3A2-6

Web閲覧履歴からのTopic Mapの抽出の支援

Extracting Topic Maps from Web browsing histories

間瀬心博[*1] 山田誠二[*2]
Motohiro Mase Seiji Yamada

[*1]東京工業大学 大学院 [*2]国立情報学研究所
Tokyo Institute of Technology National Institute of Informatics

In this paper, we propose a method of clustering to extract Topic Maps from the Web browsing history. The our method is based on the conventional agglomerative clustering with the constraint of the Web structure and the weight of link relation. Topic Maps show 2D-visualized overview graph of the browsing history, and the relations between the topics that extracted from the pages around the history pages. Using the Web browsing history, we experimentally evaluate the extracted Topic Maps.

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，ユーザにとって非常に有用であり重要なものとなっている．Webページの総量は年々増加し続けており，2005年1月現在で115億を超えるとされている[1]．そのため，ユーザがWeb上から目的の情報を探し出すのは非常に困難なタスクであるが，多くの場合Google，Yahoo!等に代表される検索エンジンを利用することでユーザの負担は大幅に軽減されている．しかし，日常的にWebを利用するユーザが抱える問題は他にも存在する．10th GVU WWW User Survey[2]によると，「一度訪れたページに再度訪れることができない」ことを調査対象のユーザ3,291人のうち547人が重要な問題であると考えており，また同様に「自分が獲得した情報を効率的に整理し，まとめることができない」と908人が考えている．前者の問題はWebブラウザのブラウジング履歴を閲覧・検索することで解決できるが，基本的にはWebページがリスト表示されるものであるため使い勝手が良いものではない．後者の問題も，ブラウザのブックマーク機能を利用してWebページを内容別に分類したり，Webページ自体をローカルに保存し整理することも可能であるが利便性が高いとは言えない．

これらの問題を解決する一つの手段として，ユーザのWebブラウジング履歴を元に収集したWebページから，ページに含まれるトピックとトピック間の関係を抽出し2次元グラフ上に可視化してユーザに提示することを考える．ユーザはブラウジングしたページやその周辺に存在するページをトピックごとに分類された状態で確認することができ，抽出されたトピック間の関係も見ることができる．そのため，ユーザは獲得した知識を手動で整理せずにすみ，再度閲覧したいページをそのページに記述されたトピックをクエリとして探すことが可能になると期待できる．本稿ではWebブラウジング履歴からトピックとトピック間の関係を抽出するために，従来の集積的な階層的クラスタリングを基本にしたWebの構造的な制約とWebページ間のリンク関係による重み付けを考慮したクラスタリング手法を提案する．ユーザのブラウジング履歴を可視化する研究は数多く行われている．ブラウジング履歴を2・3次元グラフ表示する研究[4]や，さらにWebページのサムネイルを同時に表示する研究[3]がある．また，タスクやセッションごとに分類して履歴を可視化する研究[5]も行われている．本研究はWebブラウジング履歴の可視化に加えWebページから抽出したトピックやトピック間の関係を提示することを試みる．

2. 提案手法

2.1 トピックマップ

Webページの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）がある．トピックマップは，情報資源が表すトピックとトピック間の関係を，情報資源とは独立に表現しトピックに関連する情報資源に対してリンクを張ることで表現する手法である．このトピックマップのシンタックスとしてよく用いられるのがXML表記によるXTM（XML Topic Maps）である．本稿では，Webブラウジング履歴から抽出したトピックやトピック間の関係をトピックマップの形式で表現可能な形式で抽出することを試みる．

2.2 アプローチ

トピックマップはユーザによるWebブラウジングの履歴ページやその周辺に存在するWebページにどのようなトピックが含まれているかだけではなく，トピック間に存在する関係についても提示する必要がある．Webページに記述された内容についての類似度を元にクラスタを構成した場合には，クラスタが示すトピック同士の関連の強弱については確認することはできるが，トピック間にどのような関係が存在するのかという情報は得ることができない．このようなトピック間の関係の種類も同時に抽出するためにWebページ間のリンクが持つ特徴を利用することを考える．

Webのリンク構造：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が双方のページに関連があると判断して張っている．そのためWebページのリンク関係はトピック間に存在する関係を内包していると考えられる．ユーザによってリンクが張られてもページ間の類似度が低い関係は，ページ内容の類似度によるクラスタリングでは抽出することはできない．そこで，このようなWeb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，リンク構造を制約としたクラスタリングを考える．リンク構造による制約とはクラスタリングの際にリンク関係にあるクラスタ同士のみをマージするということである．

リンクの種類：ページ間のリンク関係の種類を，双方のページが配置されているディレクトリの関係から推定することを考える．本研究では以下の3種に分類する．1) *upward/downward*：リンク元ページとリンク先のページが同一サイトに内に存在し，リンク先のページが上位階層または下位階層のディレクト

リに含まれるリンク関係である．下位階層に含まれるページのトピックは上位階層に含まれるページのトピックのサブトピックと考える．2) *crosswise*：リンク先のページとリンク元のページが同一サイトに存在し，リンク先のページがリンク元のページの上位・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．階層差が大きくなるとトピック間の類似度は下がると考える．3) *outward*：リンク先のページとリンク元の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．基本的には双方のトピックに関連があると考える．

