社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－146 (55)
2008／9／23

# Webからの将来情報の発見・分析にむけて

金澤 健介† Adam Jatowt†† 小山 聡†† 田中 克己††

† 京都大学工学部情報学科 〒606-8501 京都市左京区吉田本町
†† 京都大学情報学研究科社会情報学専攻 〒606-8501 京都市左京区吉田本町
E-mail: †kanazawa@dl.kuis.kyoto-u.ac.jp, ††{adam,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 本研究では, 将来の情報の検索に関する提案を行う. 現在, 過去情報を Web から検索するシステムはあるが, 将来のイベント情報を効率的にみつける方法はない. しかし, オンライン文書中には将来のイベントデータがある。また, 文書を分析することで過去のイベントの周期性が確認できる. 提案手法では, ニュース記事の時系列中に現れる周期的なパターンを分析して, 将来のイベントに関する情報を推測する. また, 将来情報の検索のために, 推測したイベントの確かさを測る. 本研究では, 将来情報の検索に関する様々な問題を論じて, 今後の研究についての計画を述べる.
キーワード 将来情報の検索, データマイニング, ニュースアーカイブ, 周期的イベント

# Towards Finding and Analyzing Future-related Information on the Web

Kensuke KANAZAWA†, Adam JATOWT††, Satoshi OYAMA††, and Katsumi TANAKA††

† Department of Informatics, Faculty of Enginieering, Kyoto University Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Department of Social Informatics, Graduate School of Informatics, Kyoto University Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †kanazawa@dl.kuis.kyoto-u.ac.jp, ††{adam,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

**Abstract** In this paper, we propose the framework for retrieving information about the future. Although some applications provide the capability to retrieve past information from digital collections or from the Web, currently there is no means to efficiently find the information about future events or in other words, to perform "future information retrieval". Yet, there is much data about forthcoming events contained in online documents or such information can be actually revealed through the periodical analysis of these documents. Our method is based on analyzing periodical patterns within temporal collections of news articles and on measuring the probability of predicted events. In the future information retrieval we need to measure the probability of predicted events.In the paper we also discuss various issues related to the future information retrieval and describe our plans for continuing this research.
**Key words** Future Information retrieval, Data Mining, News Archive, Periodical Events

## 1. はじめに

現在,Web 上には多くのイベントの情報がある. これらの情報には過去や現在のイベント情報だけでなく, これから起こる将来のイベントに関する情報も多く含まれている. 将来のイベントへの関心は高く,Web 上から将来のイベント情報を探したいという要望は強い. しかし, 現在将来のイベント情報を検索できるシステムで広く使われているものはなく, 既存の Web 検索エンジンを使って将来のイベント情報を発見することは難しい. また,Web 上に実際に書かれている将来情報には, 情報の種類によって偏りがある. 例えば Google News(注1)には将来の情報を記述している記事が 50 万件以上あるとされており [1], 将来情報をもつ記事が多く存在していることがわかる. しかし, その多くが数週間後, 数ヵ月後といった近い将来の情報であり, 数年後などのより遠い将来の情報は, より少ない. また, 一つの記事に情報が書かれているのみでは, その情報が確かかどうかはわからない.

そこで現在書かれていない情報や現在ある情報の確かさの支持を得るために, 過去のイベント情報から将来のイベントの推測を行う. それにより数年先などの遠い将来の書かれにくい情報

(注1)：http://news.google.com/

を得ることができ, ユーザがイベントの推測を行う助けになる. また, 推測によってもイベントの情報が得られたことにより, 記事に書いてあった将来イベントについての情報の確かさの支持が得られると考える. 将来情報を推測する際には要素として, 時間, 場所, 因果関係などが挙げられる. 本研究では時間に注目し, 周期的に起こるイベントの推測を行う. 繰り返しのないイベントや不規則に起こるイベントは本研究では扱わない. 周期的に起こるイベントとして, 例えば桜の開花, ソフトウェアのアップデート, 携帯電話の新機種発売時期等が挙げられる. 携帯電話の機種は発売時期が例年似た時期になっており, 次に発売されるのがいつごろか予測できる. このように周期的に起こるイベントの将来の生起時間を, 過去に起こったイベント群を取得し, それらの生起時間をみることで予測する. その際に, どれだけ周期的にイベントが起こっているか示す値として, 周期度を定義した. 過去のイベント群を取得する際には情報源として Web ページ・ニュースアーカイブ・メールなどが考えられるが, 本研究ではニュースアーカイブである Google News Archive(注2) を情報源とする. ニュースアーカイブを用いたのは, ある程度構造化されているため日付表現がとりやすく, ノイズが少ないためである.

