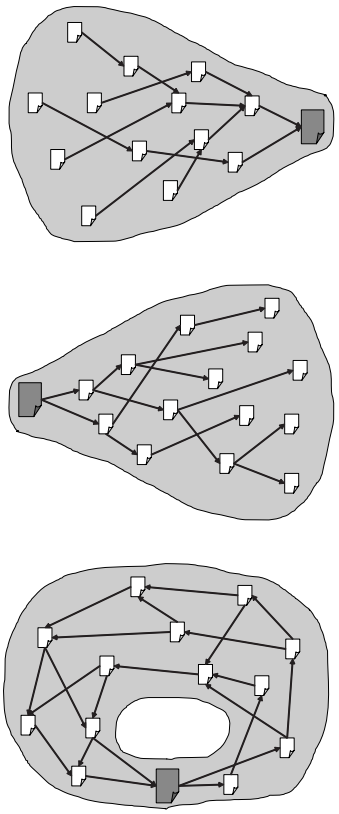

図 3 提案手法による 3 種類のクラスタリングの概念図：fan-in クラスタリング（上），fan-out クラスタリング（中），cycle クラスタリング（下）

Fig. 3 Intuitive illustration of three types of clustering: fan-in clustering (upper figure), fan-out clustering (middle figure) and cycle clustering (lower figure).

るに，出次数の和という尺度で長さが測られるパスをどのように束ねれば，Web 検索の性能向上により大きく寄与するクラスタを作れるだろうか，という問いに答えるかたちで，“出次数の大きいページを数多く経由するほど，内容的に大きく異なる Web ページへ移行しやすくなる” という先ほどの仮説が，さらに 3 つに細分化されるのである．そして，今回の実験によって，3 つの仮説のうちどれが最も強く立証されるかが分かる．なお，τ の増減にともなって，どの種類のクラスタリングにおいても，クラスタ・サイズが大小に変化する．τ は，クラスタの粒度を制御するパラメータとしての役割を果たしている．

3.3.3 計 算 量

いずれの種類のクラスタリングを実行する場合も，与えられた Web ページの総数を n，ハイパーリンクの総数を m として，提案のクラスタリング手法全体での時間計算量の上界は $O(n^2 \log n + mn)$ となる．なぜなら，この値は，すべての Web ページから，他のすべての Web ページへの最短パス長を求めた場合の計算量であり[2),5)]，提案のクラスタリング・アルゴリズムの計算量は，これを決して超えないからである．

しかし，別論文[10),11)] で論じたように，中心ページとして選ばれる Web ページの数（クラスタの個数）は，実際には Web ページの総数 n よりも相当少なく，また，最短パス長計算のための中心ページからの探索範囲は，実際には閾値パラメータ τ によって著しく制限される．そのため，上記の上界よりも実際の計算量はきわめて少なくなっていると思われる．実時間でいえば，今回の評価実験で最も良い検索性能を示した fan-out クラスタリングを，NTCIR-3 Web 検索タスクの文書データ NW100G-01（約 1,000 万件の Web ページ，約 5,500 万のハイパーリンクを含む[3)]）に対して実行するのに，Xeon 2.8 GHz，メモリ 6 GB の計算機 10 台で並列実行し，約 20 時間を費やした．比較のために述べておけば，この時間は，同文書データすべてを同じ計算機 10 台で並列して MeCab で形態素解析し終える時間よりも短い．

4. 評 価 実 験

評価実験では，NTCIR-3 Web 検索タスク[3)] のために準備された文書データ NW100G-01 と 47 個の検索質問を用いた．評価尺度も，NTCIR-3 Web 検索タスクで使われた平均適合率（average precision）を採用し，さらに，やはり同タスクにおけるサーベイ検索タスクの評価方法にならって，検索結果の上位 1,000 件をとって評価を行った．なお，評価における relevance level は rigid であり，これはあわせて NTCIR-3 において提供された relaxed relevance level に比べて，より厳しく適合の度合いを見るための正解データである．実験では，fan-in クラスタリング，fan-out クラスタリングについては，$\tau = 10$，15，20，25，30 の 5 通り，そして cycle クラスタリングについては，$\tau = 20$，25，30，35，40，50 の 6 通りの閾値パラメータの値で，提案のクラスタリング手法を実行した．

