2007－ICS－148 (10)
2007／7／4

Webページ集合からのWebページのコンテンツと構造を用いたクラスタリングによるトピックマップの抽出

間瀬 心博 †， 山田 誠二 ††， 新田 克己 †，

† 東京工業大学大学院総合理工学研究科 †† 国立情報学研究所，総合研究大学院大学

本論文では，Webページ集合からトピックマップを半自動で抽出する手法を提案する．従来のネットワーク構造に注目したクラスタリング手法に，Webページのコンテンツによる類似度，Webサイトのディレクトリ構造とWebページ間のリンクを利用した重み付けを導入する．Webページ間のリンクに内包されるトピックの関係を考慮し，Webサイトのディレクトリから Webページ間のリンクの意味を推定することで，Webページの集合からトピックだけでなくトピックの関連も併せて抽出する．Webユーザのブラウジング履歴から生成したWebページセットから，従来のクラスタリング手法と提案手法を用いてトピックマップを抽出し比較することで，提案手法の検証を行う．

Extracting Topic Maps from Web histories by clustering with Web structure and contents

MotohiroMase † SeijiYamada †† KatsumiNitta †

† IGSE, Tokyo Institute of Technology †† National Institute of Informatics, SOKENDAI

In this paper, we describe a method to semi-automatically extract Topic Maps from a set of Web pages. We introduce the following two points to the existing clustering method: The first is merging only the linked Web pages, to extract the underlying relationship of the topics. The second is introducing the contents similarity of Web pages and the weights that are based on the types of links and the distance between the directories in which the pages are located, to generate dense clusters. To evaluate the extracted topic maps and proposed method, we conduct experiment for comparing our method and the existing clustering method. We show the effectiveness of our approach using similarities of contents and weights from Web structure information.

1 はじめに

現在，膨大な量のWebページを利用した情報収集は非常に有用であり重要なものになっている．Webページの総量は年々増加し続けており，2005年1月時点で110億を超える[4]とされている．そのため，必要な時に自分の求める情報をWeb上から探し出し，またいかに獲得した情報を分類・整理するかが問題[5]となっており，様々な研究がなされている．そのうちの一つとしてトピックマップ(Topic Maps)があげられる[6]．トピックマップは，ユーザの概念や知識にそって情報を整理・分類するために用いられる国際規格であり，様々な情報リソースとユーザのもつ概念や知識とを結びつけ，またそれらの概念間の関係を表し，効率的に必要な情報へアクセスしやすくするものである．トピックマップを作成するには，トピックマップの対象となる領域や情報リソース，トピック等の選定，トピックマップの作成等の作業が必要となる．Ontopoly[1]等のトピックマップエディタを利用することで，これらの作業のコストを削減することは可能であるが，基本的には手動で行われる．また，自動でトピックマップを作成するには，既存のXML等のメタデータを用いてトピックマップに変換する方法[11, 8]が考えられる．しかし，ユーザが日常的に接しているWebではRDF，XML等の構造化されたメタデータも増えてきているものの，多くのページはHTMLで記述された半構造化されたデータであり，従来の手法をそのまま適用することは難しい．そのため，半構造データから直接トピックマップを作成する手法が必要となる．

そこで，本研究では半構造データであるWebページの集合から，トピックマップを自動的に抽出することを目指す．しかし，完全なトピックマップを自動的に抽出するのは難しいため，ユーザとのインタラクションを通じてトピックマップを完成させることとし，まずユーザに提示するためのトピックマップの雛型を自動抽出することを目的とする．Webページのコンテンツや Webページ間のリンクの構造には，様々なトピックやそれらの関係が内包されていると考えられるため，これらの点を考慮した

[1] http://www.ontopia.net/solutions/ontopoly.html

クラスタリング手法を用いて，トピックマップの抽出を試みる．ユーザが実際に閲覧したWebページや，その周辺に存在する未見のWebページからトピックマップを抽出することができれば，ユーザの獲得した情報は分類・整理され，また未見の関連する情報へもアクセスすることが容易になると期待できる．

