2005－DBS－137（Ｉ）（12）
2005／7／13

Max Flow アルゴリズムによる Web ページのクラスタリング方法

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．従って，リストが長い場合，求める情報を探すのは極めて難しい．そこで，検索結果をリストでなくクラスタリングして表示する方法を提案する．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 ページのも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 the needed information. It is often tedious and less efficient. In this paper, we propose a new link-based clustering approach to categorize search results returned from Web search engine. The maximum flow algorithm effective in finding the page sets with the link structures connected more precisely is used for the analysis of link information. In the experiments to evaluate method perfoemance quantitatively, we had good results using the data of NTCIR.

1. まえがき

インターネット上から必要な情報を得るために検索エンジンが良く利用されている．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．その検索結果の全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，順番に調べるには骨が折れる．

† 職業能力開発総合大情報工学科，相模原市
Department of Information and Computer Science ,Polytechnic University
†† 東京都立大学大学院工学研究科，八王子市
Graduate School of Engineering, Tokyo Metropolitan University
††† 岡山大学大学院自然科学研究科，岡山市
Graduate School of Natural Science and Technology, Okayama University

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 ページには他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，全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 Collec-

tion for IR Systems) のデータを利用して実験を行ったところ良好な結果を得た．

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二つに分類できる．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一つのクラスタにまとめる．従って，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．

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 [9),10)] と呼んでいる．クラスタリングでもこのリンク情報を使いストラクチャ・マイニングを行う方法が考えられる．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，例えば，Y.Wang ら [6)] 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．

具体的には，二つのページ 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 から出て行く M 個のリンクを表し，出力リンクを各次元に割り当てた M 次元ベクトルであり，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 N 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，例えばページ P と Q が相互にリンクを張っている場合と，そうでない場合との違いが鮮明でない．互いにリンクが直接張られた方がより類似内容である可能性が高いことから，リンクの共通度だけでは不充分ではないかと考えられる．

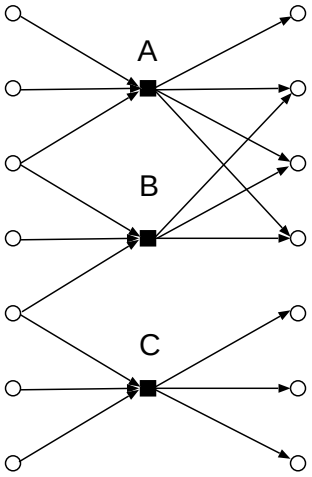

図 1 リンク先やリンク元を共有するページ

2.2 コミュニティ発見

ストラクチャ・マイニングの例としては，他にコミュニティ発見の研究がある．これは，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 ページ間のコミュニティを発見する．そこで使われる手法として，正田ら [4),5)] の連結性に基づく方法と今藤ら [2)] 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 からリンクを辿ってページ B に到達可能で，逆に B から A へも到達可能だとすると，この二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．更に，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，G.Flake ら [1)] はコミュニティを図 2 で示すような「コミュニティの外へのページへの (又は，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、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充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

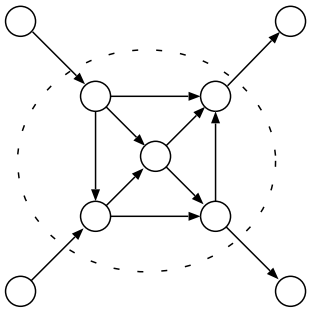

図 2 Web コミュニティの例

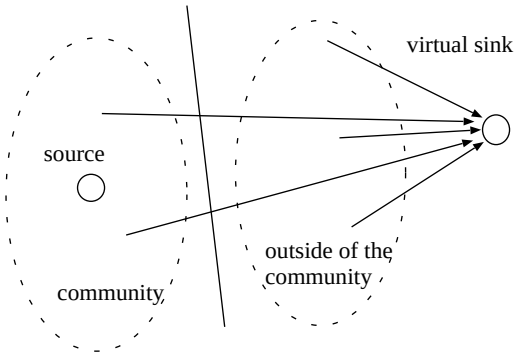

図 3 最大流と最小切断の関係

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 3). 更に，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 [3] 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