図 4 では，上にワン・クリック・ディスタンス文書モデル，下にページ単位文書モデルによる平均適合率の評価を示した．凡例の “IN”，“OUT”，“CYCLE” は，それぞれ，fan-in クラスタリング，fan-out クラスタリング，cycle クラスタリングを意味し，続く数字は閾値パラメータ τ の値を示す．グラフの縦軸は平均適合率，横軸はクラスタ情報の混合率（式 (4) における α）を表す．ワン・クリック・ディスタンス文書モデルの下での評価によれば，閾値パラメータ τ が 20 および 25 のときの fan-out クラスタリングを利用して文書ベクトルを変更した場合に，ベースライン（図中の太い実線），つまりクラスタ情報をいっさい用いない場合に比べて，$\alpha = 0.8$ のとき，10%を超える平均適合率の改善が実現され，平均適合率に関する経験則[13)] によれば，重要な差といえる（ベースラインの 10%増しの値を，太い点線で表してある）．これは，NTCIR-3 当時の水準で，全参加者中，第 3 位の成績である．また，cycle クラスタリングは，fan-out クラスタリングほどの成績は残せなかったものの，幅広い τ の値に対して安定した性能を示す．つまり，τ の値をチューニングする手間と，達成できる検索性能とのトレード・オフを考えて，fan-out クラスタリングと cycle クラスタリングのどちらを採用するかを決めることができる．具体的な数値は**表 1** を参照されたい．その一方，ページ単位文書モデルの下での平均適合率は，α を上げるほど逆に低下している．これは，ページ単位文書モデルの下では不適合であるが，ワン・クリック・ディスタンス文書モデルの下では適合判定される Web ページが，α を増加させるほど，検索結果の中により多く混ざってくるためである．図 4 の 2 つのグラフから，

- ワン・クリック・ディスタンス文書モデルは，ページ単位文書モデルとは明らかに異なる評価モデルであること，
- 評価モデルのこの違いに対応した検索手法の提案が可能であること，
- 本研究の提案する新たなクラスタリング手法が，そのような検索手法であること，

