3G1-02

Web における話題の時間変化の提示

A Visual Representation as Time-Change of Topics on the Web

森 幹彦[*1] 山田 誠二[*2]
Mikihiko Mori Seiji Yamada

[*1]京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

[*2]国立情報学研究所
National Institute of Informatics

1. はじめに

World Wide Web（以下，Web と呼ぶ）では現在，個人から企業までの様々な人々や組織が情報提供やコミュニケーションの場として Web サイトを解説している．例えば，実社会の情勢を知らせるオンラインニュース，ノウハウを提供したり交換するサイトなどがある．近年では，筆者の考えや感じたことや筆者の興味をもったものを紹介する，日記や blog と呼ばれるサイトも公開されるようになり，急激に増え続けている．

一方，提供される情報も膨大になったため，利用者にとっては必要な情報の抽出が難しくなった．この解決法として様々な Web 検索エンジンが提供され，広く利用されている．例えば，Google では PageRank[2] と呼ばれる Web ページ間のリンクを考慮した Web ページの順位付けを行うことで利用者の直感に近い順位で検索結果を提供している．また，検索結果を適合順に並べるだけではなく，Web ページ間の類似性で分類する Vivisimo[*1]や，検索を絞り込むための検索質問例と検索結果数を提示する WiseNut[*2]のような検索エンジンも登場していて，検索結果の内容把握の向上に貢献している．

Web 検索エンジンによって，必要な情報が掲載されている個々の Web ページは容易に抽出が可能になった．しかし，互いに関係し合う Web ページ群を順序立てて調べたいときに，現在の Web 検索エンジンは十分でない．例えば，2003 年の鳥インフルエンザ問題の経緯を調べたいとする．Web 検索エンジンを用いる場合なら「鳥インフルエンザ 2003」などのように検索質問を入力することになる．検索結果は，適当に順位付けされているか，さらに適当に分類されているかで提示される．しかし，このような検索結果の提示では，鳥インフルエンザの時間的な遷移を十分に考慮しているとは言えない．そのため，例えば上から順番に閲覧すると，利用者は時系列を行き来しながら情報を取得することになり，利用者に情報整理能力を要求することになる．また，どの情報が最新であるかを判断することは難しい．

本研究は，Web ページに記述されている事件などの出来事を抽出し，情報の利用者の観点で分類して時間順序と情報間の関係の表現を主とした情報の提示法を検討している．本稿では，Web ページから出来事を抽出する枠組みを提案する．また，抽出した出来事を利用する場合の表示法について検討する．

連絡先: 森幹彦, 京都大学学術情報メディアセンター, 〒606-8501
京都市左京区吉田二本松町

∗1 http://www.vivisimo.com/

∗2 http://www.wisenut.com/

2. 時間依存した情報の抽出

事件などの出来事が起きたことは，Web ページの著者に認識され，その Web ページに記述されるきっかけとなる．このような認識された出来事を本研究では，イベントと呼ぶことにする．イベントが起きてから実際にそれが著者によって記述される．この記述をイベント記述と呼ぶことにする．あるイベントについて，イベント記述が作成されるまでには，時間差が存在する．

イベントとイベント記述および Web ページの関係を図 1 に表す．図 1 において，横軸は時間経過である．縦の点線はイベントの起きたことを表し，小さい丸印はイベント記述を表す．イベント記述から左方向に延びた点線は，そのイベント記述がどのイベントに関連しているかを示すものである．同じ高さに描かれているイベント記述は同一筆者に書かれたものとしている．イベントが起きてから実際にイベント記述が書かれるまでの時間は様々である．また，場合によっては一つのイベントについて複数書かれることもある．

イベント記述は Web ページ中に一つまたは複数ある．多くの Web ページでは，複数のイベント記述によって構成されていると仮定している．図 1 では，破線の楕円で描かれている中のイベント記述が同一 Web ページ内であることを示している．

3. イベント記述

本研究の対象は，時間情報が提供された Web ページである．オンラインニュースの普及と blog の広まりにより，様々な時間情報を明示した Web ページが大量に存在する．これらの Web ページの特徴として，1 つのイベント記述を単位として，1 ページが 1 イベント記述として，内容を分けて記載されていることである．ただし，1 つのイベントを複数のページで記述することもあり，一方で類似した複数のイベントを 1 ページで記述することもある．

各イベント記述は, 時間情報を持っている．時間情報にも様々あり，イベントが起きた時刻，イベント記述の書かれた時刻，Web ページが更新された時刻などが考えられる．本研究では，イベントの起きた時刻を原則として利用することを考える．ただし実際には，イベント記述中に明確に時間情報が書かれないこともあり，その場合はイベント記述を書いた時刻はイベントの起きた時刻に最も近いと考えられるため，イベント記述で代用することも検討する．

イベントが Web ページの著者らによって注目すべきものであるほど，イベント記述は多くなる．また，多くの場合，イベント間には因果関係がある．イベント間の因果関係はイベント記述にも反映されると仮定する．そこで，因果関係のあるイベント群から生成されるイベント記述群を話題と呼ぶことにする．例えば，先に挙がった鳥インフルエンザの例において，個別の感

