Max Flow アルゴリズムを用いたWeb ページのクラスタリング方法とその評価

大野成義[†,††] 渡辺匡[††] 片山薫[††]
石川博[††] 太田学[†††]

Web 上の情報を探すために使われる検索エンジンの多くはユーザに検索結果をスコア順のリストとして返す．したがって，リストが長い場合，求める情報を探すのはきわめて難しい．そこで，検索結果をリストでなくカテゴリ表示するための新しいクラスタリング方法を提案する．クラスタリングする方法としては，ページ内の文章を解析する方法でなく Web ページの持つリンク情報を基に行う．リンク情報の解析には，より緻密に結びついたリンク構造にあるページ集合を見つけるのに有効な最大流アルゴリズムを用いる．提案方法を定量的に評価するために，適合の正解がある NTCIR のデータを使い実験を行い良好な結果を得た．

Clustering Web Pages Based on Maximum Flow Algorithm

Shigeyoshi Ohno,[†,††] Masashi Watanabe,[††] Kaoru Katayama,[††]
Hiroshi Ishikawa[††] and Manabu Ohta[†††]

While search engines are indispensable for searching on the Web, users have to check a long ordered list to locate necessary information. It is often tedious and less efficient. In this paper, we propose a new link-based clustering approach to categorizing search results returned from Web search engine. The maximum flow algorithm which is effective to find the page sets connected tightly by hyperlinks is used for the analysis of link information. In order to evaluate method perfoemance quantitatively, we conducted experiments using the data of NTCIR and had good results.

1. まえがき

インターネット上の Web には膨大な情報が存在している．情報量はとりわけ 2000 年以降，爆発的に増加しつつある．この情報空間から必要な情報を探しだすために検索エンジンが広く利用されている．ところが，検索エンジンを使ったからといって簡単に必要な情報が得られるとは限らない．多くの検索エンジンはユーザの入力した検索語に基づき検索を行い，その結果をスコア順に並べたリストとして返す．その検索結果のすべてがユーザの必要とした情報とは限らないため，ユーザは検索結果のリストを順番に調べる必要がある．短いリストであれば簡単に調べられるが，短いリストは再現率が悪くなりやすい．長い検索結果リストであれば，必要とされる情報が多く含まれている可能性は高くなるが，順番に調べるには時間がかかる．

もし，検索結果をスコア順のリストでなく，類似のページをクラスタにまとめれば，必要な情報を見つけやすくなることが期待される．

そこで，検索結果をクラスタリングする新しい方法を提案する．文書をクラスタリングする場合，共通する語や句から類似度を計算して行う方法がある．しかし，検索エンジンで扱う Web ページは通常の書籍とは異なる．Web ページによっては動画や画像を張りつけて文字情報が少ない場合もある．また，Web ページの間にはリンクが張られており，各 Web ページにはほかのページへのリンクが埋めこまれている．このリンク情報を用いてクラスタリングする方法を提案する．

類似の内容の Web ページ間はリンク構造も密になっていると期待できる．つまり，あるページのリンクは

† 職業能力開発総合大学校情報工学科
Department of Information and Computer Science, Polytechnic University
†† 東京都立大学大学院工学研究科
Graduate School of Engineering, Tokyo Metropolitan University
††† 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

元のページと何らかの関係のあるページにリンクが張られている．当然，内容が類似したページにはリンクされやすく，まったく関連のないページにリンクされる可能性は低い．このようなリンク構造は最大流アルゴリズムを使うことで調べることができる．そこで，Web ページのクラスタリング方法に，この最大流アルゴリズムを用いることを提案する．

なお，提案方法を定量的に評価するために適合する正解が存在する NTCIR（NII-NACSIS Test Collection for IR Systems）のデータを利用して実験を行った．

2. 関 連 研 究

2.1 検索結果のクラスタリング

Web 検索結果のクラスタリングは大きく 2 つに分類できる．コンテンツ・マイニングを利用したクラスタリングと，ストラクチャ・マイニングを利用したものである．前者はコンテンツを分析し，特徴語を抽出し，その特徴語に着目して似通ったページを 1 つのクラスタにまとめる[1),2)]．したがって，コンテンツに書かれている言語に依存する．また，特徴語を抽出していることから，コンテンツの内容，意味を分析していることになる．コンテンツ・マイニングによるクラスタリングは研究が進んでおり，商用の Vivisimo[3)] や WSM[4)] のようなシステムが存在する．後者はコンテンツ・マイニングだけでなくログの分析も利用してより精度を上げようとしているようである．しかし，コンテンツ・マイニングによるクラスタリングは，形態素解析などを使って特徴語を抽出する時間が必要になることや文書集合にベクトルを割り当てることによる次元の呪縛の困難など問題を発生する場合がある．さらに，動画や画像データなどで構成されていてテキストの少ないページには適用が難しい．

一方，Web ページはリンク情報を持っていることから，このリンク情報を分析することで有用な情報を取りだすことをストラクチャ・マイニング[5),6)] と呼んでいる．クラスタリングでもこのリンク情報を使う方法が考えられる．このような方法によるクラスタリングはあまり研究が進んでおらず，商用システムも知られていない．また，ストラクチャ・マイニングであれば，テキストの少ない動画や画像・音声データで構成されているページにも適用できる．そこで，本研究ではこのストラクチャ・マイニングによるクラスタリングを目指す．

リンク情報を用いたクラスタリングとして，たとえば，Wang ら[7)] の研究がある．コンテンツ・マイニングによるクラスタリングでは語や句に着目し，これらを共有するページを同一のクラスタに分類することから，語や句の代わりにリンクに着目し，同じリンクを持つページを同一のクラスタに分類する方法を提案している．

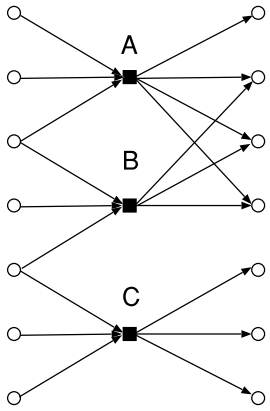


図 1 リンク先やリンク元を共有するページ
Fig. 1 Pages with shared links.

