グループ化された Web ページを用いた検索

梅沢 晃† 山名 早人‡

†‡早稲田大学大学院理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: †umezawa@yama.info.waseda.ac.jp, ‡yamana@yama.info.waseda.ac.jp

あらまし 検索エンジンは膨大な量の WWW 上から情報を入手するための有効な手段である．しかし Web ページの作者は，1 つのドキュメントを複数の Web ページとそれらを結びつけたハイパーリンクで表現することが多いため，個々の Web ページを単位として検索する検索エンジンでは適切な結果を得ることができるとは限らない．本論文では，複数の Web ページに分散して現れるキーワードに対しての検索を可能にするための，グループ化された Web ページを用いた検索方法を提案する．従来，特徴ベクトルやリンクを用いたグループ化の方法が提案されているが，2，3 個の Web サーバ内の Web ページのみを対象としたグループ化であったり，グループ化された Web ページを実際に検索に適用し，それの有効性を実証するという研究は行われてこなかった．本論文では，1000 万規模の Web ページを対象にリンク情報とパス情報を利用して Web ページのグループ化を行い，グループ化された Web ページを用いた検索方法の提案を行う．実験の結果，提案方法により再現率の向上することが示された．

キーワード Web 検索，リンク解析，Web ページのグループ化

Web Search Method Based on the Grouping of Web Pages

Akira UMEZAWA† Hayato YAMANA‡

†‡Graduate School of Science and Engineering, Waseda University

3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 JAPAN

E-mail: †umezawa@yama.info.waseda.ac.jp, ‡yamana@yama.info.waseda.ac.jp

Abstract A search engine is an effective means to retrieve information from WWW. However, since the author of a Web page puts one document by some Web pages which connected each other in many cases, we cannot necessarily obtain a correct result by the search engine which searches each Web page as a unit. In this paper, we propose the search method using the group of a Web page for enabling search to the keyword which distributes and appears in two or more Web pages. Although the technique of grouping using the feature vector or a link is proposed conventionally, in this paper, by using a lot of Web pages, we use only a link to perform grouping of a Web page and propose the method of using the grouping Web page for search.

Keyword Web Retrieval, Link Analysis, Grouping of Web Pages

1. はじめに

検索エンジンは，膨大な量の WWW 上から必要な情報を入手するための有効な手段である．しかし Web ページの作者は，1 つのドキュメントを複数の Web ページとそれらを結びつけたハイパーリンクで表現することが多いため，個々の Web ページを単位とした検索エンジンでは適切な結果を得ることができるとは限らない．

複数の Web ページに分散して現れるキーワードに対しての検索を可能とする目的や，1 つのドキュメントを構成する Web ページ集合を閲覧しやすいように提示する目的で，Web ページをグループ化する研究が行われている．Web ページのグループ化には，特徴ベクトルを用いた Web ページ同士の類似判定によりグループ化する方法[1]や，Web ページのパス情報，Web ページのメタデータ，異なる Web サーバからのリンク数，他の Web ページへのリンク数などの情報に基づいてグループ化する方法[5]が提案されている．

従来の研究では，2，3 個の Web サーバ内の Web ページのみを対象としたグループ化であったり，グループ化された Web ページを実際に検索に適用し，それの

有効性を実証するという研究は行われてこなかった．

本論文では，1000 万規模の Web ページを対象にリンク情報とパス情報を利用して Web ページのグループ化を行い，グループ化された Web ページを用いた検索方法の提案を行う．そして実験によりグループ化された Web ページを用いた検索の有効性を示す．

本論文の構成は，第 2 節で関連研究を紹介し，第 3 節で本論文で用いる Web ページのグループ化方法を述べる．第 4 節でグループ化された Web ページを用いた検索方法の提案を行う．第 5 節では，実際の Web データを用いて Web ページのグループ化を行い，提案方法の評価を行う．第 6 節で本論文のまとめを行う．