染例や感染後の対策などは，それぞれ別のイベント記述として扱うこととする．そして，特定の感染例からそれの対策までは 1 つの話題として扱う．1 つの話題には複数の著者のイベント記述が含まれることもある（図 1）．ただし，話題を抽出する段階で，情報の利用者の観点を反映するため，静的に分類して保持することは難しい．

直接因果関係のないイベントであっても，著者の発想において類似すると考えるイベントは，イベント記述間の関係が示される．すなわち，話題の間にも関係が存在することになる．関連する話題群をエピソードと呼ぶことにする．

紙媒体のニュース記事とは異なって Web 上の記事の特徴である，関連するイベント記述間の関係をハイパーリンクで繋ぐことが多い．ハイパーリンクで繋がれるイベント記述は同一の著者の中だけに留まらず，複数の著者間でも繋がれることが一般的である．これにより，イベント記述間の関係を計算するために起こった時刻や記述中に現れる単語の類似性だけでなく，ハイパーリンクを利用することが意味をなすと考えている．このハイパーリンクを用いて，特定のイベント記述では不明だったイベントの起きた時刻を他のイベント記述中にある時刻を用いることで可能にすることも考えられる．

4. 時系列表現の表示

特定の話題に注目して情報を取得したい場合，イベント記述の前後関係とイベント記述の類似性は，情報の利用者にとって重要である．このとき，直前直後のイベント記述だけではなく，長期的に見たイベント記述群が一覧できることが必要であると考える．したがって，ある話題に対してできるだけ長い期間のイベント記述を提示できることが重要である．多すぎる場合には時間を絞り込むことや，話題を絞り込むことで対応する．

また，類似した話題をエピソードとして表示されることも必要である．鳥インフルエンザの例では，1 つの感染例について発生から終了まで一覧できることも必要であるが，複数の感染例が実際には起こっていて，それらを提示することも必要である．また，場合によっては，事件が細分化されることもあるため，エピソードが分岐したり，個別の事件として考えられていたことが原因が 1 つであり収束することもあるため，エピソードが収束したり，ということを把握できる必要もある．さらに，鳥インフルエンザと重症急性呼吸器症候群の関係のように，ほとんど交わりはないが健康不安という観点で共通のエピソードは，関連性があるとして表示されることが望ましい．

これらの要件をそろえた表示法の一例を図 2 に示す．縦軸はエピソード記述の近さを表している．横軸が時間軸である．小さい丸印はイベント記述を表していて，1 つの話題はイベント記述の固まりになっている．エピソードが分岐していく過程は，イベント記述が離れて話題の固まりが分離していくことで表示される．イベント記述の描画だけであっても，利用者には直感的に話題の分布と関連性が把握できる．

5. 関連研究

Allan らは Topic Detection and Tracking (TDT) を提案している [1]．この提案では，ニュース記事から話題を抽出して追跡することを目的としている．一方，南野らは blog のような日時と記事本文という形式が続く Web ページに特化して，記事の検索や流行のキーワードの抽出などを提供する BlogWatcher を提案してシステムを公開している [3].

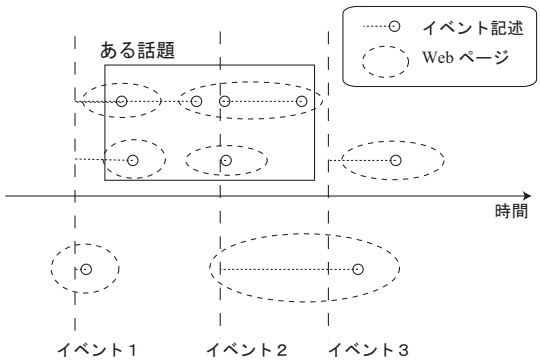

図 1　イベント，イベント記述，Web ページの関係

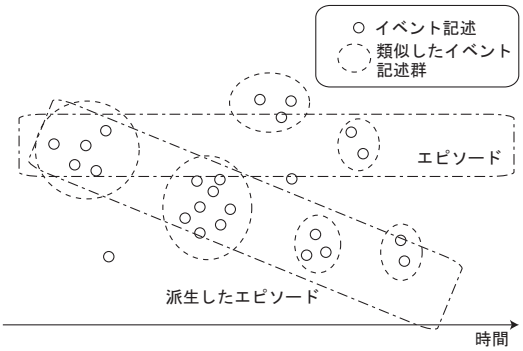

図 2　イベント記述の時間的前後関係と類似性の表示法

6. まとめ

本稿では，Web における時間情報に依存したイベントの記述表現について提案した．イベントとイベント記述の関係，イベント記述と Web ページの関係を定義した．

イベント記述を Web の利用者が閲覧するとき，出来事の前後関係を見ながら全体を閲覧するために必要な表現法について提案した．特定の話題に絞ってイベント記述を概観するとき，エピソードが見えてくる．エピソードは複数の系統があり，分岐したり収束したりすることを想定している．今後，この表現法をシステムに実装し，その有用性を確認したい．

参考文献

[1] Allan, J., Papka, R. and Lavrenko, V.: *On-line New Event Detection and Tracking*, Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 37–45 (1998).

[2] Brin, S. and Page, L.: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, vol. 30, pp. 107–117, (1998).

[3] Nanno, T., Suzuki, Y., Fujiki, T. and Okumura, M.: *Automatic Collection and Monitoring of Japanese Weblogs*, WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004).