具体的には，2 つのページ P,Q の類似度を以下のように定義し，K-平均法を使ってクラスタリングを行う．

$$Cosine(P,Q) = (P \cdot Q)/(\| P \| \| Q \|) \quad (1)$$
$$= \frac{((P_{out} \cdot Q_{out}) + (P_{in} \cdot Q_{in}))}{(\| P \| \| Q \|)}$$
$$\| P \|^2 = (\Sigma_1^N P_{out}^2 + \Sigma_1^M P_{in}^2) \quad (2)$$
$$\| Q \|^2 = (\Sigma_1^N Q_{out}^2 + \Sigma_1^M Q_{in}^2) \quad (3)$$

P_{out} はページ P から出てゆく N 個のリンクを表し，出力リンクを各次元に割り当てた N 次元ベクトルである．P_{in} はページ P に張られているリンクを表し，入力リンク数 M に対応する次元を持ったベクトルである．

ページ A，B，C とそのリンク構造が**図 1** のようになっていたとする．ページ A，B の類似度より，ページ B，C の類似度の方が小さい．また，ページ A，C には共通するリンク元やリンク先がないため，その類似度は 0 となる．この方法はクラスタの数を調節することができるという特徴がある．しかし，**図 2** のページ P と Q のように相互にリンクを張っていても共通するリンクがないと，式 (1) で計算される類似度は 0 になってしまう．ページ R と Q には共通するリンク先やリンク元があるため同じクラスタに分類されることがあっても，リンクによる参照関係が考慮されないと P と Q が同じクラスタに分類されることはない．リンクが張られているのは何らかの関連性があるからであり，この参照関係をクラスタリングするときに利用

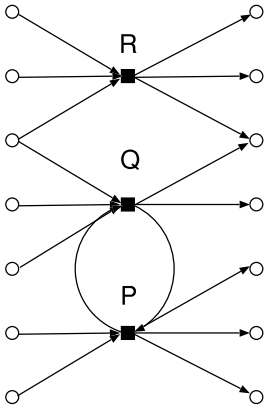


図 2 リンク先（元）を共有するページと共有するリンク先（元）はないがリンクによる参照関係のあるページ
Fig. 2 Pages with shared links and pages with an unshared link and reference links.

すべきである．

2.2 クローリングによって集められる巨大な Web ページ集合を対象としたクラスタリング

ストラクチャ・マイニングの例としては，ほかにクラスタリングの対象を検索結果に制限せず，クローリングによって集められる巨大な Web ページ集合に広げた研究がある．この世界中に存在する Web ページのリンク構造を分析することで，Web ページをクラスタリングする手法として，正田ら[8),9)] の連結性に基づく方法があり，コミュニティを発見する手法として Imafuji ら[10)] の最大流アルゴリズムを使う方法があげられる．

ページ A からリンクをたどってページ B に到達可能で，逆に B から A へも到達可能だとすると，この 2 つのページは密接な関係があると考えられる．つまり，リンク構造を有向グラフと見たときの強連結成分は，密接な関係にあるということである．さらに，正田らは到達可能なページ間の離れ具合や間に存在するページの持つリンク数（ハブ値やオーソリティ値）からページ間の距離を導入し，強連結成分の細分化を行っている．このページ間の距離をページの類似度とし，クラスタリングすることができる．この場合，同一のクラスタにする距離を調節することで，クラスタの大きさを制御できる．

一方，Flake ら[11)] はコミュニティを**図 3** で示すような「コミュニティの外へのページへの（または，からの）リンクよりもコミュニティ内のページどうしのリンクを多く持つ」という条件を満たす Web ページの集合と定義し，最大流アルゴリズムを使うことで，近似的に計算できることを発見した．ページを頂点，

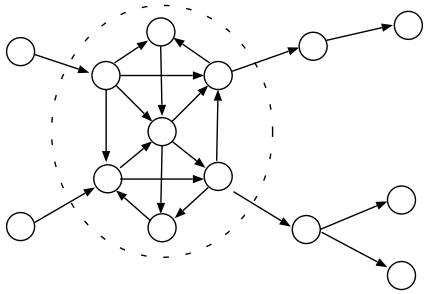


図 3 Web コミュニティの例
Fig. 3 An example of web community.

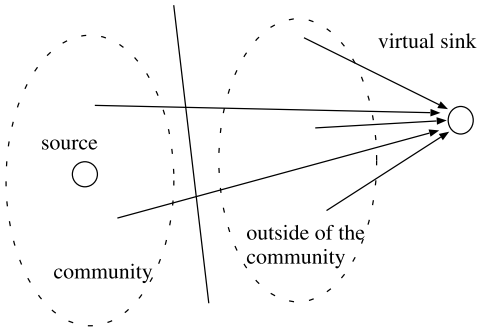


図 4 最大流と最小切断の関係
Fig. 4 Max Flow and Min cut.

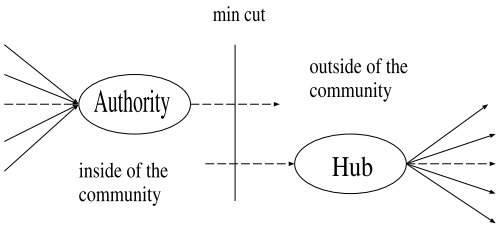


図 5 オーソリティページ・ハブページと最小切断
Fig. 5 Authority / Hub pages and Min cut.