2. 関連研究

本節では複数の Web ページを 1 つの情報単位としてグループ化する研究を紹介する．

[1]では，カットという概念を用いて，ネットニュースの記事や検索エンジンの出力結果の Web ページを個々のトピックに対応している部分グラフへ分割する方法を提案している．カットは，特徴ベクトルを用いて Web ページ同士の類似判定を行うことで，Web ページを意味的に繋がっている部分グラフへと分割する．

LI と Wu は“information unit”[2]という概念を導入した．検索の単位として，多数の物理的に異なる Web ページからなる 1 つのドキュメントである“information unit”が有効であることを示した．

Web ページの作者は，全ての訪問者が同じリンクを通って Web ページ間を移動するという想定の元に Web ページを作成する．そのため，順路の途中の Web ページで表示した情報はそれ以降の Web ページにおいて省略されることがある．しかし，そのような省略は個々の Web ページを単位とする検索エンジンの精度を低下させる原因となる．[3]では，省略された情報を補完するために，Web ページの作者が想定している訪問者のリンク順路を発見する方法を提案している．ある Web ページに対して，その Web ページへの順路として適切なリンク元となる Web ページを，それぞれの Web ページの属する Web サーバ，パス情報などを用いて決定する方法を提案している．

Mizuuchi らは，キーワード発生頻度および分布を見なすことにより“information unit”の概念の拡張を行っている[4]．検索結果として得られる Web ページから，Web ページの内容とリンク情報に基づいて検索結果のランキングを行う方法を提案している．

Ayan らは，Web サーバ内から個人で作成している Web ページや大学のプロジェクトといった論理的なまとまりを持つ Web ページ集合を自動的に発見する方法を提案している[5]．論理的なまとまりを持つ Web ページ集合を発見する方法は，Web ページ集合の入口となる Web ページを発見する段階と，入口ページからその論理的なまとまりを持つ Web ページ集合に含まれる Web ページの境界を識別する段階に分かれる．論理的なまとまりを持つ Web ページ集合の入口ページは，Web ページのパス情報，Web ページのメタデータ，異なる Web サーバからのリンク数，他の Web ページへのリンク数などの情報に基づいてランキングを行うことで決定される．論理的なまとまりを持つ Web ページ集合の境界を識別する方法は，入口ページからのリンク数やパス情報を用いて行われる．

3. Web ページのグループ化

従来の Web ページのグループ化方法では，Web ページの特徴ベクトルから類似度を判定し，グループ化に利用する方法[1]，Web ページのメタデータ，URL の文字列，アンカーテキストを解析することで，Web ページのグループ化を行う方法[5]，Web ページ作者の想定しているリンクの順路を Web ページの集成に利用する方法[3]が提案されている．

本論文で行う Web ページのグループ化の目的は，複数の Web ページによって 1 つのドキュメントを示す論理的なまとまりを持つ Web ページの集合，1 人の作者により作られた Web ページの集合を発見することである．[5]の方法を基にして Web ページのグループ化を行う．

グループ化される Web ページ集合には入口となる Web ページが必ず 1 つ存在すると考えられる．入口となる Web ページの存在が論理的に 1 つのまとまりを持つ Web ページ集合の存在を示すものとなる．1 つのグループにまとめられる Web ページ集合の入口ページは，そのグループに含まれる Web ページの中で最も上位のディレクトリに存在する Web ページの 1 つであると考える．また，一般的に同一のディレクトリには関連した内容を持つ Web ページを配置すると考えられることから，1 つのディレクトリに入口ページは 1 つであると考える．以上から Web ページグループは入口ページを最上位のディレクトリに持ち，グループのメンバーとなる Web ページは入口ページ以下のディレクトリに存在すると考える．

入口ページの発見の際には異なる Web サーバからのリンク数のみを用いる．理由は，アルゴリズム自体の単純化と，異なる作者により作成された Web ページからのリンクという客観的な情報のみを用いることで，Web ページ作者の癖，Web サーバや Web ページ作成ツールといった環境の違いに依存しないでグループ化が可能であると考えられるためである．