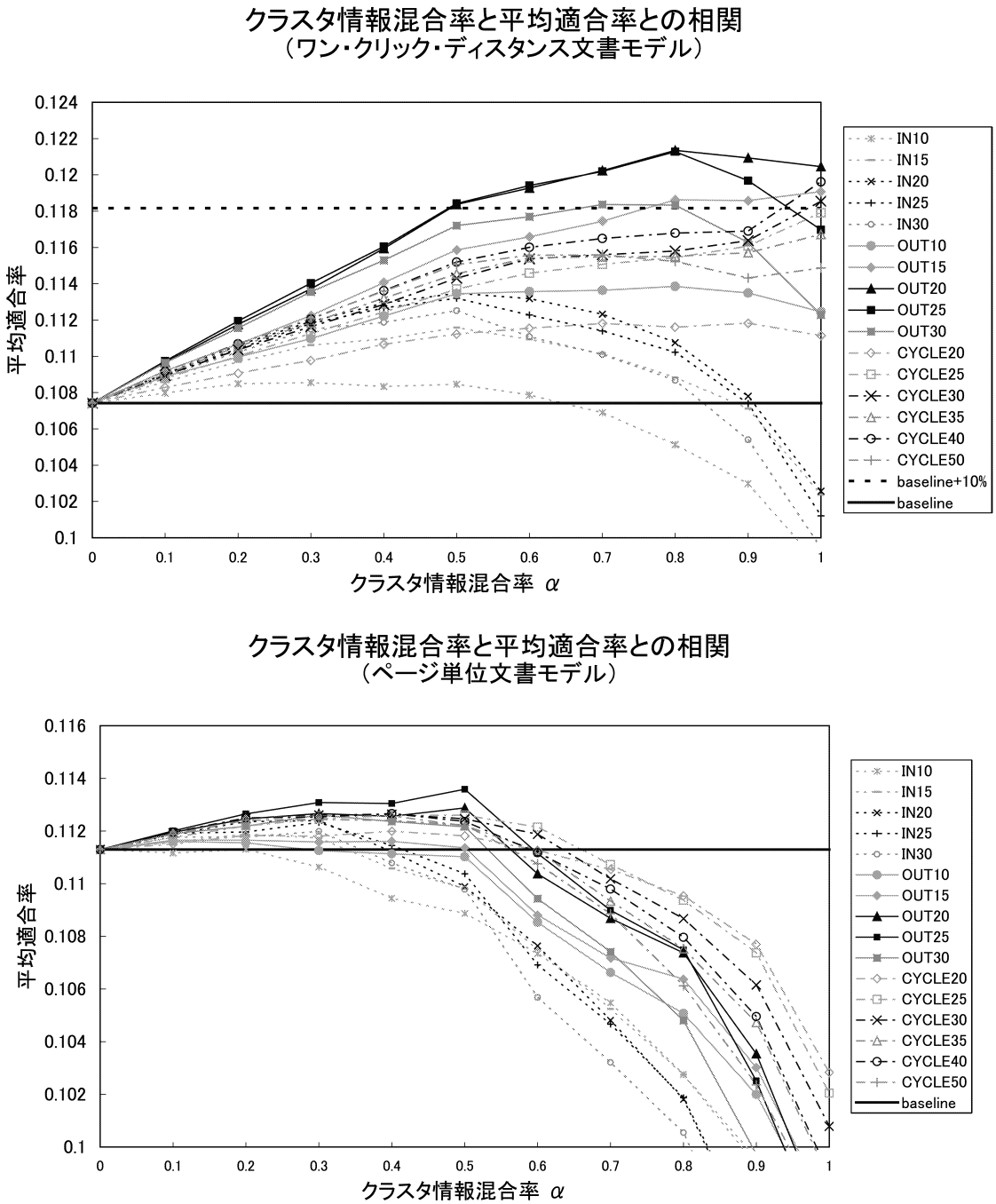

図 4 ワン・クリック・ディスタンス文書モデルによる評価（上）と，ページ単位文書モデルによる評価（下）の比較

Fig. 4 Comparison between the evaluation with one-click-distance document model (the upper graph) and that with page-unit document model (the lower graph).

以上 3 点が分かった．

ところで，NTCIR-3 Web サーベイ検索タスクでは，平均適合率以外の尺度による評価も取り入れられている．**表 2** に，rprec，dcg(100)，dcg(1K) という 3 つの尺度による評価の結果を，クラスタ情報をまったく使わない場合（ベースライン），fan-out クラスタリン

表 1 閾値パラメータが 20，25 であるときの fan-out クラスタリングを使って，文書ベクトルを変更した場合の，クラスタ情報混合率と平均適合率との相関

Table 1 Correlation between cluster information mixture ratio and average precision when we use fan-out clustering results with threshold parameter 20 and 25 to modify document vectors.

α	OUT20	OUT25
0.0 (baseline)	0.1074	0.1074
0.1	0.1097	0.1097
0.2	0.1118	0.1194
0.3	0.1137	0.1140
0.4	0.1159	0.1161
0.5	0.1183 (+11%)	0.1184 (+11%)
0.6	0.1193 (+11%)	0.1194 (+11%)
0.7	0.1202 (+12%)	0.1202 (+12%)
0.8	0.1213 (+13%)	0.1213 (+13%)
0.9	0.1209 (+13%)	0.1197 (+11%)
1.0	0.1205 (+12%)	0.1170

表 2 閾値パラメータが 20 であるときの fan-out クラスタリング，および，閾値パラメータが 40 であるときの cycle クラスタリングを使って，文書ベクトルを変更した場合の，クラスタ情報混合率と平均適合率以外の尺度による評価との相関

Table 2 Correlation between cluster information mixture ratio and evaluation results other than average precision when we use a fan-out clustering result with threshold parameter 20 and a cyclic clustering result with threshold parameter 40 to modify document vectors.