Webからの情報抽出に関しては，Webネットワークからのグラフ構造の抽出に関する研究が多く行われているが，トピックマップの抽出を直接の目的にしたものはない．Webから構造を取り出す研究としては，Webコミュニティの抽出に関する研究が多く行われている．Broderらは大規模なWebデータから関連するWebページ集合の特徴的なリンク構造である完全二部グラフ構造を探索することで，Webコミュニティの発見を行っている[1]．Flakeらは最大流量最小カット定理をWebのネットワーク構造に適用し近似的に計算することで，Webコミュニティを発見している[2]．Girvanらはグラフ構造のエッジの媒介性であるbetweennessに注目し，betweennessの高いエッジを削除していくことで，密なグラフ構造を発見している[3]．しかし，Girvanらの手法はネットワークのノード数をn，エッジ数をmとするとき，時間計算量は$O(n^3)$もしくは$O(m^2n)$であり，大規模ネットワークへの適用が困難であった．この問題に対して，Newmanは同手法の評価関数の最大値問題として階層的クラスタリングを行うことで，計算量を減らし同程度の精度の結果を得ている[10]．これらの研究はネットワーク構造のみに注目したアプローチであるが，本研究ではNewman法をベースにし，Webページのリンクによるネットワーク構造だけでなく，Webページのコンテンツ やWebサイトのディレクトリ構造による重みを導入している．

2 トピックマップ

Webページ集合で表現されるトピックとそのトピックの関係を表現するのに適した手法として，ISO/IEC JTC1 SC34 WG3で策定されたトピックマップ（ISO/IEC 13250 Topic Maps）[6]がある．トピックマップとは，知識を記号化し，記号化された知識を関連する情報リソースに結びつけるための技術である．トピックマップは，Fig. 1に示すように，情報リソースによって表現される**トピック（topic）**，それらトピック間の**関連（association）**，トピックと関連する情報リソースを結びつける**出現（occurrence）**の3つの要素で構成される．トピックマップは自由度が高く，トピックや関連等をユーザが自由に定義することが可能であり，様々なトピックや関連を内包しているであろうWebに適していると考えられる．

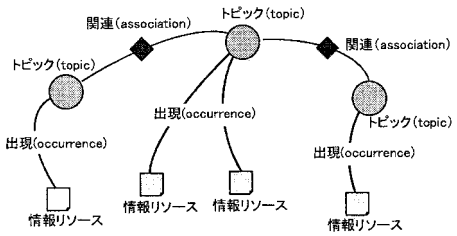

Fig. 1 トピックマップの概要

トピックマップの構文として用いられるのが，XML表記によるXTM（XML Topic Maps）[14]である．XTMを用いてトピックマップを表記するには，いくつかの情報項目や名前付き特性が必要となるが，全ての特性を自動的に抽出するのは非常に難しい．そのため，本研究では，トピック，トピック名，関連，出現の特性について抽出する．基本的には，Webページ集合をクラスタリングし，クラスタをトピック，クラスタ間のエッジを関連，クラスタを構成するWebページをそのクラスタで表現されるトピックの出現として抽出を行い，まずトピックマップの雛型を作成することを目指す．

3 Webページ集合からのトピックマップの抽出

3.1 概要

トピックマップは，Webページ集合に含まれているトピックだけでなく，それらのトピックの関連についても提示する必要がある．既存のcontents-basedなクラスタリング手法[7]のように，Webページに記述された内容の類似度を基にクラスタリングを行った場合には，構成されたクラスタが示すトピック同士の関連の強弱は確認できるが，トピック間にどのような関連が存在するのかという情報を獲得することはできない．一方，一般的なstructure-basedなクラスタリング手法は，ネットワークの構造に注目したアプローチであり，Webページの内容の類似度等は考慮されていない．

そこで，本研究ではstructure-basedなクラスタリング手法として良く知られるNewman法をベースにして，Webページのリンクによるネットワーク構造とWebページの内容の類似度を考慮したク