リンクを辺容量 1 の有向辺とする有向グラフにおいて，あるページをソースとしそのページから十分離れたページをシンクとする．最大流アルゴリズムによってソースからシンクに流れる最大流量と各有向辺に流れる流量を求めることができる．このとき最大流量は最小切断容量となり，最小切断はソースページを含むコミュニティとそのコミュニティの外との境界になる（**図 4**）．また，**図 5** のようにオーソリティページの出力リンクは最小切断になりやすく，オーソリティページはコミュニティに含まれやすい．一方で，ハブページは出力リンク数より入力リンク数が少なく，入力リンクが最小切断になりやすいことからコミュニティに含まれ難くなる．さらに，今藤らは最大流アルゴリズムを使ったコミュニティの特徴分析[12)] を行っており，

コミュニティのトピックがより詳細になると報告している．つまり，最大流アルゴリズムを使うことで，より細かな点まで似たページを集めてくることができると期待できる．

そこで最大流アルゴリズムを使う方法を検索結果のクラスタリングに利用することを提案する．

3. 最大流アルゴリズムを用いたクラスタリング

3.1 クラスタリングの方法

検索結果が得られたとする．それらのページ間にはリンクが張られている．直接リンクされていることもあれば，1 つ以上のページを経由して間接的にリンクされていることもある．そのような，直接・間接のリンクが密に存在することで複数のページが連結されていれば，これらのページは共通の内容を持つ可能性が高い．これらの関係を最大流アルゴリズムを使うことで見つけだす．

提案方法は以下の手順で検索結果をクラスタリングする．

1. 検索結果リストの n 個のページを，$p_i(1 < i < n)$ とし，その集合を P で表す．

2. p_i からリンクが張られているページを調べ，そのページの集合を $FP1$（FP=Forward Pages）する．さらに，$FP1$ の各要素ページからリンクの張られているページを調べ，そのページの集合を $FP2$ とする．

3. ページ $P \cup FP1 \cup FP2$ を頂点とし，その間のリンクを有向辺とする有向グラフ $G(V, E)$ を考える（**図 6**）．

4. $k = 1$ とする．

5. 3. で得られた有向グラフの範囲で，頂点 p_k から有向辺をつたってたどれる頂点 p_i の集合をシード集合 Pk とする．

6. 周辺グラフ $G'(V', E')$ を以下の手順で構築する（**図 7**）．V のうち Pk から有向辺をつたってたどれる頂点および Pk を V' とし，その間のリンク（辺容量=1）を E' とする．さらに，仮想ソース s を V' に，辺容量が無限大の有向辺 $(s, pk_i), pk_i \in Pk$ を E' に加える．また，仮想シンク t を V' に加え，$V' \cap FP2$ の各頂点から t への有向辺（辺容量=1）を E' に加える．図 7 では $Pk = \{p_1, p_2, p_3, p_4\}$，$V' \cap FP1$ を白丸印で，$V' \cap FP2$ を白三角印で表す．

7. 最大流アルゴリズムを実行する．**図 8** で飽和辺を点線で表す．

8. p_k から不飽和辺をたどって到達可能な頂点 p_i と p_k からなる集合を C_k とする．**図 9** では，$C_k = \{p_1, p_2, p_3\}$ である．

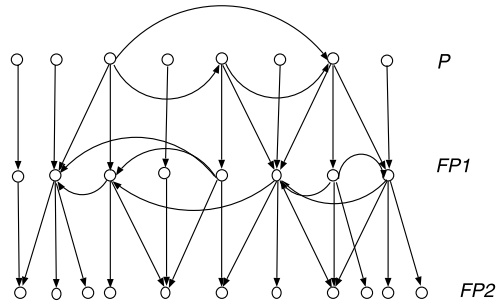


図 6 有向グラフ $G(V, E)$
Fig. 6 A directed graph $G(V, E)$.

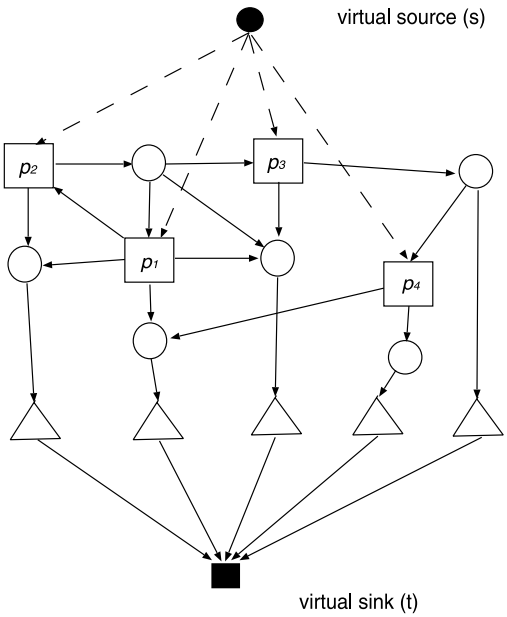


図 7 周辺グラフ $G'(V', E')$
Fig. 7 A vicinity graph $G'(V', E')$.

