7C-1

Web 文書における時間と文脈に依存するイベントの抽出

森 幹彦† 山田 誠二‡

京都大学学術情報メディアセンター† 国立情報学研究所‡

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．

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 ページ群を系統立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き，Web 検索エンジンを用いるなら「鳥インフルエンザ 2003」などのように検索質問を入力する．検索結果は，適当に順位付けされている場合や適当に分類されている場合がある．しかし，時間的な流れを考慮しておらず，上から順番に閲覧すると時系列を行き来することになる．

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，情報の利用者の文脈で分類して時間順序を中心とした情報の提示法を見当している．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．

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が認識されると，Web ページの著者の記載のきっかけになる．このように，認識された出来事をイベントと呼ぶことにする．イベントが起き，実際にそれが著者によって記述されると，イベント記述が生成される．本研究では，イベント記述が Web ページ中に一つまたは複数あり，それらの組み合わせで Web ページの内容が構成されていると仮定している．

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，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．縦軸が同じ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．

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，内容が分かれていることにある．また，これらの Web ページは Web の特徴を利用して，関連するイベント記述間の関係をリンクで繋ぐなどの紙媒体のニュース記事とは異なる性質を持っている．

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利用す

Time-dependent event extraction on contexts from Web pages

† Mikihiko Mori, Academic Center for Computing and Media Studies, Kyoto University

‡ Seiji Yamada, National Institute of Informatics

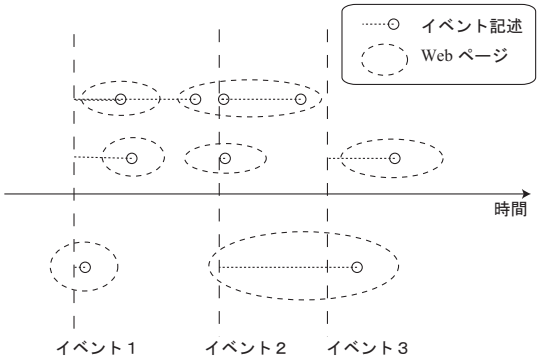

図 1　イベント，イベント記述，Web ページの関係

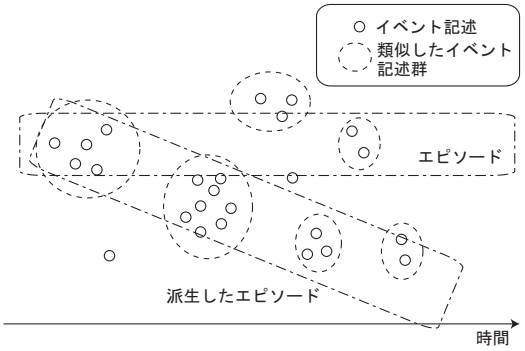

図 2　イベント記述の時間的前後関係と類似性の表示法

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る．そこで，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．

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．そこで，関連する話題群をエピソードと呼ぶことにする．

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や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，長期的に見えることが必要となる．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．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ため，1 つの話題は固まりになり，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いくことで表示される．エピソード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]．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．また，南野らは blog のような形式の記事に特化して，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].

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).