3.1. 入口ページの決定

グループ化される Web ページの集合には入口となる Web ページが必ず 1 ページ存在すると考えられる．入口ページの存在が Web ページグループの存在を示すものとなる．入り口ページの発見は，異なる Web サーバ上にある Web ページからのリンク数で決定する．複数の入口ページ候補が同一ディレクトリに存在する場合は，異なるサーバ上にある Web ページからのリンク数が最も多い Web ページを入口ページとする．また，入口という性質から，自らが属するディレクトリ以下に存在する Web ページへのリンクを 1 つ以上持つ必要がある．

3.2. グループの境界の決定

本論文では，[5]で提案されているパス情報とリンク情報を用いた方法とほぼ同一の手順で Web ページグループの境界を決定する．しかし，Web ページグループのメンバーとなるページ数が少ない場合に行われるグループの統合作業は行わない．本論文ではたとえ所属ページ数が少ない場合であっても，他の Web ページからの十分な数のリンクを持つ Web ページはグループとしての価値があると考える．

入口ページから他の Web ページグループを通過することなく到達可能な Web ページを同一グループの Web ページとし，入口ページより上位のディレクトリへのリンクや，既に通過した Web ページへのリンクは無視する．

図 1 のようなリンク情報を持つ Web ページ集合が存在するとする．図の点線はディレクトリの境界を表し，“html”のディレクトリ以下に“press”，“help”が存在している．入口ページ A と同一のグループとなる Web ページは，A と同一のディレクトリに存在し，A からリンクにより到達可能な C, D, E の Web ページ．さらに，A のディレクトリのサブディレクトリに存在し，リンクを辿ることにより到達可能な F，I の Web ページとなる．G は H を辿ることで到達可能であるが，その際に他の Web ページグループを通過するため A のグループのメンバーとはならない．

4. グループ化された Web ページを用いた検索

ユーザが複数のキーワードを用いて検索を行う際，特定の領域を表すキーワードと，詳細な情報をあらわすキーワードを併用して利用することが多いと考えられる．このような場合，特定の領域を表すキーワードを Web ページグループの検索に利用し，検索されたグループに対して詳細な情報を表すキーワードで検索を行うことにより，従来の Web ページ単位の検索より適切な検索が行えると考えられる．本節では，複数の Web ページに分散して現れるキーワードに対しての検索を可能にするための，Web ページグループを用いた検索方法を提案する．

例として，山名研究室から情報検索に関連している論文を検索するために，キーワード(山名研究室,論文,情報検索)を用いて AND 検索を行うとする．しかし，キーワード中の全ての単語が同一 Web ページ上に存在しない場合，Web ページを単位とする検索エンジンでは検索結果を得ることができない．[3]で述べられているように，多くの Web ページ作者は既に訪問者に示した情報をそれ以降のリンク先では省略することがあり，検索エンジン利用者が求めている適切な Web ページがあるにもかかわらず，全てのキーワードを含まないために除外されることがある．

以上の問題を解決するために，Web ページ単位の検索では適切な検索結果を得られないキーワードに対して，本論文では以下の手順で検索を行う．

1. Web ページグループの中から，領域を表すキーワードを用いて適切な Web ページグループを検索する．
2. 検索された Web ページグループに対して，残りのキーワードを用いて検索を行い，その結果を検索結果として得る．

それぞれの手順を詳細に示す．手順 1 のグループの決定では，検索クエリから領域を表す 1 つのキーワードを用いて検索を行い，得られたそれぞれの Web ページが属するグループを発見する．そして手順 2 で，得られたグループ内の Web ページに対してのみ，残りのキーワードを用いて検索を行い，それによって得られた Web ページを検索結果として得る．