提案方法は以下の手順で行う．

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 $FP1$ とする．更に，$FP1$ 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 $FP2$ 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 $G(V, E)$(図 4) を考える．

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辿れる頂点 p_i の集合をシード集合 Pk とする．

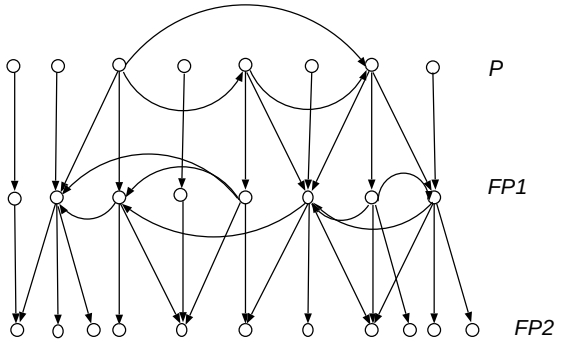

図 4 有向グラフ $G(V, E)$

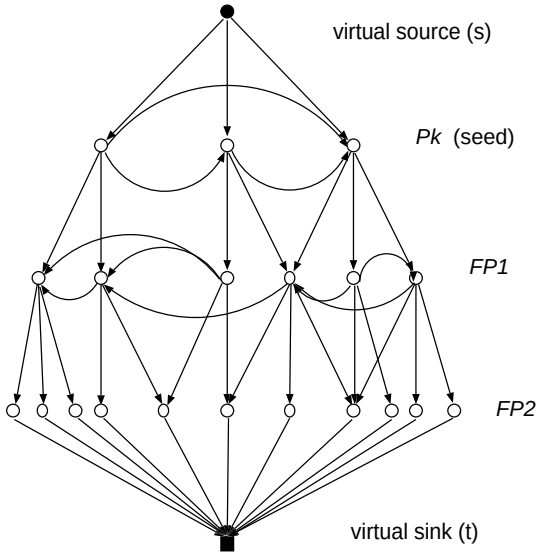

図 5 周辺グラフ $G'(V', E')$

6. 周辺グラフ $G'(V', E')$(図 5) を以下の手順で構築する．V のうち Pk から有向辺をつたって辿れる頂点および Pk を V' とし，その間のリンク (辺容量$= 1$) を E' とする．更に，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，$V' - \{V' \cap FP1\} - \{s \cup Pk \cup t\}$ の各頂点から t への有向辺 (辺容量 $= 1$) を E' に加える．

7. 最大流アルゴリズムを実行する．

8. p_k から不飽和辺を辿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．

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 について既に計算済みの場合は 6. 7. の処理を省略する．

10. 全ての 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

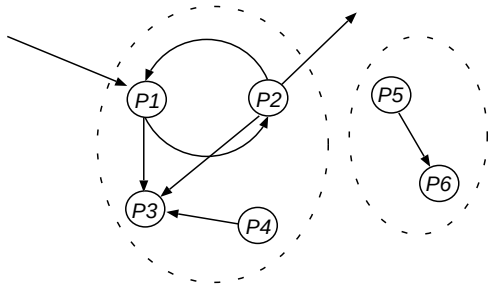

図 6 クラスタリングされたページとそのリンク関係

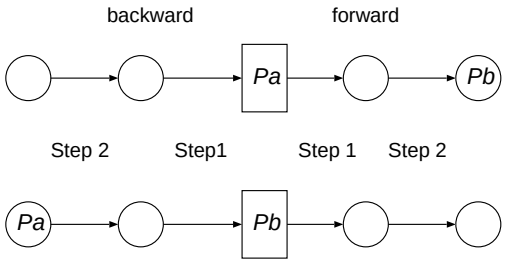

図 7 リンク方向とリンクの逆方向

代りにその和集合がクラスタになる．

例えば，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 6) ができる．

図 6 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成できない．しかし，ページ p_3 はページ p_1 と p_2 からのみ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 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，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，今藤ら [2] はシードから前後に 2 ステップで辿れるページ集合 (頂点) とそのリンク (有向辺) から周辺グラフを構築したが，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 ステップで辿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充分であり，そのために必要最小限の周辺グラフがあれば良い．例えば，図 7 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良い．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，辿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．