ラスタリング手法を用いて，トピックとトピックの関連を併せて抽出する．

3.1.1 Webページのリンク構造の利用

Web上に存在するページ間のリンクは基本的にページ作成者によって作成されているため，リンクされているページの内容が関連していると考えられる[9]．そのためWebページのリンク関係にはトピック間に存在する関係が内包されていると考えられる．ページ作成者によってページ間にリンクが張られていて，何らかの関係が存在していたとしても，ページ間の類似度の低い場合には，contents-basedのクラスタリングでは抽出することができない．

そこで，Webのリンク構造に内包されたトピック間の関係を抽出するために，リンクされているWebページのみをマージしてクラスタリングを行い，最終的にクラスタ間に残ったWebページのリンクからトピックの関連を抽出する．

3.1.2 ディレクトリ構造とWebサイト間のリンクの利用

リンクされたWebページ同士をマージしてクラスタを構成する際に，Webページを繋ぐリンクの強さを考慮することで，より密なクラスタを作成することを考える．リンクの強さは，Webページの特徴ベクトルの類似度を用いることで計算することも可能だが，ここではさらに，双方のWebページが配置されているディレクトリによるリンクの分類やディレクトリの距離から推定し，一般的な内容の類似度からは抽出することが難しい，Webページの関連を抽出する．

Parasite[13]では，リンク先のページがリンク元のページと同一サイト内に存在する場合にはディレクトリの階層が上位階層，同一階層，下位階層の3種類，同一サイトに存在しない場合の計4種類に分類することで，Webページ間のリンクの意味を推定するヒューリスティックを報告している．

本研究では，Webページ間のリンクを以下に示す3種類のタイプに分類し，Webページの配置されたディレクトリ距離と併せて，Webページの内容に依存せずに，Webページ間の関連の強さを推定する．ここで，ディレクトリ距離を，Webページとディレクトリをノードとした木構造で，(任意の2つのWebページ間の最短パス上にあるディレクトリの数 −1) とする．

upward/downward リンク元ページとリンク先ページが同一サイト内に存在し，リンク先ページがリンク元ページが含まれるディレクトリの上位階層または下位階層のディレクトリに含まれている場合のリンクである．

crosswise リンク元ページとリンク先ページが同一サイトに存在し，リンク先ページがリンク元ページのディレクトリの上位・下位階層のディレクトリに含まれない場合のリンク関係である．

outward リンク元ページとリンク先ページが同一サイトに存在しない場合のリンク関係である．

リンクの種類を以上の3つに分類し，ディレクトリ距離を次の方針で重み付けを行う．

- ディレクトリの構造においてcrosswiseのリンクをupward/downwardのリンクよりも優先する重み付けを行う．

  基本的にWebサイトのディレクトリ構造はページ作者によって構成されているため，ディレクトリ毎に関連する内容のWebページがまとめられていると考えられるためである．また，upward/downwardでリンクされているWebページは，上位概念，下位概念の内容が記述され，crosswiseの関係にあるWebページの内容よりも関連が薄いと考えられるからである．

- upward/downwardのリンクよりも，outwardのリンクを優先させる．

  outwardのリンクはページ作成者が，明示的にWebページのトピックに関連する他の情報源へのリンクを張っていることを意味し，基本的には関連するトピックが記述されていると考えられる．これに対してupward/downwardについては，そのページのトピック及びその他のトピックを包括的に含んだ上位トピックあるいは特殊化された下位トピックへのリンクであるため，outwardのリンクによって示されるトピックよりも関連が薄いと考えられるからである．

- ディレクトリ距離が小さいリンクを優先する重み付けを行う．

  これは，同一ディレクトリ内のページのトピックは関連度が非常に高く，同一サイト内でディレクトリ距離が大きくなるにつれ，関連度は低くなると考えられるからである．

以上のように重み付けを行い，リンクされたWebページやクラスタをマージしてクラスタを構成し密なクラスタを作成する．また，それらのクラスタ間

に残っているエッジを抽出することで，クラスタによって示されるトピック間の関連を抽出することが可能となる．重み付けの詳細を次節で述べる．

3.2 クラスタリング

本論文で提案する手法は，Newman 法をベースにして，Web ページのコンテンツの類似度や Web ページ間のリンクの種類や Web ページが含まれるディレクトリ間の距離を考慮した重み付けを加えたものである．Newman 法[10] は structure-based なクラスタリング手法の一手法であり，各クラスタの結合の強さを表す指標である Modularity Q を最大にするように新たなクラスタを構成していく階層的クラスタリング手法である．Q が高い状態とは，各クラスタ内のノード同士のリンクが密であり，クラスタリングが適切に行われている状態である．Q は以下の式で求められる．