以上の 3 種にリンクを分類し，次の方針で重み付けを行う．1）同一ディレクトリのページのトピックは関連が高くなり，ディレクトリの階層差が大きくなるにつれ関連は低くなると考え，階層差の少ないリンクを優先した重み付けをする．2）ディレクトリの階層構造において兄弟関係にあるリンクを親子関係にあるリンクよりも優先する．

2.3 クラスタリング

本稿で提案する手法が従来の基本的な階層的クラスタリング手法と異なるのは以下の点である．1）本来の Web 構造においてリンク関係にあるページ，それらのページを含むクラスタ同士でのみをマージして新たなクラスタを構成する．2）クラスタ間の類似度を計算する際にクラスタの特徴ベクトルのみを用いるのではなく，クラスタに含まれるページ間のリンクの種類やディレクトリの階層差を考慮した重み付けを行う．3）新たにマージしたクラスタの凝集性が設定した閾値を下回った場合に，そのクラスタを今後マージするクラスタ候補から除外する．これらの点について従来の階層的クラスタリング手法を拡張したものが提案手法である．

3. 実験

提案手法を用いてブラウジング履歴を元に収集した Web ページ集合からトピックマップを抽出する実験を行った．トピックマップを抽出するための Web ページ集合は履歴ページからインバウンド・アウトバウンドリンクを問わず 4 ステップで到達できるページで構成される．また，各ページから展開するリンク数は 3 とした．Web ページ集合から提案手法を用いて抽出したトピックとトピックの関係をばねモデルを用いて 2 次元グラフ上に描画した．図 1 に実際にトピックマップの抽出結果を示す．図中のノードは Web ページで構成されたクラスタであり，クラスタのラベルはトピックを表している．リングで囲まれたノードはブラウジングが行われた履歴ページが含まれていることを示す．

Meadow(UNIX 等で動作する Emacs を Windows に移植したエディタ) のフォント設定についてブラウジングを行った履歴からトピックマップを抽出した結果を図 1 に示す．図 1 より，Meadow・Emacs に関連するトピックのクラスタが実際に閲覧した Web ページが含まれるクラスタの近傍に配置されていることがわかる．中央上部にはフォントに関するクラスタが存在し，その左側に Adobe に関連するクラスタが配置されている．これは Adobe 社がフォント販売を行っているため，フォントについて記述された Web ページからリンクされていたものと考えられる．また，中央部には Cygwin(Windows 上に移植された UNIX ツール群) に関するクラスタがあり，Meadow や Emacs に関するクラスタとリンクされている．これらは，コンテンツの類似度は低いもののページ作成者によって「Windows 上で動作する UNIX ツール」というビューポイントから Meadow 関連のページからリンクが張られたことによって，抽出することができたと考えられる．このようなページ作成者のトピック

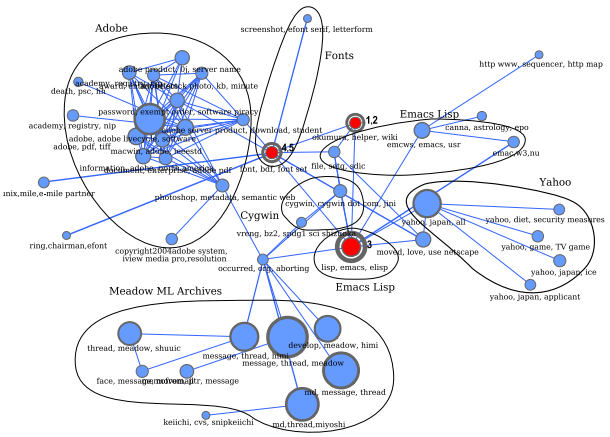

図 1: 実験 1：Meadow のフォント設定の検索

間の関連付けはノイズにもなりうるが，新たなビューポイントの発見として期待できる．

4. まとめ

本稿ではユーザの Web ブラウジング履歴とその周辺に存在する Web ページの集合からトピックマップを抽出する手法を提案した．提案手法を用いて実際のブラウジング履歴からトピックマップを抽出する実験を行い，抽出されたトピックマップからはいくつかの興味深いトピック間の関係を発見することができた．今後は，ユーザがトピックマップの閲覧，トピックやトピック間の関係への任意のラベルの付与，トピックやトピック間の関係の変更・削除等の作業をできるインタフェースの実装が必要となる．

参考文献

[1] Gulli, A.,Signorini, A. The Indexable Web is More than 11.5 Billion Pages, Proceedings of the 15th WWW conference, 2005.

[2] GVU's WWW Surveying Team, GVU's 10th WWW User Survey: Problem Using the Web, Gerogia Tech, Atlanta, GA, 1998.

[3] Gandhi, R., Kumar, G., Bederson, B. B., Shneiderman, B., Domain Name Based Visualiazation of Web Histories in a Zoomable User Interface, Proceedings of WebVis2000, pp. 591-598, 2000

[4] Cugini and J. Scholtz, VISVIP: 3D Visualization of Paths through Web Sites, Proceedings of the International Workshop on Web-Based Information Visualization (WebVis'99), pp. 259-263, 1999.

[5] Matthias, M., Benjamin, B, B., Browsing Icons: A Task-Based Approach for a Visual Web History, HCIL-200119, CS-TR-4308, UMIACS-TR-2001-85, HCI Lab, University of Maryland, Maryland, USA, 2001.