clustering data	α	rprec	dcg(100)	dcg(1K)
(baseline)	0.0	0.1470	6.2910	13.2428
OUT20	0.8	0.1515	6.9676	**14.1969**
OUT20	1.0	0.1492	6.7916	14.0224
CYCLE40	0.8	0.1476	6.5832	13.4364
CYCLE40	1.0	0.1459	6.4277	13.2877

グで $\alpha = 0.8$，1.0 とした場合，cycle クラスタリングで $\alpha = 0.8$，1.0 とした場合のそれぞれについて示した．これら評価尺度の定義は NTCIR3 Web タスクのオーバビュー[3] を参照されたい．特に dcg(1K) は，$\tau = 20$ の fan-out クラスタリングで，$\alpha = 0.8$ のときに NTCIR-3 当時の水準で第 3 位の値である．

参考までに，各クラスタに含まれる Web ページの個数の分布を，主なクラスタリング結果について，**図 5** に示した．上のグラフは，fan-in クラスタリングについて τ を 20，30 とした場合，fan-out クラスタリングについては，$\tau = 20$ の場合のみ，cycle クラスタリングについて τ を 40，50 とした場合の分布を表している．下のグラフは，fan-out クラスタリングについて，τ を 10，20，30 と変化させたときの分布を表している．τ が大きいとクラスタの粒度は粗くなり，τ が小さいとクラスタの粒度は細かくなることが分かる．なお，fan-out クラスタリングの場合，クラスタに含まれる Web ページ数の平均は，$\tau = 10$，20，30 のとき，それぞれ 2.7 個，4.4 個，6.2 個となっている．

なお，今回の実験では，慎重を期するため，検索性能への寄与に関して，次のようなクラスタリングとの比較も行った．それは，Web ページ u が Web ページ v へのリンクを持っているときは，つねに u と v とが同じクラスタに属するクラスタリングである．このクラスタリングを文書ベクトルの変更に用いると，たとえば $\alpha = 1.0$ とした場合は，すべての Web ページのスコアが，必ず，その Web ページからリンクされている Web ページのスコア以上の値となる．このクラスタリングは，一見，ワン・クリック・ディスタンス文書モデルに非常に適したクラスタリングのように見える．しかし，実際に評価すると，α の値をどのように調整しても，ベースラインと比較して 4%弱の性能改善しか得られなかった．このことから，ワン・クリック・ディスタンス文書モデルの下で，性能の良い Web 検索を実現するためには，リンク先の Web ページの高いスコアをリンクもとの Web ページへと伝播させるだけでは不十分であり，ハイパーリンクによる Web ページ間の無数の指示関係から，Web 検索という応用に対して性質の良いものを，選択的に抽出する必要のあることが分かる．

また，次のようなクラスタリングとの比較も行った．今回良い成績を示した fan-out クラスタリングについて，パス長を，出次数の和ではなく，単純にパスを構成するリンクの本数と定義した場合である．この場合，各クラスタは，THP の大きい順に選ばれたページを始点とする通常の意味での幅優先探索によって構成されることになる．そして，閾値パラメータ τ は，この幅優先探索の深さとなる（$\tau = 1$ のときは，前の段落で述べたクラスタリングに合致する）．しかし，クラスタ情報の混合率 α をどのように調整しても，性能の大きな改善は得られなかった．たとえば，$\tau = 3$ のとき，つまり，中心ページからリンク 3 本をたどって到達できる Web ページをメンバとして各々のクラスタを構成していった場合，最善でも，$\alpha = 0.3$ の場合にベースラインの 1.5%増しの平均適合率しか得られなかった．また，$\tau = 2$ のときは，やはり最善でも，$\alpha = 0.3$ のときにベースラインの 0.7%増しの平均適合率しか得られなかった．参考までに示せば，$\tau = 2$，3 のとき，1 つのクラスタに含まれる Web ページの平均個数は，それぞれ 2.8 個，10.9 個となった．これによって，3.3 節で述べたように，クラスタ内でのテキスト情報の均一性をできるだけ保つようにという意図から，出次数の和によってパス長を定義したことは，