$$Q = \sum_i (e_{ii} - a_i^2) \tag{1}$$

$$\Delta Q_{ij} = 2(e_{ij} - a_i a_j) \tag{2}$$

ここで，e_{ij} はネットワークのエッジの総数におけるクラスタ i とクラスタ j 間のエッジの数の割合であり，a_i はクラスタ i 内のノードに接続しているエッジの本数の割合である．Q は，各クラスタに含まれるエッジの総数の割合とその期待値との差である．ノードが属しているクラスタが考慮されない，エッジがランダムに張られたネットワークにおいては，$e_{ij} = a_i a_j$ であり，$Q = 0$ となる．エッジがランダムに張られたネットワークの期待値よりも各クラスタの結合が強い場合には，$Q > 0$ となる．$Q = 1$ の時が最大であり，最も強いクラスタ構造となる．

Newman 法には基本的にネットワークの構造のみを考慮した手法であり，Web ページのコンテンツについては何ら考慮されていない．そのため，本研究では Web ページのコンテンツによる類似度と前節で示した方針に従った重み付けの 2 つを用いたリンクの重みを求め，Newman 法に適用した．

3.2.1 Web ページ間のリンクの重み

Web ページ間のリンクの重みは，Web ページのコンテンツによる特徴ベクトルの類似度 $s(p,q)$ と，Web ページ間のリンクの種類とディレクトリ間の距離による重み $w(p,q)$ の重み付き線形和で求める．Web ページ p，q の類似度 $S(p,q)$ は次式で求められる．

$$S(p,q) = \alpha s(p,q) + (1-\alpha) w(p,q) \tag{3}$$

コンテンツによる類似度 Web ページの特徴ベクトル v_i は HTML ファイルから抽出した単語とその出現頻度に対して TF・IDF 法[12] による重み付けを用いることで得られ $v_i = (w_{i1}, w_{i2}, ..., w_{in})$ で表される．ここで w_{ij} とは Web ページ i における単語 j の重みである．Web ページ p，q の類似度 $s(p,q)$ は下式のように余弦で定義される．

$$s(p,q) = \frac{v_p \cdot v_q}{\| v_p \| \| v_q \|} \tag{4}$$

リンクの種類による重み付け リンクの種類による重み付けは，Web ページ間のリンクの種類と Web ページが含まれるディレクトリ距離から求める．Web ページ p と q のリンクに対する重み $w(p,q)$ は次式で表される．

$$w(p,q) = \begin{cases} \dfrac{0.25}{d+1} C_{ld} & \text{(upward/downward)} \\ \dfrac{0.5}{d+1} C_{ld} & \text{(crosswise)} \\ 0.4\, C_{ld} & \text{(outward, } \tau_s \le s(p,q)\text{)} \\ 0 & \text{(outward, } s(p,q) < \tau_s\text{)} \end{cases} \tag{5}$$

ここで C_{ld} はリンクの方向によって決定され，ページ間で双方向にリンクしている場合には 2，片方のページからのみリンクしている場合は 1 である．d は 3.1.2 で示した Web ページのディレクトリ距離である．τ_s は全ての outward タイプのリンクの関係にある Web ページ間のコンテンツの類似度の平均値とする．

これらの重みは，3.1.2 で示した方針で決定されており，同一ディレクトリのリンクが最も優先され，次いで外部リンク，上位下位階層へのリンクと続き，以降はディレクトリ距離に依存した重み付けとなる．