本研究は将来イベントの情報の推測のための導入であり, ニュースの記事数に基づく分析・推測を簡単なモデルで行うのみである. しかし, 今後はテキストの分析などの手法で補完することで, より良い結果を得ることを目指す.

以下,2 章で関連研究を示し, 3 章で過去の同種のイベントの取得と判別, 4 章で将来イベントの予測, 5 章で評価実験, 6 章でテキスト情報の利用, 7 章で本研究のまとめについて述べる.

## 2. 関 連 研 究

将来情報の検索としては,Baeza-Yates の研究 [1] が挙げられる. Baeza-Yates は, 将来の時間を記述している記事をニュースから検索して, 結果をランキングして提示する概要を提案している. 将来情報の検索では, 記事中の書かれている時間を判別し, 現在の時間と比較している. また, ランキングはユーザから与えられたトピックとの適合性と信用度をもとに行っている. これはニュース上の明示的に将来情報を書かれている記事を検索しているのみであり, 本研究のようにイベント情報の抽出を行っていない. また, 過去の情報の分析や, 将来情報を得るための推測は行われていない.

Wuthrich らの研究 [2] は, 株価指標を予測しようとするものである. 過去の経済ニュースと過去の指標の変化を訓練データとして規則を生成し, その日のニュースと前日の指標を用いて, 指標の変化を予測している. ニュースからキーワードを抽出して, その特徴ベクトルを使って規則の重みを決めている. Choudhury らの研究 [4] では, ブログでのコミュニケーションの変化を分析し, サポートベクターマシンを用いて株価の変動との関連を求め, 変動を予測している. これら研究では, 株価というオブジェクトの値の変動の予測を行っており, その予測も株価に特化したものである. 本研究では, イベントの生起時間というオブジェクトの発生について予測を行っている.

## 3. 過去のイベントの取得と判別

イベントの予測を行うために, 予測したいイベントの過去の発生パターンを調べる. 予測したいイベントがユーザによってクエリで表され, そのクエリで検索を行う. 取得された文章には時間を付与する. 今回はニュースアーカイブを用いて検索を行い, 時間は記事の書かれた日付とした. 検索結果には, 複数のイベントが混在している. そこで, 過去のイベント群の生起時間を見るために, 一つ一つのイベントを判定して, 分割を行った. また, イベントの判定が行いやすいように, 正規化を行った.

### 3.1 イベントの取得

Google News Archive では上位にランキングされる記事の時間分布の偏りがみられるため, ユーザに入力されたクエリで検索を行う際に正規化をする必要がある. 時間的に記事の生成頻度が変わらないと考えられるストップワード『and』を含む記事の検索結果で、ランキングの偏りがあることを示す。図 1 は, クエリ『and』の上位 500 件を年ごとに集計して時系列に並べたものである. そして図 2 は,Google News Archive のクエリ『and』を含む文書の全ての集計結果である. Google News Archive では, 図 2 のような全ての結果の時系列グラフを検索結果のページに表示する. 図 1 では 1980 年代に結果がまったく得られていないが, 全ての結果記事においては 1980 年代にも他の年代と同程度の文書数がある. これからランキングに時間による偏りがあることがわかる. 図 3, 図 4 のように他のクエリの検索結果からも同様のことがいえる.