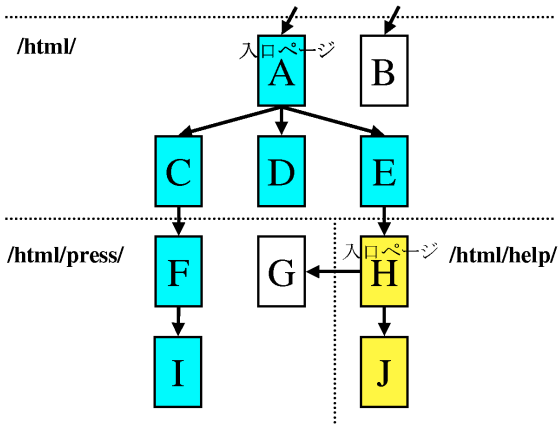

図 1：Web ページ集合のグループ分割

5. 提案方法の評価実験

本節では第 3 節，第 4 節で示した方法を用いて，Web ページのグループ化と，そのグループを用いた検索方法の評価を行う．実験に用いる Web データには，NTCIR-4[6]の Web タスクを用いる．NTCIR の Web データから，データ中に存在する Web ページ間のリンクのみを抽出する．表 1 に Web データの URL 数とリンク数を示す．

表 1：実験用の URL 数とリンク数

URL 数	11,017,992
リンク数	53,964,444

5.1. Web ページのグループ化

パス情報とリンク情報による Web ページのグループ化では，入口となる Web ページからのリンクを辿ることにより，グループのメンバーとなる Web ページが決定される．表 2 は異なる Web サーバからのリンク数を条件とした入口ページを用いたときに抽出される Web ページグループの数，グループに属する Web ページ数，何れかのグループに属する Web ページの Web データ全体に対する割合を示している．[5]では入口ページ数，Web ページグループの最小サイズ，入口ページから辿るリンク数のパラメータを変化させて Web ページのグループ化を行っており，その際 1 グループあたりの平均 Web ページ数は，16.1～233.8 になる．この値は今回リンクのみを用いて行ったグループ化による 1 グループの平均 Web ページ数とほぼ同等であり，グループの大きさとしては適切であると考えられる．

グループへの分割例として，表 3 に異なる Web サーバからのリンク数が 1 以上ある Web ページを入口ページとしたときの，“www.waseda.ac.jp” 内のグループを示す．Web サーバ内には 122 の Web ページが存在し，11 のグループが発見された．太字で示された URL がそれぞれのグループの入り口ページとなる．

異なる Web サーバからのリンクの閾値が 1 以上のときの Web ページグループの大きさの分布(図 2)を見てみると，構成する Web ページ数が少ないグループに偏っている．2 個の Web ページで構成されるグループがグループ全体の 8%あり，半分のグループが 17 個以下の Web ページで構成されていることがわかる．このような小さなグループに対して[5]では，パス情報とリンク構造を用いて Web ページグループを統合していくが，本論文では既に述べたように Web ページグループの統合は行わない．

また Web サーバ内に含まれるグループ数の分布(図 3)から，Web ページグループを持つ Web サーバのうち約 72%において，Web サーバ内に 1 つのグループしかもっていないことがわかる．一方で 100 を超えるグループに分割された Web サーバもいくつか存在する(表 4)．

表 2：Web ページグループ

異なる Web サーバからリンク数	グループ数	1 グループの平均 Web ページ数	Web ページ全体に占める割合
1 以上	170,264	53.71	83.01%
2 以上	116,960	75.16	79.79%
3 以上	90,264	93.09	76.26%
4 以上	75,552	106.27	72.87%
5 以上	64,651	118.24	69.38%
6 以上	56,825	128.40	66.22%
7 以上	50,121	138.69	63.09%
8 以上	45,089	148.11	60.61%
9 以上	40,941	156.73	58.24%
10 以上	37,437	165.76	56.32%