9. k を n 以下の間 1 増加させて 5〜8 を繰り返す．ただし，シード集合 Pk についてすでに計算済みの場合は 6，7 の処理を省略する．

10. すべての C_k が互いに素であればそれぞれがクラスタになり，要素が重なった集合があれば，それらの代わりにその和集合がクラスタになる．

たとえば，ページ p_1，p_2，p_3，p_4，p_5，p_6 があり，9. までの処理を行って 6 つの集合 $C_1 = \{p_1, p_2, p_3\}$，$C_2 = \{p_1, p_2, p_3\}$，$C_3 = \{p_3\}$，$C_4 = \{p_3, p_4\}$，$C_5 = \{p_5, p_6\}$，$C_6 = \{p_6\}$ を得たとすると，$C_1 \cup C_4$ と C_5 の 2 つのクラスタ（**図 10**）ができる．

図 10 のページ p_1，p_2 は互いにリンクを張っており，強連結グラフを構成している．一方，ページ p_3，p_4 は一方的なリンクしかないため強連結グラフを構

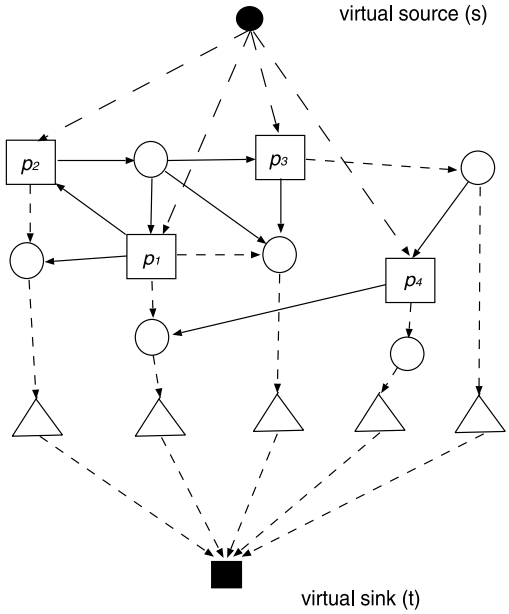


図 8 最大流アルゴリズムを実行（点線は飽和辺）
Fig. 8 Maximun flow algorithm is executed (The dotted lines are saturated edges).

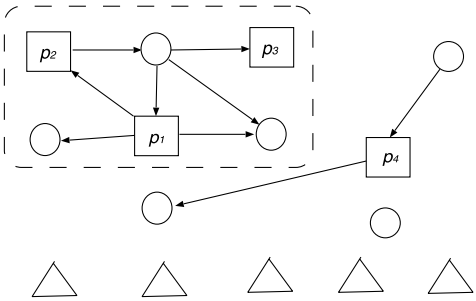


図 9 飽和辺を削除した不飽和辺のみのグラフ
Fig. 9 Graphs with unsaturated edges.

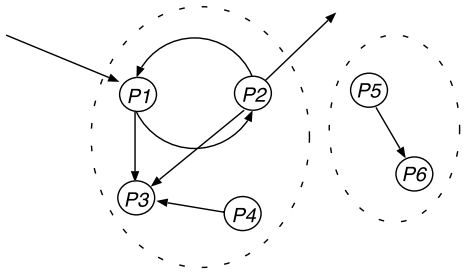


図 10 クラスタリングされたページとそのリンク関係
Fig. 10 The clustered pages.

成できない．しかし，ページ p_3 はページ p_1 と p_2 からリンクが張られており，提案方法では同じクラスタ（C_1 または C_2）に分類される．ページ p_4 も同様にページ p_3 とクラスタ C_4 を構成する．クラスタ C_1 と

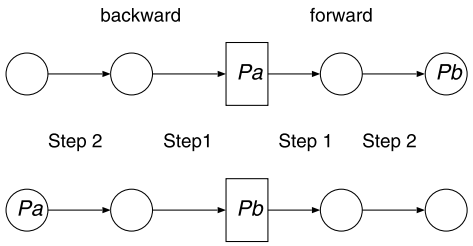


図 11 リンク方向とリンクの逆方向
Fig. 11 Forward links and backward links.

C_4 はページ p_3 を共有することから，ページ p_1，p_2，p_3，p_4 からなるクラスタ $C_1 \cup C_4$ ができる．ページ p_5 と p_6 も同様に強連結グラフは構成しないが，p_5 から p_6 へのリンクのみ存在することからクラスタ C_5 を構成する．

3.2 高速化のための周辺グラフの制限

コミュニティ発見と違い，検索結果をクラスタリングすることに時間はかけられない．より早くクラスタリングのための計算を行い，検索結果をユーザに返す必要がある．そこで，ページ数を V，リンク数を E とすると最大流アルゴリズムは $O(VE)$ であること[13]から，構築する周辺グラフに制限を加え小さくすることでスループットを上げる．制限方法として以下の 3 つの方法を採用した．

1. コミュニティを発見するために，Imafuji ら[10]はシードから前後に 2 ステップでたどれるページ集合（頂点）とそのリンク（有向辺）から周辺グラフを構築した．しかし，前節で説明したように提案方法はシードからリンクの前方方向に 2 ステップでたどれるページの集合に制限する．ページ p_i をクラスタリングできれば十分であり，そのために必要最小限の周辺グラフがあればよい．たとえば，**図 11** のようにシードページ p_b がシードページ p_a からリンクされていることを見つけるには，p_b からリンクの逆方向に探索しなくても p_a のリンク方向を調べるだけでよい．リンクの逆方向，リンクで参照されていることを調べるのは現在それほど難しくないが，今回はリンクの方向に探索することにした．また，ステップ数を増やすことは，たどれるページ数が指数的に増加するため，2 ステップに限った．

2. 図 6 のページ集合 $FP1$ の要素ページにおいてページ集合 $FP2$ の要素ページへのリンクのみでページ集合 Pk や $FP1$ の要素ページへのリンクがないものは周辺グラフから削除する．削除されたページからのみリンクされている $FP2$ のページも一緒に削除する．Pk や $FP1$ のページへのリンクがないというこ

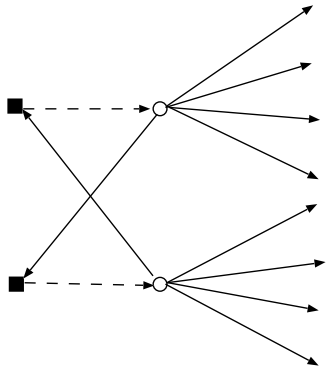


図 12 ハブが介在することで強連結になる例
Fig. 12 An example of strongly connected by hub pages.