この偏りを解消するために, 検索結果の取得の際に期間ごとの正規化を行う. まず, ある期間 $T$ を定める. 今回は $T = 10$ 年とした. 検索する期間を, 長さ $T$ ごとの期間に区切り, その期間の結果の総記事数 $X_{T_i - T_{i+1}}$ を求める. 総取得件数が $N$ 件である場合に, 期間 $[T_i - T_{i+1}]$ の文書を $N * X_{T_i - T_{i+1}} / \sum_k X_{T_k - T_{k+1}}$ 件を取得する. これをすべての期間について行い, ほぼ $N$ 件の記事を取得する.

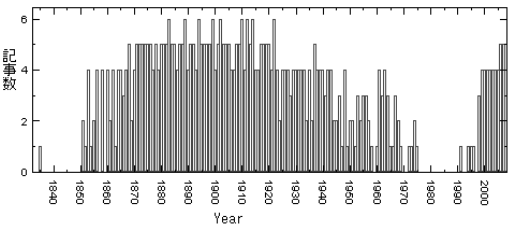


図 1 クエリ and の検索結果 (上位 500 件)

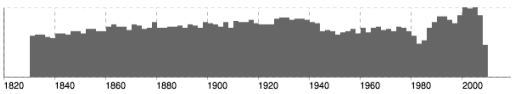


図 2 クエリ and の検索結果全ての時系列グラフ

(注2)：http://news.google.com/archivesearch/

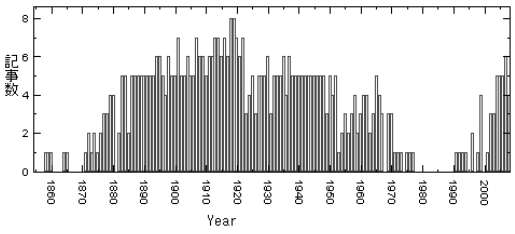


図 3　クエリ only の検索結果 (上位 500 件)

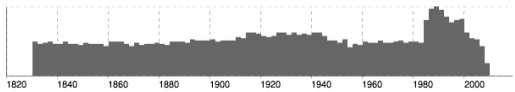

図 4　クエリ only の検索結果全ての時系列グラフ

### 3.2　イベントの判定

イベントの判定には, バースト検出を用いた. バースト検出では, 移動平均に基づく判定を行った. 1 単位時間ごとに重みを与えて, 重みの時間幅 $w$ での移動平均を $\mathrm{MA}_w$ とあらわす. 各時間の重みは, その時間に含まれる記事の数に基づいて計算を行う. そして, 以下の式でカットオフ値を定める.

$$\begin{aligned}\mathrm{cutoff}(time) &= \mathrm{MA}_w(time) + \alpha * var(\mathrm{MA}_w) \\ &\quad + \beta * average(\mathrm{MA}_w)\end{aligned} \tag{1}$$

Vlachos らの研究 [3] では, 時間一定のカットオフ値を用いているが, それを修正して本研究では時間によって可変のカットオフ値を定義している. 連続してこのカットオフ値より値が大きい区間を一つのイベントとして判断し, その時間にイベントが生起しているとする. 時系列の両端付近の移動平均を計算する際は, 端点以降に端点と同じ値が続くとして計算を行っている.

### 3.3　正　規　化

Google News Archive では, データベースに含まれている記事数に時間的な偏りがみられる. 図 2・図 4 にみられるように, ほとんどの検索結果では, 古い時期の記事ほど数が少なくなっている. それを解消するために, 各時間の重みの正規化を行う. 本研究では, ストップワードを用いて正規化を行った. ストップワードが含まれている文章の実際の生成頻度は時間によって偏りがないので, ストップワードの時間ごとの重みの偏りは記事数の偏りを表していると考えられる. そこで, ストップワードの時間ごとの重みが, 時間によって偏りがないように正規化を行う. 表 1 に示す 20 個のストップワードについて 500 件の検索結果を取得し, それらの平均を 1 期ごとにとり, その逆数を求める. 検索には 3.1 で述べた方法を用いた. 求めた値を各時間に含まれる記事の数に掛け, その時間の正規化された重みとする. これをバースト検出に用いる.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| a | about | among | an | and | as | at | but |
| by | for | from | in | it | of | on | only |
| or | that | the | to | | | | |