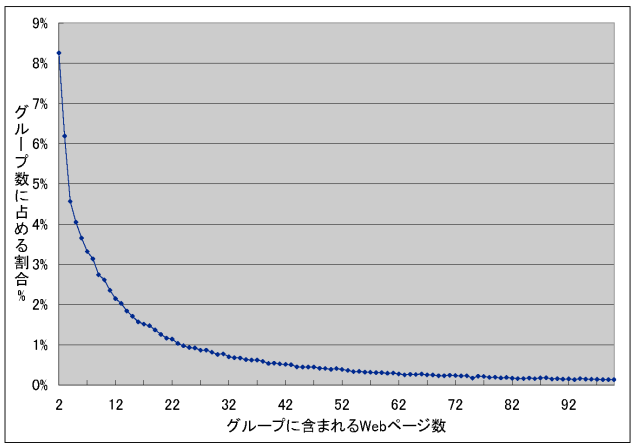

図 2：1 グループあたりの Web ページ数の分布

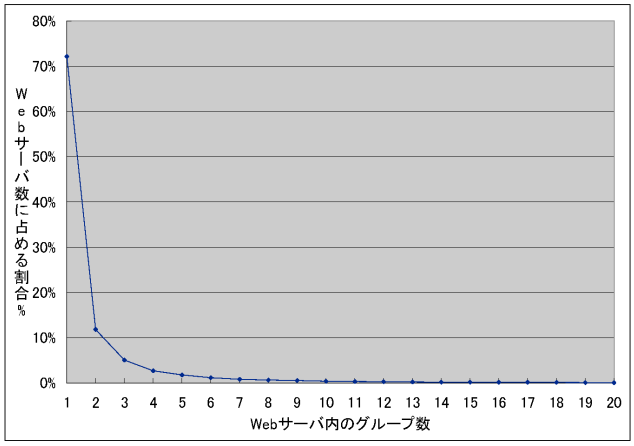

図 3：Web サーバに含むグループ数の分布表

表 3：www.waseda.ac.jp 内のグループ

http://www.waseda.ac.jp/alumni/index.html
http://www.waseda.ac.jp/alumni/index-e.html
http://www.waseda.ac.jp/cie/international.html
http://www.waseda.ac.jp/cie/icl/index.htm
http://www.waseda.ac.jp/cie/index.html
http://www.waseda.ac.jp/ecocampus/index-e.html
http://www.waseda.ac.jp/ecocampus/index.html
http://www.waseda.ac.jp/gradcom/index.html
http://www.waseda.ac.jp/gradcom/access/index.html
http://www.waseda.ac.jp/gradcom/calendar/index.html
http://www.waseda.ac.jp/gradcom/degree/index.html
http://www.waseda.ac.jp/gradcom/entrance/index.html
http://www.waseda.ac.jp/gradcom/exchange/index.html
http://www.waseda.ac.jp/gradcom/facilities/index.html
http://www.waseda.ac.jp/gradcom/history/index.html
http://www.waseda.ac.jp/gradcom/lectures/index.html
http://www.waseda.ac.jp/gradcom/members/index.html
http://www.waseda.ac.jp/gradcom/news/index.html
http://www.waseda.ac.jp/gradcom/open/index.html
http://www.waseda.ac.jp/gradcom/publications/index.html
http://www.waseda.ac.jp/gradcom/seminars/index.html
http://www.waseda.ac.jp/gradcom/services/index.html
http://www.waseda.ac.jp/index-j.html
http://www.waseda.ac.jp/
http://www.waseda.ac.jp/WIAPS/index.html
http://www.waseda.ac.jp/archives/index.html
http://www.waseda.ac.jp/bonn/index.html
http://www.waseda.ac.jp/gradlaw/
http://www.waseda.ac.jp/gsaps/index.html
http://www.waseda.ac.jp/hiken/index.html
http://www.waseda.ac.jp/index-b5.html
http://www.waseda.ac.jp/index-gb.html
http://www.waseda.ac.jp/index-k.html
http://www.waseda.ac.jp/intl-ac/index.html
http://www.waseda.ac.jp/kikou/
http://www.waseda.ac.jp/mailto.html
http://www.waseda.ac.jp/mnc/index.html
http://www.waseda.ac.jp/organization.html
http://www.waseda.ac.jp/placement/
http://www.waseda.ac.jp/projects/LAW/index.html
http://www.waseda.ac.jp/schl/edu/index.html
http://www.waseda.ac.jp/seikei/index.html
http://www.waseda.ac.jp/student/award/
http://www.waseda.ac.jp/student/expo/
http://www.waseda.ac.jp/student/index.html
http://www.waseda.ac.jp/student/shp/index.html
http://www.waseda.ac.jp/syogakukin/
http://www.waseda.ac.jp/whatsnew.html
http://www.waseda.ac.jp/intl-ac/bulletin/c12-05.html
http://www.waseda.ac.jp/intl-ac/bulletin/c01.html
http://www.waseda.ac.jp/intl-ac/bulletin/c02.html
http://www.waseda.ac.jp/intl-ac/bulletin/c03.html
http://www.waseda.ac.jp/intl-ac/bulletin/c04.html
http://www.waseda.ac.jp/intl-ac/bulletin/c05.html
http://www.waseda.ac.jp/intl-ac/bulletin/c06.html
http://www.waseda.ac.jp/intl-ac/bulletin/index.html
http://www.waseda.ac.jp/koho/index.html
http://www.waseda.ac.jp/koho/databook/
http://www.waseda.ac.jp/koho/festa/index.html
http://www.waseda.ac.jp/koho/ichizu/index.html
http://www.waseda.ac.jp/koho/media/index.html
http://www.waseda.ac.jp/koho/ob/index.html
http://www.waseda.ac.jp/open/index.html
http://www.waseda.ac.jp/open/internships/keijiban/index.html
http://www.waseda.ac.jp/open/network/index.html
http://www.waseda.ac.jp/open/new/index.html
http://www.waseda.ac.jp/open/openkamoku/keijiban/index.html
http://www.waseda.ac.jp/rps/faculty/index-e.html
http://www.waseda.ac.jp/rps/faculty/index.html
http://www.waseda.ac.jp/schl/edu/staff/index.html
http://www.waseda.ac.jp/schl/edu/staff/associ/index.html
http://www.waseda.ac.jp/schl/edu/staff/office/index.html
http://www.waseda.ac.jp/schl/edu/staff/office/ogata/index.html
http://www.waseda.ac.jp/waseda125/index.html
http://www.waseda.ac.jp/waseda125/etc/index.html