3.2.2 アルゴリズム

以下の手順でクラスタリングを行う．

1. Web ページ集合のリンク構造をネットワークの初期状態とする．1 つの Web ページを 1 つのクラスタと考え，Web ページ間のリンクをエッジとする．

2. Web ページ集合に対して，Web ページ間に張られた全てのリンクについて，3.2.1 で示した重みを計算し，対応する各エッジにその重みを適用する．

3. エッジの張られているクラスタの組み合わせの中から，クラスタを構成した際に最大の ΔQ_{ij} をとる組み合わせを選択する．もし最大の ΔQ_{ij} が負の値であれば，クラスタリングを終了する．

4. 選択されたクラスタをマージして新たなクラスタを作成し，そのクラスタに関連する e_{ij}，a_i を再計算して，(3) へ．

3.3 トピックマップの抽出と可視化

Web ページ集合をクラスタリングし，構成されたクラスタ，エッジ，クラスタを構成する Web ページからトピックマップの各要素を抽出する．

トピックの抽出 クラスタに含まれる Web ページによって表現される主題をトピックとする．つまり，1 つのクラスタが 1 つのトピックを表しているとする．トピックの名前を適切にかつ自動的に決定することは非常に困難であるため，クラスタの概念ベクトルから選択したいくつかの特徴語や，クラスタを構成する Web ページのタイトル，URL，コンテンツ等をユーザに提示することで，手動でトピックの適切な名前を決定してもらう．

関連の抽出 :トピック間の関連は，クラスタ間に存在するエッジとして抽出する．エッジは，最後まで残っている Web ページ間のリンクによって構成されている．エッジが残っているということは，エッジの両端のクラスタによって表現されるトピックの間に何らかの関係があると考えられる．しかし，トピックと同様に，トピック間にどのような関連があるのかを自動的に判断することは難しいため，関連のある 2 つのトピックや，エッジを構成している Web ページ間のリンクのアンカーテキストやその周辺の文字列，アンカー URL，Web ページのコンテンツ等の情報を提示し，ユーザにトピック間の関連の説明を決定してもらうこととする．

出現の抽出 :クラスタを構成する全ての Web ページは，これらのページによって表現されるトピックに関連づけられた情報リソースであり，この関連づけを出現として抽出する．

以上のように抽出した，トピック，関連，出現を用いてトピックマップの雛型を作成する．トピックマップを二次元平面上にグラフとして可視化する際には，ばねモデルを用いてノードの配置を決定する．本研究では，グラフ可視化のツールの一つである graphviz[2]を用いた．

4 評価実験

前章で示した提案手法によりトピックマップを抽出して，そのトピックマップの妥当性を検証し，提案手法の評価をするために，被験者による評価実験を行った．提案手法は，structure-based のクラスタリング手法である Newman 法をベースにして，Web ページのコンテンツの類似度や，Web ページ間のリンクの種類，Web ページが配置されているディレクトリ構造を考慮した重み付けを用いた手法である．そのため，ベースとなった Newman 法との比較を行うことで，重み付けを加えることの効果を検証することができる．同一の Web ページ集合から 2 つの手法で抽出したトピックマップについて，トピックマップのトピックと関連の評価をしてもらった．16 名の被験者（大学院生 12 名，社会人 4 名）により，実験を行った．各被験者には，Web ページ集合を収集するために用いるブラウジング履歴を 2 つ提供してもらい，各手法でトピックマップを抽出したものを提示し，計 4 つのトピックマップについて評価をしてもらった．なお，4 つのトピックマップを評価する順序はランダムに決定した．

4.1 Web ページ集合の収集

本実験では，トピックマップを抽出する Web ページ集合として，被験者がブラウジングした Web ページからリンクをたどり収集した Web ページの集合を用いた．この Web ページ集合は，基本的に被験者が閲覧した Web ページに関係する Web ページで構成されると考えられるため，被験者は抽出したトピックマップの妥当性を判定しやすく，また関係する情報を発見できるなどの効果が期待できる．