表 1　ストップワードリスト

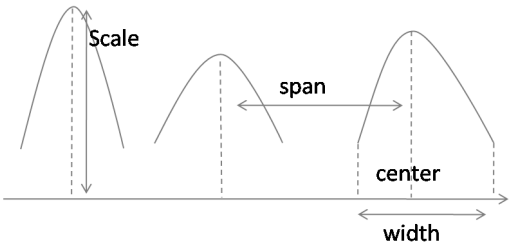


図 5　イベント群の特徴量

## 4.　将来イベントの予測

### 4.1　イベント群の特徴量

イベント群の特徴を表す値として,4 個の値を考える. イベント $e_k$ の期間を $\mathrm{Width}(e_k)$, 期間の最頻値を $\mathrm{Mode}(e_k)$, 規模を $\mathrm{Scale}(e_k)$ とする. ここで $\mathrm{Mode}(e_k)$ は, イベントのうちで重みの最も大きい時間とした. そして, イベント $e_k$ と $e_{k+1}$ との生起間隔を $\mathrm{Span}(k, k+1) = \mathrm{Mode}(e_{k+1}) - \mathrm{Mode}(e_k)$ と表す.

### 4.2　周期度の導入

周期度は, あるイベント群がどの程度周期的に起こっているかを示すものである. 周期的かどうかを決める要素として, イベントの数, イベントの生起間隔, イベントの期間, イベントの規模を考える. $N$ をイベントの数として,4 個の要素 $C = \{\mathrm{Num}, \mathrm{Span}, \mathrm{Width}, \mathrm{Scale}\}$ を正規化した値を以下のように定義する.

$$elem_{\mathrm{Num}} = \frac{1}{N}$$

$$elem_{\mathrm{Span}} = \frac{\sqrt{var(\mathrm{Span})}}{average(\mathrm{Span})}$$

$$elem_{\mathrm{Width}} = \frac{\sqrt{var(\mathrm{Width})}}{average(\mathrm{Width})}$$

$$elem_{\mathrm{Scale}} = \frac{\sqrt{var(\mathrm{Scale})}}{average(\mathrm{Scale})}$$

これらは周期的である場合 0 で, 周期的でないほど大きくなる. これらを用いて周期度 $P$ を定義する.

$$\begin{aligned}P &= \textstyle\prod_{x \in C} \mathrm{Exp}(-w_x * elem_x) \\ &= \mathrm{Exp}(-\textstyle\sum_{x \in C} w_x * elem_x)\end{aligned} \tag{2}$$

周期度 $P$ は完全に周期的である場合は 1 となり, 周期的でないイベント群ほど値が小さくなり, そのイベントの周期からのずれが大きいほど小さくなる. また,$w_x$ は重みで, 大きいほど $elem_x$ の値による変動が大きくなる. イベント数, イベント間の生起間隔, イベントの期間, イベントの大きさの順に影響が大きいとして, $w_{\mathrm{Num}} > w_{\mathrm{Span}} > w_{\mathrm{Width}} > w_{\mathrm{Scale}}$ となる.

今回の定義では, 時系列で隣あうイベント同士の生起間隔しかみておらず, 簡単な分散しか考慮していない. この定義は, さらに今後考えていく必要がある. より複雑な周期をもつイベントに関して, モデルをたてて周期度を計算することが必要である.