表 4 ：100 以上のグループを持つ Web サーバ

グループ数	Web サーバの URL	サーバの内容
222	www.ryokan.or.jp	社団法人国際観光旅館連盟
167	www.shop.wan.ne.jp	犬のためのショッピングモール
132	www.e-na.co.jp	ショッピングモール
127	www.pcassist.co.jp	パソコンスクール
124	www.edu.net-kochi.gr.jp	高知県教育情報通信ネットワークシステム
117	www.hotwired.co.jp	雑誌
115	www.tbs.co.jp	テレビ局
109	www.edu.city.kyoto.jp	京都市教育委員会
108	www.tv-asahi.co.jp	テレビ局
104	www.ans.co.jp	保育園・幼稚園を対象とした Web サーバ運営
103	shinshu.online.co.jp	信州観光案内
102	www.i-love-epson.co.jp	エプソンの製品情報
100	www.eee.co.jp	ショッピングモール
100	www.rtpro.yamaha.co.jp	ヤマハ製ルータの情報
100	www1.kyoto-be.ne.jp	京都府教育委員会

5.2. グループ化された Web ページを用いた検索

NTCIR-4 の Web データを対象とした検索を行う．検索には日本語全文検索システムである Namazu を利用する．検索を行うキーワードは NTCIR-4 の Web タスクで用いられた検索課題を利用する．検索課題には，従来の学術文書や新聞記事を対象とした Ad Hoc 型検索(ad hoc search)あるいは主題検索(subject search)に相当し，固定された文書データセットに対して，所与の話題に関する文書集合を発見する情報指向検索タスクと，既知事項検索におけるシステム評価を対象としたナビゲーション指向検索タスクがある．それらの検索課題から，複数のキーワードにより構成される検索クエリを実験では利用する．2 語のキーワードで構成される検索クエリは 100 個，3 語のキーワードで構成される検索クエリは 60 個，4 語のキーワードで構成される検索クエリは 2 個存在した．本論文で提案する検索方法の目的は，少ない検索結果しか得られない検索クエリに対して，検索に使うキーワードの Web ページ中での出現条件を緩和することで，単純な AND 検索で

は発見できなかった Web ページを発見することである．単純な AND 検索による検索結果数が 0～9 である 21 種類の検索クエリ(付録)を対象として本論文で提案する検索方法を適用する．検索課題にある検索キーワードは検索課題作成者が重要であると考えた順番に並んでいる。Web ページのグループ化の際に利用する，異なる Web サーバからのリンク数の閾値は 1 とする．