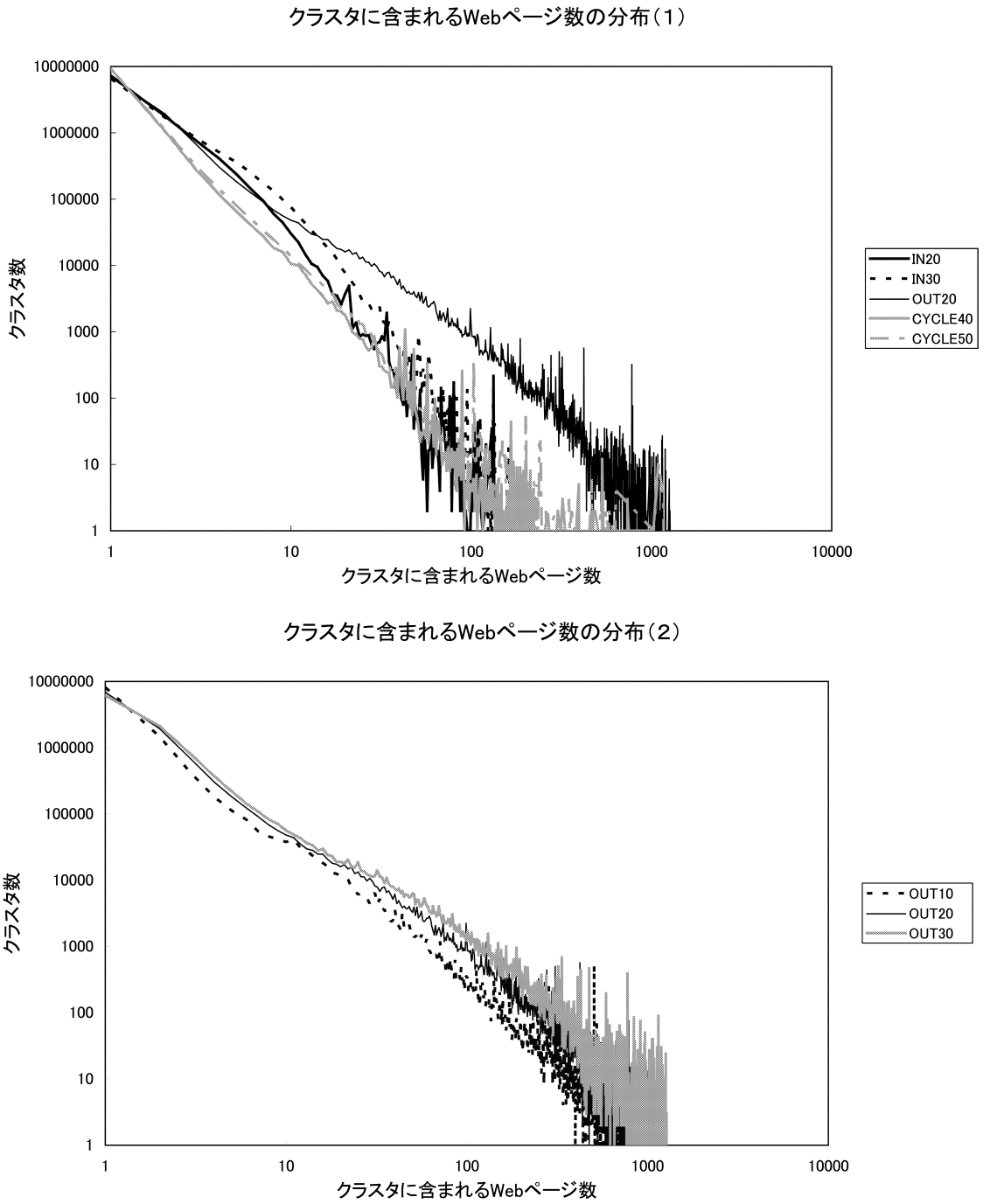

図 5 クラスタに含まれる Web ページ数の分布
Fig. 5 Distributions of the number of Web pages included in each cluster.