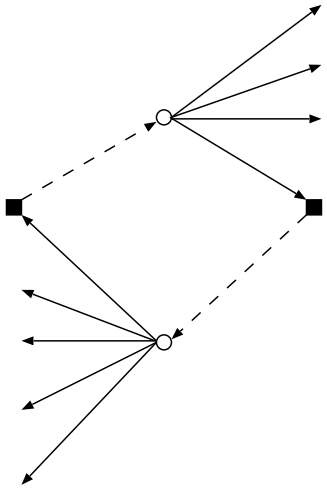

図 8 ハブが介在することで強連結になる例

2. 図 4 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 を越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 8 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．例えば，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全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程，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

表 1　rigid 判定による結果 (%)

	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 (%)

	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

うことは，そのページに辿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身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．

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[7] のデータを使って実験を行った．Formal Run で使われたのは 15 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. 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. で行ったさらなる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，適合，部分的適合，不適合の 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 relax は不適合以外を全て適合ページと判断したことを表している．

4.2 考　　察

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他に，比較のためコンテンツ・マイニングによってクラスタリングした METAL[8] 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全てコンテンツ・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 3　クラスタサイズの比較

	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

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であることがわかる．一方，relax 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であることがわかる．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．

提案方法は relax 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極めて多い．これは提案方法の類似度の判定が厳しく，類似のページが見つからないページが多くなり，もし，類似のページがあったとしても多くはないということを意味している．逆に類似度の判定が厳しいがために，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も少なく，表 2 のように提案方法が最も良い適合率・再現率を示しているといえる．

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 8 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言える．

提案方法は，図 6 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と考えられる．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．しかし，実際の実験で p_3 や p_4 のようなページが存在することは確認できたが，期待した程は絶対数が多くなく，クラスタのサイズを大きくするような効果はあまり果たせなかった．

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) 以外の有効辺 (リンク) の辺容量を全て 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．更にハブ値やオーソリティ値を決めるには反覆計算が必要となり処理時間が増大してしまう．

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．両者は全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，色々な経験が参考になると期待できる．従って，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．

3.2 の「高速化のための周辺グラフの制限」で説明した 3 つの方法のうち最後の方法は近似であり，出力リンク数を 50 としたことに対する正当性を確認する必要がある．そのためこの制限を適用しない方法についても実験して確認を行ったところ，クラスタリングの結果は微妙に異なるが，適合率や再現率は同じ値であった．実験結果から，出力リンク数を 50 としたことは悪影響をおよぼさないことを確認した．しかし，周辺グラフに制限をかけるための出力リンク数の最適値はいくらであるかまでは判定できなかった．

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全く考慮しない，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もつことが示された．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．

参考文献

1) G.W.Flake, S.Lawrence and C.L.Giles, “Efficent Identification of Web Communities,” In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 150-160, 2000.
2) N.Imafuji and M.Kitsuregawa, “Finding Web communities by Maximum Flow Algorithm Using Well-Assigned Edge Capacities,” IEICE transactions on Information and Systems, Vol.E87-D No.2 pages 407-415, 2004.
3) 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
4) 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.
5) 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.
6) Y.Wang and M.Kitsuregawa, “Use link-based Clustering to Improve Search Results.” Proceedings of the 2nd International Conference on Web Information System Engineering, IEEE Computer Society, Dec. 2001.
7) K.Eguchi, “Overview of the Topical Classification Task at NTCIR-4 WEB,” Working Notes of the 4th NTCIR Meeting, Supplement volume 1, pp.48-55, June 2004.
8) M.Ohta, H.Narita and S.Ohno, “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, June 2004.
9) J.Kleinberg, “Authoritative sources in a hyperlinked environment,” Journal of the ACM, Vol.46, No.5, September 1999, pp.604-632.
10) R.Kumar et.al.”Trawling the Web for emerging cyber-communities,” Proceeding of the 8th international conference on World Wide Web, 1999, pp.1481-1493.