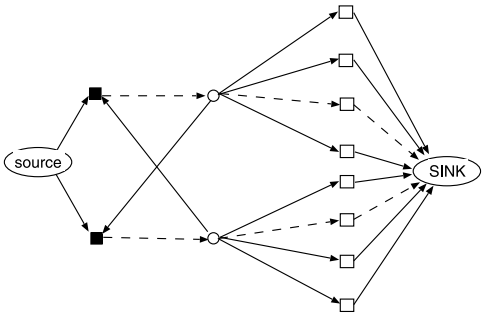


図 13 ハブへの入力リンクが飽和辺（波線矢印）になる様子
Fig. 13 In-links to hub pages are saturated.

とは，シードページの間のつながりには無関係であり，p_i をクラスタリングすることには影響しないからである．

3. 出力リンク数が 50 を超えるページをハブページと考え，周辺グラフから削除した．**図 12** のようにハブページが存在すると簡単に強連結グラフが構成される．しかし，ハブページを介在するとページ間の類似性は早く薄れていってしまう．たとえば，総合大学のトップページからリンクされているページは逆にトップページにリンクしていることが多く強連結グラフを構成するが，同じ大学という以外にはまったく類似性のない場合がある．出力リンク数が多ければ多いほど，そのリンクによる関係は弱い．最大流アルゴリズムでは，このようなハブページが存在するとハブページへの入力リンクが簡単に飽和辺（**図 13**）になってしまい，クラスタリング時に存在しない辺となる．入力リンクが存在しないということは，そのページにたどり着くことができない．つまり，そのページが存在しないのと同じである．このようにハブページの存在自体がクラスタリングにはあまり影響しないことから，最大流アルゴリズムを適用する前に周辺グラフから削除することにする．

表 1 rigid 判定による結果（%）
Table 1 Results based on rigid relevance judgment (%).

	平均適合率	適合率	再現率
METAL	36.0	44.9	75.4
強連結グラフ	20.3	42.1	48.3
提案方法	21.4	44.9	50.3

表 2 relaxed 判定による結果（%）
Table 2 Results based on relaxed relevance judgment (%).

	平均適合率	適合率	再現率
METAL	30.0	48.0	53.2
強連結グラフ	25.4	46.2	52.1
提案方法	33.5	52.4	60.5

4. 実験結果と考察

4.1 実 験 結 果

提案方法を定量的に評価するために，NTCIR-4 の Web タスク D[14] のデータを使って実験を行った．Formal Run で使われたのは 11 の topic についての検索結果である．各 topic についての検索結果はそれぞれ 200 位まで Web ページがリストされている．これをクラスタリングする．

比較のため強連結グラフによるクラスタリングと提案方法の 2 つの実験を行った．強連結グラフはもとになるグラフを 3.1 節で説明した提案方法と同じ周辺グラフに制限し，そこから抽出した．しかし，3.2 節で説明した制限は行っていない．

クラスタリングの評価方法も NTCIR-4 の Web タスク D と同じ方法を採用する．それは，利用者が明確な検索要求を抱いてブランジングするような場面を想定して検討された評価方法である．ここでは，分類結果における適合ページの分布を分析することとし，適合ページの多いクラスに含まれるページ群のランキングに関する精度と再現率を算出する．具体的には適合ページを多く含むクラスタを順番にソートする．上位 20 個のページを取りだし，平均適合率，20 位までの適合率（表では適合率と表記），20 位までの再現率（表では再現率と表記）を計算する．

検索結果は高適合（S），適合（A），部分的適合（B），不適合（C）の 4 段階で評価される．**表 1** の rigid は高適合と適合のみを適合ページと判断したことを表しており，**表 2** の relaxed は不適合以外をすべて適合ページと判断したことを表している．各表には，実験した強連結グラフによるクラスタリングと提案方法のほかに，比較のためコンテンツ・マイニングによって

表 3 クラスタサイズの比較
Table 3 Comparison of cluster size.

	クラスタ数	最大サイズ	非クラスタ数
METAL	48.5	51.4	8.7
強連結グラフ	23.5	13.2	124.6
提案方法	27.1	9.5	143.7

表 4 クラスタリング結果の例（METAL）
Table 4 An example of created clusters (METAL).

ページ ID	順位	ラベル名	適合判定
NW011452773	1	宇宙	A
NW008449606	2	宇宙	B
NW011452772	3	宇宙	B
NW011452783	4	宇宙	B
NW001673028	5	宇宙	C
NW013290383	6	宇宙	C
NW006585020	7	宇宙	C
NW011452812	8	宇宙	B
NW011452811	9	宇宙	B
NW012004625	10	宇宙	C
NW012004595	11	宇宙	C
NW008449608	12	宇宙	A
NW013290356	13	宇宙	C
NW003387031	14	宇宙	C
NW002658037	1	天文学	A
NW009614525	2	天文学	C
NW001443580	3	天文学	C
NW002280945	4	天文学	S
NW011649170	5	天文学	S
NW009382688	6	天文学	A
NW008449515	7	天文学	C
NW012004638	1	研究	C
NW001088770	2	研究	B
NW002658037	3	研究	A
NW008450014	4	研究	C
NW012004640	5	研究	C
NW008449782	6	研究	C
NW008450141	7	研究	C
NW008450173	8	研究	C
NW008449433	9	研究	C