有効だったと分かる．

以上の議論をふまえると，本研究の提案したクラスタリング手法の特徴は，次のようになる．第 1 に，出次数によってパス長を定めることで，Web ページのテキスト内容上の逸脱を Web ページ間の距離に換算して，クラスタ内部でのテキスト内容の一貫性を保持する工夫をしている．なぜなら，ワン・クリック・ディスタンス文書モデルの下では，検索質問に適合する Web ページへのリンクを持っていればどんな Web ページでも適合するとされているわけではなく，そこには，

実際に検索結果として提示されて有用かどうかという観点が介入しており，そのうえで正解データの作成が行われているからである．第 2 に，提案手法では，クラスタの中心ページを THP の値の大きい順に選んでいる．THP は，Web のリンク構造が密になっている部分で中心的な位置を占める Web ページにおいて高い値を持ち，よって，検索者が求める Web ページへの案内役の Web ページとしての適切さを代弁する特徴量を採用していることになっている．第 3 に，提案手法はクラスタの粒度を制御するパラメータ τ を備えており，この τ の値を適切に設定することで，各クラスタの “まとまりの良さ” を制御できる．たとえていえば，“目次” としての役割を果たす Web ページに，個別の “章” がぶら下がっているようなリンク構造，また場合によっては，さらにその下に個別の “節” がぶら下がっているような構造を，閾値パラメータの値を調節することで，1 つのまとまりとして括り出すことができ，このような性質の良い部分構造が，リンクによる指示関係全体から選択的に抽出されてはじめて，情報検索の性能向上につながっているものと考えられる．

5. ま と め

検索者の求める Web ページを直接検索結果として提示するのではなく，そのような Web ページへのリンクを含む Web ページを提示することは，WWW に含まれる膨大な情報の一種の “要約作業” として有用と思われる．そのため，従来の，各 Web ページを独立した情報の単位と見なす検索評価モデルだけでなく，ワン・クリック・ディスタンス文書モデルのような評価モデルも提唱される．しかしながら，この新しい評価モデルの下で，大きな性能改善を示す手法は，これまで知られていなかった．この意味において，本研究は，Web 検索研究における新しい領域への第一歩であり，文書モデルそのものの洗練も含めて，より有用な Web 検索の実現へ向けた端緒である．

謝辞 まず，数多くの貴重なコメント・ご批判によって，この論文をより良くすることに多大な貢献をしてくださった査読者の方々に，深く御礼を申し上げます．なお，この研究は，文部科学省科学研究費補助金特定領域研究 13224087（2001-2005）の補助を受けています．また，実験データは NTCIR から提供され，実験環境は国立情報学研究所の大山敬三教授の協力の下に準備されました．

参 考 文 献

1) Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison Wesley Longman (1999).
2) Dijkstra, E.W.: A Note on Two Problems in Connexion with Graphs, *Numer. Math.*, Vol.1, pp.269–271 (1959).
3) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop, *Proc. 3rd NTCIR Workshop* (2003).
4) Eguchi, K., Oyama, K., Aizawa, A. and Ishikawa, H.: Overview of the Informational Retrieval Task at NTCIR-4 WEB, *Working Notes of the 4th NTCIR Workshop Meeting* (2004).
5) Fredman, M.L. and Tarjan, R.E.: Fibonacci Heaps and Their Uses in Improved Network Optimization Algorithms, *J. ACM*, Vol.34, No.3, pp.596–615 (1987).
6) Kanazawa, T., Aizawa, A., Takasu, A. and Adachi, J.: The Effects of the Relevance-Based Superimposition Model in Cross-Language Information Retrieval, *Proc. 5th European Conference on Digital Libraries*, pp.312–324 (2001).
7) Kanazawa, T., Takasu, A. and Adachi, J.: A Relevance-Based Superimposition Model for Effective Information Retrieval, *IEICE Trans. Inf. Syst.*, Vol.E83-D, No.12, pp.2152–2160 (2000).
8) Kanazawa, T., Takasu, A. and Adachi, J.: Improving the Relevance-Based Superimposition Model for IR with Automatic Keyword Extraction, *Proc. Recherche d'Information Assistée par Ordinateur* (*RIAO 2004*), pp.449–462 (2004).
9) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).
10) Masada, T., Takasu, A. and Adachi, J.: Decomposing the Web Graph into Parameterized Connected Components, *IEICE Trans. Information and Systems*, Vol.E87-D, No.2, pp.380–388 (2004).
11) Masada, T., Takasu, A. and Adachi, J.: Web Page Grouping Based on Parameterized Connectivity, *Proc. 9th International Conference on Database Systems for Advanced Applications* (*DASFAA 2004*), pp.374–380 (2004).
12) Ramaswamy, T., Gedik, B. and Liu, L.: Connectivity Based Node Clustering in Decentralized Peer-to-Peer Networks, *Proc. 3rd Interna-*