被験者から Web 上でサーチング（ここでは，明確な目標を持ったブラウジングのこととする）を行った履歴から，任意の連続する 5 ページ分を選択してもらった．ただし，履歴の途中でブラウザのアドレスバーから URL を直接入力したり，ブックマークを使用したものは不可とした．これは，提案手法が Web ページのリンク関係によって重み付けを計算する手法であるので，各 Web ページが必ず他の Web ページとリンクされている必要があり，リンクされずに独立した Web ページが存在するのは適当ではないためである．

この 5 ページの履歴ページを 1 セットの履歴データとした．これらの履歴データに含まれる Web ペー

[2] http://www.graphviz.org/

ジから，リンクを4つたどることで到達可能なWebページを収集し，これをトピックマップを抽出するためのWebページ集合とした．また，各Webページからたどるリンクの数は3つに制限し，ランダムに選択した．収集した各Webページ集合のWebページ数は約600であった．これらのWebページからテキストデータのみを抜き出し，形態素解析器であるMeCab[3]を用いて品詞と未知語を抽出し，TF・IDF法により各語の重み付けを行うことで，特徴ベクトルを作成した．各Webページの特徴ベクトルの次元数の平均は約270であった．

4.2 トピックマップの抽出

各被験者から提供してもらった履歴データからWebページ集合を収集し，提案手法，Newman法を用いてトピックマップを抽出した．3.2.1で示した提案手法の式(3)のパラメタは$\alpha = 0.5$とした．トピックマップは2次元グラフで可視化され，トピックはノードで，関連はトピックを結ぶエッジで表現される．また，被験者が実際に閲覧したWebページの含まれるトピックは外側にリングがかかったノードで表示される．トピックのノードの面積は，出現によってトピックに関連づけれられているWebページの数に比例する．各トピック，関連を表す円状のアイコン，エッジ状の菱形のアイコンをクリックすると情報表示・評価用ウィンドウが表示される．被験者はこのウィンドウ上に表示される情報を参考に，トピックと関連の評価を行った．また，評価の際には全てのトピックの評価が終了した後に，関連の評価を行った．

4.3 トピックマップの評価

トピックマップが抽出された状態では，トピック，関連共に名前がつけられていない．被験者には，この状態のトピックマップを提示し，トピックと関連の適切な名前，説明を決定してもらうことで，トピックマップの評価を行った．

4.3.1 トピックの評価

トピックは関連づけられたWebページ集合によって構成されているため，これらのWebページによって表現される主題や概念の名称がトピックの名前となる．被験者には，Webページ集合の主題は何であるかを判断してもらい，名前をつけてもらった．各トピックの評価にあたり，次の情報を被験者に提示した．1）トピックを構成するWebページのタイトルとURL，必要であればそのWebページの内容．2）クラスタの概念ベクトルから求めたトピックの特徴語である3つの単語．

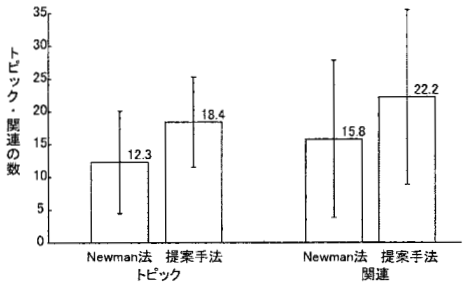

Fig. 4 トピック・関連の数の平均

トピックを構成するWebページ集合に，異なるトピックを表現するWebページ集合が混在している場合には，最も多くのWebページによって表現される主題の名前を選択してもらった．

そして，名前をつけることのできたトピックは“妥当なトピック”，そうでないものは“妥当でないトピック”と判定した．

4.3.2 関連の評価

関連は，2つのトピックを構成するWebページ間にリンクが張られている場合に，それらのトピックの間には何らかの関係があると判断し，構成されている．被験者には，関連が示す2つのトピックの間に説明可能な関係の有無を判定してもらい，説明可能であればその関係の説明をしてもらった．各関連の評価にあたり，次の情報を被験者に提示した．1）2つのトピックの名前，必要であれば各トピックの情報．2）2つのトピックのWebページ間に張られたリンクのリンク元，リンク先のWebページのタイトル，URL，アンカーテキスト．