クラスタリングした METAL [15] の結果も並べて表示する．NTCIR-4 の Web タスク D に参加したチームはすべてコンテンツ・マイニングによるクラスタリングを行っていた．そのなかで METAL は Formal Run において最も良い結果を記録しているため，コンテンツ・マイニングによるクラスタリング方法の最も成功した例として比較することにした．

表 4，**表 5**，**表 6**，**表 7** はクエリ 3（**図 14**）に対するクラスタリング結果である．表 4 の METAL はコンテンツ・マイニングによる方法であるため，クラスタ名として意味のあるラベルが得られている．一方，表 5 と表 6 はコンテンツを分析していないので意味のあるラベル名が得られない．そこでラベル名を便宜上 C1，C2，C3，C4，C5 とする．表で順位とはクラスタ内での順位である．提案方法も強連結グラフによるクラスタリング方法もクラスタへの分類だけでクラスタ内での順位は計算していない．そこで，クラスタリングする前の検索結果リストでの順位を利用している．

表 5 クラスタリング結果の例（強連結グラフによるクラスタリング）
Table 5 An example of created clusters (strongly connected graph).

ページ ID	順位	ラベル名	適合判定
NW011452773	1	C1	A
NW011452772	2	C1	B
NW003295673	3	C1	C
NW011452783	4	C1	B
NW011452812	5	C1	B
NW011452811	6	C1	B
NW003295675	7	C1	C
NW011453148	8	C1	C
NW008449520	1	C2	B
NW008449505	2	C2	A
NW007125925	3	C2	C
NW008449782	4	C2	C
NW008450141	5	C2	C
NW008450173	6	C2	C
NW008449480	7	C2	C
NW008449794	8	C2	C
NW008449608	9	C2	A
NW007125916	10	C2	C
NW008449609	11	C2	A
NW008449515	12	C2	C
NW012004638	1	C3	C
NW012004596	2	C3	C
NW012004608	3	C3	C
NW007125920	4	C3	A
NW007125730	5	C3	A
NW012004597	6	C3	C
NW007125770	7	C3	B
NW012004601	8	C3	C
NW001088770	1	C4	B
NW001088769	2	C4	C

表 4 から表 6 は上位 30 ページまでとした．表 1 や表 2 の平均適合率，適合率，再現率はすべて上位 20 位までで行っている．さらに，クエリ 3 の適合ページ総数は 24 であるため，30 位以下では適合ページの割合が小さいからである．クラスタ単位での適合ページの割合は表 7 で確認できる．この表は上位 30 位を含むクラスタまでを記載した．

4.2 考　　察

rigid で判定した場合，表 1 の結果から，適合率，再現率ともに，METAL，提案方法，強連結グラフによるクラスタリングの順番に良いことが分かる．一方，relaxed で判定した場合は，表 2 の結果から，適合率，再現率ともに提案方法，METAL，強連結グラフによるクラスタリングの順番に良いことが分かる．

表 6 クラスタリング結果の例（提案方法）
Table 6 An example of created clusters (proposal method).

ページ ID	順位	ラベル名	適合判定
NW011452773	1	C1	A
NW011452772	2	C1	B
NW003295673	3	C1	C
NW011452783	4	C1	B
NW011452812	5	C1	B
NW011452811	6	C1	B
NW003295675	7	C1	C
NW008449520	1	C2	B
NW008449606	2	C2	B
NW008449505	3	C2	A
NW008450014	4	C2	C
NW008449782	5	C2	C
NW008449480	6	C2	C
NW008449608	7	C2	A
NW008449609	8	C2	A
NW007125920	1	C3	A
NW007125730	2	C3	A
NW007125770	3	C3	B
NW007125916	4	C3	C
NW002951265	1	C4	C
NW002935649	2	C4	A
NW002471253	3	C4	C
NW002885232	4	C4	C
NW002935655	5	C4	A
NW001088770	1	C5	B
NW001558702	2	C5	A
NW001088843	3	C5	C
NW001088769	4	C5	C
NW001001549	5	C5	C

表 7 クラスタ単位での適合率
Table 7 Precisions in each cluster.

ラベル名	サイズ	S	A	C	適合率
METAL					
宇宙	14	0	2	5	50.0
天文学	7	2	2	0	57.1
研究	14	0	1	3	28.6
強連結グラフによるクラスタリング					
C1	8	0	1	4	62.5
C2	12	0	3	1	33.3
C3	8	0	2	1	37.5
C4	4	0	1	1	50.0
提案方法					
C1	7	0	1	4	71.4
C2	8	0	3	2	62.5
C3	4	0	2	1	75.0
C4	5	0	2	0	40.0
C5	5	0	1	1	40.0

どちらの場合も提案方法は強連結グラフによるクラスタリングより良い結果になっている．また，提案方法は METAL と比較して同程度の適合率，再現率をあげており，コンテンツ・マイニングによる様々なクラスタリング方法と比較しても同程度以上の能力があると考えられる．

<NUM>0003</NUM>
<TITLE> コペルニクス，地動説，キリスト教 </TITLE>
<DESC> コペルニクスの地動説がキリスト教社会でどのように受容されていたかを調べたい． </DESC>
<BACK> 史料を忠実に追うとコペルニクスの地動説（太陽中心説）は 16 世紀にすでにキリスト教に認められていた，という科学史家もいるようだ．この問題について事実関係を知りたい． </BACK>
<RELE> 文書内にて地動説に関するコペルニクス本人の著作や発言などとキリスト教の関係に言及されていれば適合文書とする．コペルニクス本人とは関係なく，地動説とキリスト教の関係を述べるに留まった文書，またはコペルニクス本人とキリスト教の関係を述べたのみで地動説に触れていない文書は部分的適合とする．コペルニクスの地動説の説明のみでキリスト教との関連が言及されていないものは不適合とする．</RELE>

図 14 クエリ 3 の検索課題
Fig. 14 Retrieval query 0003.