表 5 は単純な AND 検索による結果と，検索課題作成者が最も重要と考える検索ワードをグループ検索に利用したときの提案方法による検索結果を示す．結果数は検索された Web ページの数，正解数は適切な Web ページの数，精度は検索の精度を示す．表 5 よりほとんど全ての検索結果数が単純な AND 検索より増加していることが分かる．

次に，検索精度を見てみる．AND 検索の結果と提案方法の結果の全ての Web ページを手作業で確認した．AND 検索にしろ提案方法による検索にしろ，適切な文書を見つけられることが少ないことがわかる．これは検索課題が不適切であるというより，収集された Web データが不完全であっただめだと考えられる．NUM13 以外の検索課題では提案方法による検索精度の向上は見られない．

しかし，AND 検索と提案方法による検索結果の正解 Web ページ数の比較を示す図 4 から，NUM5, NUM13, NUM15, NUM18 の検索課題において，検索要求を満たす Web ページ数が増加していることが分かる．また，NUM13 では特定の和菓子店のトップページを検索要求として指定しているため，要求を満たす Web ページが 1 つしかないが，AND 検索では発見できなかったその Web ページを提案方法では発見することができた．

6. おわりに

本論文では，複数の Web ページに分散して現れるキーワードに対しての検索を可能とするため，1000 万規模の Web ページを対象にリンク情報とパス情報を利用して Web ページのグループ化を行い，グループ化された Web ページを用いた検索方法の提案を行った．実験の結果，検索精度は従来の検索方法に比較して下がったが，単純な AND 検索では発見できない適切な Web ページを発見することができた．

今後の課題としては，Web ページのグループ化方法の洗練，検索クエリからグループ化に決定するキーワードを自動的に抽出する方法の開発などが考えられる．

謝　辞

本研究の一部は，科学技術振興費「e-Society」，及び，文科省２１世紀 COE「プロダクティブ ICT アカデミア」によるものである．

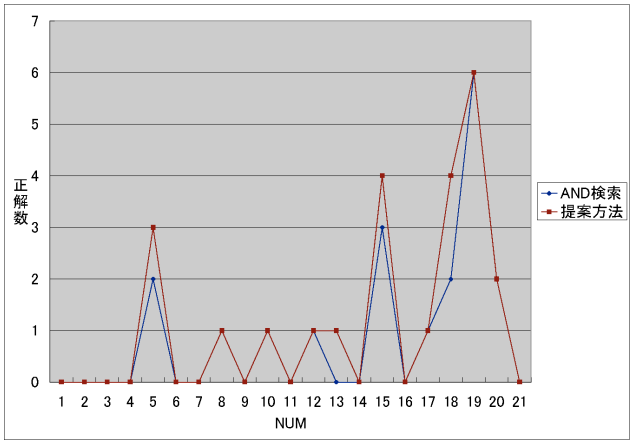

図 4：AND 検索と提案方法の検索の正解数

表 5：AND 検索と提案方法の検索結果

NUM	AND 検索			提案方法		
	結果数	正解数	精度	結果数	正解数	精度
1	0	0	0.0%	0	0	0.0%
2	0	0	0.0%	0	0	0.0%
3	1	0	0.0%	47	0	0.0%
4	2	0	0.0%	67	0	0.0%
5	2	2	100.0%	8	3	37.5%
6	2	0	0.0%	31	0	0.0%
7	2	0	0.0%	12	0	0.0%
8	3	1	33.3%	3	1	33.3%
9	3	0	0.0%	75	0	0.0%
10	3	1	33.3%	51	1	2.0%
11	3	0	0.0%	7	0	0.0%
12	4	1	25.0%	11	1	9.1%
13	4	0	0.0%	136	1	0.7%
14	5	0	0.0%	9	0	0.0%
15	6	3	50.0%	14	4	28.6%
16	6	0	0.0%	624	0	0.0%
17	6	1	16.7%	12	1	8.3%
18	7	2	28.6%	60	4	6.7%
19	7	6	85.7%	8	6	75.0%
20	8	2	25.0%	86	2	2.3%
21	9	0	0.0%	19	0	0.0%