tional Conference on Peer-to-Peer Computing (*P2P-2003*), pp.66–73 (2003).
13) Voorhees, E.M.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '98*), pp.315–323 (1998).
14) 麻生英樹，津田宏治，村田 昇：パターン認識と学習の統計学—新しい概念と手法，岩波書店 (2003).
15) 風間一洋，原田昌紀，佐藤進也：サーチエンジンの検索結果のマルチレベルグルーピング，*WIT '99* (1999).
16) 風間一洋，原田昌紀，佐藤進也：Web ディレクトリ拡張の自動化手法，情報処理学会論文誌：データベース，Vol.45, No.SIG 7 (TOD22), pp.218–229 (2004).
17) 杉山一成，波多野賢治，吉川正俊，植村俊亮：ハイパリンクで結ばれた隣接ページの内容に基づく Web ページのための TF-IDF 法の改良，電子情報通信学会論文誌，Vol.J87-D-I, No.2, pp.113–125 (2004).
18) 高野 元，久保信也：サイテーション・エンジン：リンク解析を用いた WWW 検索ランキングシステム，情報処理学会研究報告：データベースシステム，Vol.2000, No.010, pp.9–16 (2000).
19) 段 一為，佐野綾一，波多野賢治，田中克己：極小部分マッチグラフを基本とした Web 文章群の検索機構，電子情報通信学会データ工学ワークショップ（DEWS '99）論文集 (1999).
20) 正田備也：リンク情報を利用した Web 文書クラスタリングに関する研究，東京大学大学院情報理工学系研究科電子情報学専攻，博士論文 (Sep. 2004).

(平成 16 年 12 月 20 日受付)
(平成 17 年 4 月 4 日採録)

(担当編集委員 春本 要)

正田 備也（正会員）

1970 年生．1995 年東京大学大学院理学系研究科情報科学専攻修士課程修了．1999 年東京大学大学院総合文化研究科広域科学専攻相関基礎科学系修士課程修了（科学史・科学哲学研究室）．1999 年富士写真光機（株）（現フジノン（株））入社．2004 年東京大学大学院情報理工学系研究科電子情報学専攻博士課程修了．情報検索等の研究に従事．情報理工学博士．日本現象学会会員．

高須 淳宏（正会員）

1984 年東京大学工学部航空学科卒業．1989 年同大学院工学系研究科博士課程修了．工学博士．同年学術情報センター研究開発部助手．1993 年より同センター助教授．2000 年より国立情報学研究所助教授．2003 年より同研究所教授．データ工学，電子図書館，機械学習の研究に従事．電子情報通信学会，人工知能学会，日本データベース学会，ACM，IEEE 各会員．

安達 淳（正会員）

1981 年東京大学大学院工学系研究科博士課程修了．工学博士．東京大学大型計算機センター助手，文部省学術情報センター研究開発部助教授，教授等を経て，現在，国立情報学研究所教授．東京大学大学院情報理工学系研究科教授を併任．データベースシステム，データマイング，情報検索，電子図書館システム等の開発研究に従事．電子情報通信学会，IEEE，ACM 各会員．