トピックの場合と同様に，関連が説明できたものは“妥当な関連”，説明できないものは“妥当でない関連”とした．

4.4 実験結果

提案手法は，Webページ集合からのトピックマップ抽出を支援するための手法であり，抽出したトピックマップの雛型をユーザに提示し，最終的な調整はユーザが行うことを前提としている．そのため，ユーザが作業を行うにあたって，トピックマップの雛型はトピック，関連が網羅的かつ適切に抽出されていることが理想であり，それが最もユーザがコストを抑えてトピックマップの作成することができる状態であると考えられる．

そこで，できるだけ適切な説明がされ妥当と判定

[3] http://mecab.sourceforge.net/

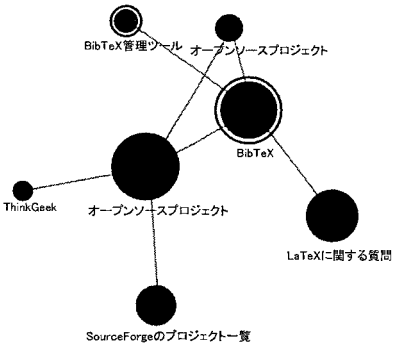

Fig. 2　トピックマップ（Newman 法）

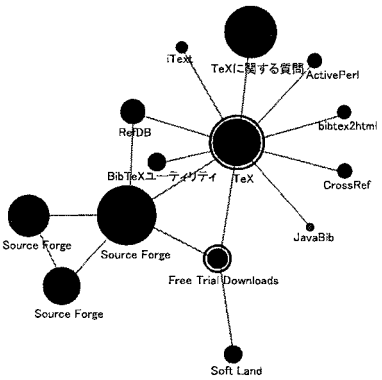

Fig. 3　トピックマップ（提案手法）

されたトピック，関連の数について評価した．Fig. 4 に実験結果を示す．Fig. 4 は，各手法の妥当なトピックと関連の数の平均を示している．Newman 法によって抽出されて，かつ適切な名前を付けられた妥当なトピックの数の平均は 12.3 個（標準偏差は 7.84）提案手法によるものは 18.4 個（6.89）である．また，関連の数の平均については，Newman 法は 15.8 個（12），提案手法では 22.2 個（13.22）である．

Fig. 4 より明らかなように，全般的に，Newman 法よりも提案手法の法が，より多くの適切なトピック，関連を抽出できているのがわかる．検定を行ったところ，トピックの数の平均に関しては，2 群のデータに正規性が確認できないためウィルコクソン符号付順位和検定により，有意差（p=0.000012, α=0.05）が認められた．また，関連の数の平均については，対応のある t 検定を行った結果，有意差（p=0.002，α=0.05）が確認できた．

5　考察

5.1　実験結果の分析

実験結果より，提案手法は Newman 法に比べより多くの数のトピックや関連を抽出することが可能であり，雛型として十分なトピックマップを抽出できることがわかった．最終的にユーザが処理を行う時点での作業については，既に提示されているトピックや関連からノイズを取り除く作業の方が，不足しているトピック等を新たに考え，付け足す作業よりもコストが低いと考えられる．特に，本研究で目指すユーザにとって既知または未見のトピックが含まれるトピックマップを作成する場合，未見のトピックを新たに探し出し付け足す作業は，非常にコストが高いためである．

提案手法によるトピックマップは，より多くのトピックや関連を提示することが可能であり，それらの中からユーザは有用なものを発見できると期待できる．提案手法が Newman 法と異なるのは，コンテンツの類似度や Web ページの配置されている Web サイト及びディレクトリの分類を考慮した重み付けを利用している点である．以下に実際に提案手法によって抽出されたトピックマップのうち，改善の結果興味深いトピックや関連を発見できるものが含まれている例を示す．

Fig. 2，Fig. 3 に Newman 法，提案手法によって抽出されたトピックマップを示す．この例は被験者が BibTeX に関連するツールについて検索した履歴から抽出したトピックマップである．提案手法によるトピックマップでは，Newman 法によるものに比べて，トピックの数が多く，内容的にも細分化されていることがわかる．Fig. 3 の右中央の「TeX」の周りにある「BibTeX ユーティリティ」，「bibtex2html」，「JavaBib」等は，Fig. 2 のトピックマップでは，右中央のトピック「BibTeX」に含まれてしまっている．