提案方法は relaxed では METAL に優り rigid では劣る．これは適合ページのリンク的近傍に部分的適合ページが存在している可能性が高いことを意味している．提案方法はこのような部分的適合ページをクラスタに含みやすく，rigid では不適合と判断されてしまい，METAL に劣ってしまうと考えられる．

適合率と再現率以外の定量的評価としてクラスタのサイズやクラスタの数が考えられる．**表 3** に，クラスタリングでできたクラスタの数（表ではクラスタ数），最も大きいクラスタのサイズ（表では最大サイズ），独自のクラスタを構成しないページ（表では非クラスタ数）を示す．最も大きいクラスタのサイズを比較すると提案方法は最も小さく，クラスタに分類されないページの数もきわめて多い．表 4 のクラスタ「宇宙」のサイズは 14，クラスタ「天文学」のサイズは 7 である．クラスタ「研究」のサイズは 9 のように見えるが，表 7 から実際には 14 であることが分かる．一方，提案方法は，表 6 で確認できるようにクラスタ C1 のサイズは 7 であり，C2 のサイズは 8 であり，C3 のサイズは 4 しかない．これは，提案方法が図 3 のようにリンク構造が密になったページ集合をクラスタとして切りだせており，そのサイズは小さいことを

意味している．一方，表 2 や表 7 で適合率が良いことから，適合ページと不適合ページが同じクラスタに分類されることは少ないことが分かる．つまり，同じクラスタに分類されたページは似通ったものが多い．

クラスタサイズに関して，同じストラクチャ・マイニングに基づく強連結グラフによるクラスタリングは提案方法と同様の傾向が見られる．しかし，この方法は図 12 のようなハブページの存在によって，あまり類似していないページも同じクラスタに分類されるというノイズを生じてしまう．ノイズのために提案方法よりクラスタのサイズが大きくなるが，逆に適合率・再現率は下がってしまったといえる．表 5 において *NW011453148* のページはクラスタ C1 に，*NW008450141* のページは C2 に分類されているが，どちらも不適合ページでありノイズの例である．提案方法でこれらのページはノイズと判断され，表 6 の C1 や C2 に分類されていないことが確認できる．

提案方法は，図 10 のページ p_3 や p_4 ような強連結グラフでは同じクラスタに含まれるとは判断できないようなページもクラスタに参加させることができる．この効果で，提案方法は強連結グラフによるクラスタリングよりもクラスタのサイズが大きくなると期待した．実際の実験で p_3 や p_4 のようなページの存在は確認できた．具体的に，ページ *NW008449606* は提案方法の表 6 でクラスタ C2 に分類されているが，クラスタ C2 のほかのページと強連結でないため表 5 のクラスタ C2 には含まれていない．しかし，このようなページは期待した程絶対数が多くなく，クラスタのサイズを大きくするような効果はなかった．

クラスタのサイズを大きくするための方法として，有向辺（リンク）の辺容量を調節することが考えられる．提案方法では仮想ソース s からシード pk_i への有向辺（仮想リンク）以外の有向辺（リンク）の辺容量をすべて 1 としたが，入出力リンク数やページのハブ値，オーソリティ値などから決める方法が考えられる．しかし，これは，クラスタのサイズを大きくする可能性があると同時に適合率や再現率を下げる可能性もある．さらにハブ値やオーソリティ値を決めるには反復計算が必要となり処理時間が増大してしまう．

処理時間の増大を防ぐために，sink への有向辺以外の辺容量をすべて 1 より大きな一定値にすることも考えられる．これは不飽和辺になりにくくなりクラスタのサイズは大きくなる．しかし，FP2-t 間が最小切断になり，連結グラフと同値になってしまう．連結グラフは強連結グラフよりも連結度が低い．あまり関連性のないページも同じクラスタに分類することになり，それらはノイズとなる．つまり，適合率は低くなる．

表 8 制限するページの出力リンク数の比較
Table 8 Comparison of numbers of restricted out-links.

	40	50	100	制限なし
適合率	46.1	52.4	52.5	52.3
再現率	51.3	60.5	60.5	60.6
処理時間	53 秒	55 秒	2 分 26 秒	15 分 39 秒

クラスタのサイズを大きくするための方法として最も期待されるのは，コンテンツ・マイニングによるクラスタリング方法との統合である．両者はまったく異なる観点からクラスタリングを行っているためどのように統合するかという問題があるが，コンテンツ・マイニングによるクラスタリングは，提案方法のようなストラクチャ・マイニングによるクラスタリングに比べて従来から検討されてきているため，参考にできることが多い．したがって，提案方法とコンテンツ・マイニングによるクラスタリング方法の統合は今後の課題とする．

4.3 近似方法の検証実験

3.2 節の「高速化のための周辺グラフの制限」で 3 つの方法の説明を行った．このうち最後の方法は近似方法である．出力リンク数が 50 以上のページをハブページと見なしてクラスタリングの結果には影響しないと仮定し，周辺グラフから削除した．この仮定に対する正当性を確認する必要がある．そこで，出力リンク数の大きいページを周辺グラフから削除することによるクラスタリング結果への影響を実験から検証した．そのために削除するページの出力リンク数を 40，50，100 とする場合，それぞれ適合率，再現率，処理時間（クエリが入力されてからクラスタリング結果が出力されるまでの時間）がどのように変化するか調べる．比較のために，出力リンク数による削除をしない場合も行った．判定は relaxed とした実験結果が**表 8** である．出力リンク数が 40 以上のものを削除すると適合率，再現率ともに悪くなるが，50 以上のものを削除してもあまり変わらないことが分かる．一方で，周辺グラフに制限をかけないと処理時間は長くなり，検索エンジンとして現実的な応答ができなくなってしまう．なお，処理時間の測定に使用したコンピュータは SunFire280R，UltraSPARC III 800 MHz CPU，1 GB MainMemory，OS は Solaris 8，言語は Perl である．CGI でよく使われることから使用言語を Perl としたが，より高速な C 言語を使えば表 8 の処理時間も短縮できる．

