

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の提案

Clustering Web Pages Based on Maximum Flow Algorithm

大野 成義[▽] 渡辺 匡^{◇◇}
片山 薫[▽] 石川 博[▽] 太田 学[△]

Shigeyoshi OHNO Masashi WATANABE
Kaoru KATAYAMA Hiroshi ISHIKAWA
Manabu OHTA

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。

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 page sets connected tightly by hyperlinks is used for the analysis of link information. In order to evaluate method performance quantitatively, we conducted experiments using the NTCIR data and had good results.

1. はじめに

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。

[▽] 正会員 職業能力開発総合大学校情報システム工学科
ohno@uitec.ac.jp

^{◇◇} 非会員 東京都立大学大学院工学研究科修士課程
wata300@jcom.home.ne.jp

[▽] 正会員 東京都立大学大学院工学研究科 {katayama, ishikawa}@eei.metro-u.ac.jp

[△] 正会員 岡山大学大学院自然科学研究科
ohta@suri.it.okayama-u.ac.jp

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。

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出すことをストラクチャ・マイニング[9]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、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。そこで、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。

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十分ではないかと考えられる。

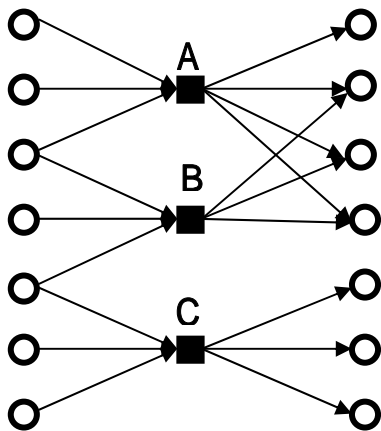


図1 リンク先やリンク元を共有するページ
Fig.1 Pages sharing links.

2.2 Web 全体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、他にクラスタリングの対象を検索結果に制限せず Web 全体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 [4] [5] の連結性に基づく方法とがあり、コミュニティを発見する手法として今藤ら [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イ値)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 [1] はコミュニティを「コミュニティ外のページへの(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2)。

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検索結果のクラスタリングに利用することを提案する。

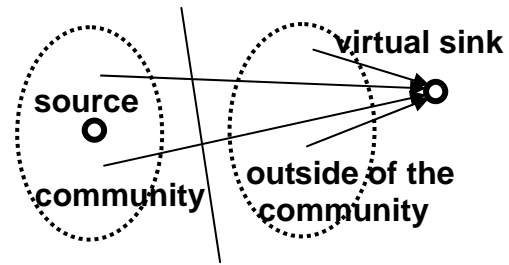


図2 最大流と最小切断の関係
Fig.2 Max Flow and Min cut.

3. 最大流アルゴリズムを用いたクラスタリングの方法

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。

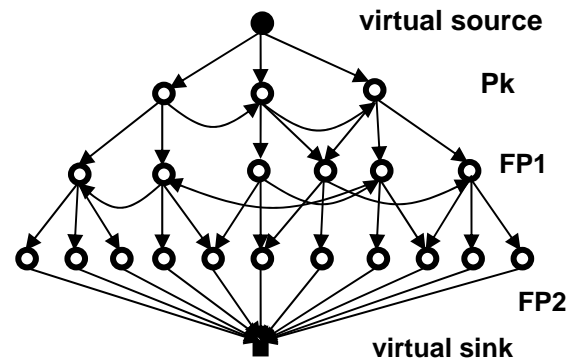


図3 周辺グラフと仮想ソースと仮想シンク

Fig.3 The vicinity graph with virtual source and virtual sink.

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 $FP1$ $FP2$ 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。
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迎れる頂点 p_i の集合をシード集合 P_k とする。
6. 周辺グラフ $G'(V', E')$ を以下の手順で構築する(図3)。 V のうち P_k から有向辺をつたって迎れる頂点および P_k 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, p_k) 、 p_k P_k を E' に加える。また、仮想シンク t を V' に加え、 V' $FP2$ 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 からの集合を 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 P_k 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, C_4 と C_5 の2つのクラスタ (図4)ができる .

図4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, 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 .

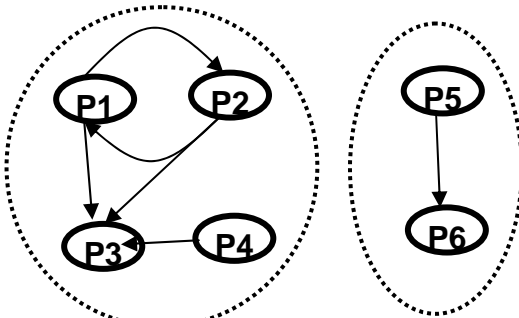


図4 クラスタリングされたページとそのリンク関係
Fig.4 The clustered pages.

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もになるグラフを 3. 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 .

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ウ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タ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 relaxed 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 .

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

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

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]の結果も併せて表示する . 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 .

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わかる . 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 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4のページ p_3 や p_4 の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。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 p_{k_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。

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る。

[文献]

- [1] G.W.Flake, S.Lawrence and C.L.Giles: "Efficient Identification of Web Communities", In 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 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 pp 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.

大野 成義 Shigeyoshi OHNO

金沢大学大学院自然科学研究科修了。博士(理学)。職業能力開発総合大学校に勤務、同時に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。情報処理学会、日本データベース学会各会員。

渡辺 匡 Masashi WATANABE

東京都立大学大学院在学中。クラスタリングの研究に従事。

片山 薫 Kaoru KATAYAMA

東京都立大学大学院工学研究科助手。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。情報処理学会、日本データベース学会各会員。

石川 博 Hiroshi ISHIKAWA

東京都立大学大学院工学研究科教授。富士通研究所を経て2000年より現職。1979年東大・理卒。博士(理学)。データベースシステムの研究開発に従事。情報処理学会、ACM、IEEE各会員。International Journal on Very Large Data Bases Editorial Board, ACM Sigmod Japan評議員、日本データベース学会理事。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)編集委員長。ナント大学(フランス)エコールポリテクニック招聘教授。著書に「E-ビジネス技術入門教科書」「次世代データベースとデータマイニング」(CQ出版)など。

太田 学 Manabu OHTA

岡山大学大学院自然科学研究科助教授。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。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)。情報検索、データマイニングとそのWebへの応用に興味を持つ。情報処理学会、電子情報通信学会、日本データベース学会、IEEE各会員。