これらのトピックの構成する Web ページは URL の Web サイトが異なっており，提案手法ではその点が考慮されているため，個々のトピックとして抽出することができている．そのため，提案手法によるトピックマップからは，「TeX」関連のツールとして「bibtex2html」等が存在することが見て取れる．また，これらのツールの Web ページは被験者にとって未見のページであり，関連する情報を新たに発見できると期待される．

5.2 課題

Webページ集合にWiki系のページを多く含んでいるケースでは，トピックや関連の抽出が適切に行われていない例もあった．これらのWikiページは，Webページ間のリンクが密に行われていたり，動的に生成されるURLであるといった特徴がある．提案手法では，基本的にWebページ間のリンクには何らかの関係が含まれているという前提を置いているため，リンクの選別，削除等の前処理は行っていない．また，Webサイトやディレクトリ構造を考慮した重み付けをしているため，CGIファイルとページタイトル等から生成されるURLには適応することが難しい．そのため，今後も増え続けるであろうWikiやblogなどのWebページに対応する必要がある．また，Newman法に導入したリンクの重みを始めとする各種パラメタの調整も検討する必要があり，今後の課題である．

6 まとめ

本論文では，Webページのコンテンツの類似度，Webページ間のリンクの種類，ディレクトリ構造を考慮した重み付けを用いたクラスタリング手法を用いて，トピックマップの雛形を抽出する方法を提案した．重み付けの効果を検証するために他手法との比較実験を行い，それぞれのトピックマップのトピックや関連の数について検証を行った．実験結果より，提案手法は多くのトピックや関連を抽出でき，雛形として十分なトピックマップを抽出可能であることが示された．今後の課題としては，ユーザとのインタラクションを通じてトピックマップを完成させるための情報提供の仕方や枠組み等の検討があげられる．

参考文献

1) Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J.: Graph structure in the web: experiments and models, in *5th WWW Conference* (2000)

2) Flake, G. W., S., L., and Giles, C. L.: Efficient identification of Web communities, in *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 150–160, New York, NY, USA (2000)

3) Girvan, M. and Newman, M. E. J.: Community structure in social and biological networks, `http://arxiv.org/abs/cond-mat/0112110/` (2001)

4) Gulli, A. and Signorini, A.: The indexable web is more than 11.5 billion pages, in *Special interest tracks and posters of the 14th WWW conference*, pp. 902–903, New York, NY, USA (2005)

5) GVU's WWW Surveying Team: GVU's 10th WWW User Survey: Problem Using the Web, `http://www.gvu.gatech.edu/user_surveys/` (1998)

6) International Standard Organization: ISO/IEC 13250 Topic Maps: Information Tecknology Document Description and Markup Language (2000)

7) Jain, A. K. and Dubes, R. C.: *Algorithms for clustering data*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1988)

8) Kerk, R. and Groschupf, S.: How to Create Topic Maps, `http://www.media-style.com/gfx/assets/HowtoCreateTopicMaps.pdf` (2003)

9) Menczer, F.: Lexical and semantic clustering by web links, *Journal of American Society Information Science and Technology*, Vol. 55, No. 14, pp. 1261–1269 (2004)

10) Newman, M. E. J.: Fast algorithm for detecting community structure in networks, *Physical Review E*, Vol. 69, 066133 (2004)

11) Reynolds, J. and Kimber, W. E.: Topic Map Authoring With Reusable Ontologies and Autoated Knowledge Mining, in *XML 2002 Conference* (2002)

12) Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523 (1988)

13) Spertus, E.: ParaSite: mining structural information on the Web, in *Selected papers from the 6th WWW conference*, pp. 1205–1215 (1997)

14) TopicMaps.Org: XML Topic Maps (XTM) 1.0, `http://www.topicmaps.org/xtm/1.0/` (2001)