5. おわりに

最大流アルゴリズムを用いてリンク構造を解析することにより検索結果をクラスタリングするという新しいクラスタリング方法を提案した．この方法はコンテンツをまったく考慮しない，ストラクチャ・マイニングのみによるクラスタリング方法である．この提案方法は，NTCIR-4 のデータを使った定量的評価実験によって適合率・再現率に関して良好な性能を持つことが示された．一方で，クラスタのサイズがかなり小さいことも確認された．この問題を克服するために，コンテンツ・マイニングによるクラスタリング方法と統合することが考えられ，今後の課題である．

謝辞 この研究の一部は科学研究費補助金（特定領域研究，研究課題番号：16016273）による．

参考文献

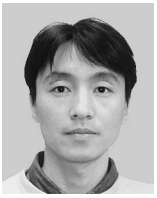
1) Zamir, O., et al.: Grouper: A Dynamic Clustering Interface to Web Search Results, *Proc. WWW8* (1999).
2) Zeng, H.J., et al.: Learning to Cluster Web Search Results, *ACM SIGIR04* (2004).
3) Vivisimo. http://vivisimo.com/
4) WSM. http://wsm.directtaps.net/
5) Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).
6) Kumar, R., et al.: Trawling the Web for Emerging Cyber-communities, *Proc. 8th international conference on World Wide Web*, pp.1481–1493 (1999).
7) Wang, Y. and Kitsuregawa, M.: Use Link-based Clustering to Improve Search Results, *Proc. 2nd International Conference on Web Information System Engineering*, IEEE Computer Society (Dec. 2001).
8) 正田備也，高須淳宏，安達 淳：パラメータ化された連結性に基づく Web ページのグループ化，*DBSJ Letters*, Vol.1, No.1 (2002).
9) 正田備也，高須淳宏，安達 淳：新しい連結性概念と Web ページのグループ化への応用，*DBSJ Letters*, Vol.2, No.1 (2003).
10) Imafuji, N. and Kitsuregawa, M.: Finding Web Communities by Maximum Flow Algorithm Using Well-Assigned Edge Capacities, *IEICE Trans. Inf. Syst.*, Vol.E87-D, No.2, pp.407–415 (2004).
11) Flake, G.W., Lawrence, S. and Giles, C.L.: Efficent Identification of Web Communities, *In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.150–160 (2000).
12) 今藤紀子，喜連川優：Max-Flow コミュニティグラフとその特徴分析，*DBSJ Letters*, Vol.3, No.1 (2004).
13) Sedgewick, R.: *ALGORITHMS*, 2nd ed., 近代科学社 (1993).
14) Eguchi, K.: Overview of the Topical Classification Task at NTCIR-4 WEB, *Working Notes of the 4th NTCIR Meeting*, Supplement volume 1, pp.48–55 (2004).
15) Ohta, M., Narita, H. and Ohno, S.: Overlapping Clustering Method Using Local and Global Importance of Feature Terms at NTCIR-4 Web Task, *Working Notes of the 4th NTCIR Meeting*, Supplement volume 1, pp.102–110 (2004).

(平成 17 年 9 月 21 日受付)
(平成 18 年 1 月 6 日採録)

(担当編集委員 江口 浩二)

大野 成義（正会員）
金沢大学大学院自然科学研究科博士課程修了．博士（理学）．滋賀職業能力開発短期大学校を経て，現在，職業能力開発総合大学校に勤務．同時に東京都立大学社会人博士課程に在学中．Web マイニングの研究に従事．日本データベース学会，日本物理学会各会員．

渡辺 匡
東京都立大学大学院工学研究科在学中．クラスタリングの研究に従事．



片山 薫（正会員）
首都大学東京システムデザイン学部研究員．2000 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．データベースシステムに関する研究開発に従事．日本データベース学会会員．

石川 博（正会員）

東京都立大学大学院工学研究科教授．東京大学理学部情報科学科卒業．富士通研究所を経て 2000 年より現職．東京大学博士（理学）．著書に“Object-Oriented Database System”（Springer Verlag），“Database and Data Communication Network Systems: Techniques and Applications”（共著，Elsevier），『e–ビジネス技術入門教科書—ビジネスモデルと情報技術（IT）IT TEXT』（CQ 出版），『次世代データベースとデータマイニング—DB& DM の基礎と Web・XML・P2P への適用』（CQ 出版）等．国際論文誌 ACM TODS，IEEE TKDE，国際学会 VLDB，IEEE ICDE 等，学術論文多数．1994 年情報処理学会坂井記念特別賞，1997 年科学技術庁長官賞（研究功績者）受賞．現在，情報処理学会データベースシステム研究会主査．情報処理学会論文誌（データベース）共同編集委員長．International Journal Very Large Data Bases Editorial Board．日本データベース学会理事．仏国ナント大学招聘教授．電子情報通信学会，ACM，IEEE 各会員．

太田 学（正会員）

岡山大学大学院自然科学研究科助教授．1999 年東京大学大学院工学系研究科電気工学専攻博士課程修了，博士（工学）．東京都立大学大学院工学研究科助手を経て 2005 年より現職．情報検索，データマイニングとその Web 応用の研究に従事．電子情報通信学会，日本データベース学会，IEEE 各会員．