文　献

[1] Keishi Tajima, Yoshiaki Mizuuchi, Masatsugu Kitagawa, and Katsumi Tanaka: "Cut as a Querying Unit for WWW, Netnews, and E0mail", Proceeding of 9th ACM Conference on Hypertext and Hypermedia, pp.235-244, (1998.6)

[2] Wen-Syan Li, Yi-Len Wu: "Query relaxation by structure for document retrieval on the Web", Proceedings of 1998 Advanced Database Symposium, (1999.11)

[3] Yoshiaki Mizuuchi, and Keishi Tajima: "Finding Context Paths for Web Pages", Proceeding of 10th ACM Conference on Hypertext and Hypermedia, pp.13-22, (1999.2)

[4] Keishi Tajima, Kenji Hatano, Takeshi Matsukura, Ryoichi Sano, Katsumi Tanaka: "Discovery and retrieval of logical information units in Web", Proceedings of the 1999 ACM Digital Libraries Workshop on Organizing Web Space, (1999.8)

[5] Necip Fazil Ayan, Wen-Syan Li, Okan Kolak:

"Automating Extraction of Logical Domains in a Web Site", In International Journal of Data and Knowledge Engineering, 43(2), Elsevier Science, pp.179-205, (2002.11)

[6] -:"NTCIR Workshop", http://research.nii.ac.jp/ntcir-ws4/index-ja.html

付　録

NUM	キーワード	検索の目的
1	ビデオチャット Web カメラ PC カメラ	PC カメラを使った流行のビデオチャットについて知りたい.
2	六本木小田島 ワイン講習会	六本木小田島のワイン講習会について知りたい.
3	アイメイク やり方	アイメイクのやり方を詳細に知りたい.
4	チェックディジット アルゴリズム	チェックディジットを計算するアルゴリズムにどのようなものがあるかを知りたい.
5	ドラえもん スネツグ	ドラえもんに出てくるスネ夫の弟のスネツグについて調べたい.
6	海外化粧品 ネット通販	海外化粧品をネット上で販売しているサイトを探したい.
7	ナショナル掃除機 紙パック	ナショナル掃除機の紙パックを調べたい.
8	クレーンゲーム UFO キャッチャー コツ	クレーンゲームのコツを知りたい.
9	非木材紙 材質 製法	木材以外でつくられている紙はどのような材質でどのように生産されるのかを知りたい.
10	イコールアクセス アメリカ 米国	アメリカのイコールアクセス制度に関する知識を知りたい.
11	相対性理論 ローレンツ短縮	相対性理論におけるローレンツ短縮の原理を知りたい
12	護身術 痴漢 強盗	日常の生活の中で痴漢や強盗などに襲われたときの護身術を知りたい
13	神保町 最中 和菓子店	神保町にある最中の有名な和菓子店の連絡先を知りたい.
14	電子情報通信学会 総合大会 期日	電子情報通信学会総合大会の期日を調べたい.
15	とんかつ まい泉 メニュー	とんかつまい泉のメニューを知りたい.
16	讃岐うどん 定義	『讃岐うどん』というものがあるが普通のうどんとはどう異なるものなのか知りたい.
17	作家の値打ち 福田和也	『作家の値打ち』という本を出し, 話題を作った福田和也という人物について知りたい.
18	宇宙開発の歴史 ロケット ミサイル	宇宙開発は主に冷戦下で行われたが, その文化史的歴史（月面着陸や火星探査計画など）および技術的歴史（ミサイル技術とロケット技術の関わりなど）に関して知りたい.
19	初級シスアド 初級システムアドミニストレータ 平均年齢	シスアドの受験者, 合格者等に関する平均年齢を調べたい.
20	東京地方裁判所 強制競売	東京地方裁判所で行われる強制競売不動産情報を知りたい.
21	就職活動 記録 合否結果	個人の就職活動に関する合否結果などの記録について記述されている文書を探したい.