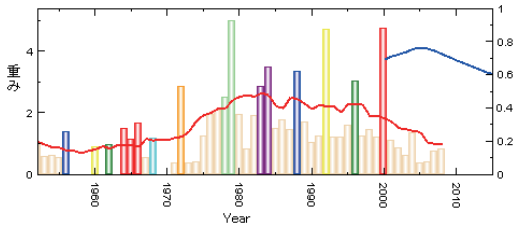


図 6　クエリ:Olympic Summer

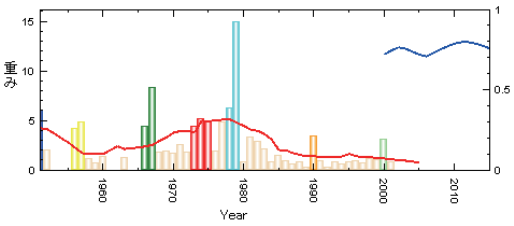


図 7　クエリ:Oil Crisis

### 4.3　将来イベントの予測

前章で過去のイベント群が取得された. これを用いて, 将来のイベントの推測を行う. 時間と過去イベント群が与えられたときに, それらから与えられた時間を中心とするイベントの生起しやすさを計算する. 予測するイベントは周期的に決まるので, 周期的であるほど尤もらしい予測であると考えられる. そこで, 前章で定義した周期度が高いものほど生起しやすいので, 周期度を生起のしやすさとする. 過去のイベント群を $E = \{e_i\}$ として, 与えられた時間にイベント $e_{\mathrm{new}}$ が起こったとする. このとき,$P(e_{new} \cap E)$ を求める.

時間 $time$ が与えられた場合に, イベント $e_{new}$ の最頻値はその値 $time$, 期間はイベント群 $E$ の Width の平均値, 規模を $E$ の Scale の平均値とする.

こうして生成されたイベントと検出したイベント群の集合 $P$ の周期度を求める. 次に, 最も周期度を下げる原因となっているイベントを取り除く.　そのイベントの決め方は,$\mathrm{Span}(i-1,i) + \mathrm{Span}(i,i+1)$ が最小となるものを取り除く. これを繰り返し, 得られた周期度の中から最も高い周期度をその年の周期度とする. イベントを取り除いて繰り返し計算することで, 検出されたイベント群からノイズを除いて計算ができる. しかし, これは周期が長くなる傾向がある.

## 5.　評 価 実 験

前述したイベントの判定, 及びそれを用いた将来イベントの予測の評価実験を行った.1 単位時間は 1 年として分析・推測を行っている. すべての例においてバースト検出の式 (1) では,$width = 0.5$, $\alpha = 0.3$, $\beta = 0$ とした.　予測での重みは,$w_{\mathrm{Num}} = 1$, $w_{\mathrm{Span}} = 0.75$, $w_{\mathrm{Width}} = 0.3$, $w_{\mathrm{Scale}} = 0$ とした. 検索結果は,2008 年 7 月 31 日のものである.

### 5.1　Olympic Summer

提案するイベント判定と推測方法を用いた結果の正しさを示すために, 既知の周期的におこるイベントの将来情報に対する推測を行う. その例として, 四年という周期が明らかな夏季オリンピックの予測を行った. クエリを『Olympic Summer』として 1950 年から 2008 年の時間幅で検索し, バースト検出および推測を行った結果を図 6 に示す. 図中の赤い線はカットオフ値, 青い線は周期度を示し, 図中の色つきの棒グラフはイベントと判断された部分である. 1980 年のオリンピックを,1978-9 年と間違えて判定しているほか,1962 年,1964-6 年のイベントの判定が実際とは異なっている. 1983-4 年のイベント判定は幅が大きくなっている. 検索を 7 月に行ったため,2008 年のペキンオリンピックは検出されていない. 推測については, 周期度が最も高いのは 2008 年となっており, 推測は成功しているといえる.

| 年 | 出来事 |
|---|---|
| 1948 | 第一次中東戦争 |
| 1956 | 第二次中東戦争 |
| 1967 | 第三次中東戦争 |
| 1973 | 第四次中東戦争 |
| 1979 | イラン革命 |
| 1990 | 湾岸戦争 |
| 2000 | 英国での燃料代抗議 |

表 2　主な Oil Crisis

### 5.2　Oil Crisis

ほぼ周期的におこるイベントの推測の例として,『Oil Crisis』というクエリでオイルショックの予測を行った. 時間幅は 1950 年から 2000 年とした. 結果を図 7 に示す. オイルショックの生起は表 2 のようになっており, これらを見比べるとイベントの判別がうまくいっているといえる. イベントの生起がほぼ周期的なので推測を行うと,2003 年・2012 年にピークがある. 2003 年は 2008 年現在まで続く石油の値上がりが始まった時期であり, 適当な推測である.

### 5.3　Windows New Release

ほぼ周期的におこるイベントの二つ目の例として,『windows new release』というクエリで予測を行った. 時間幅は,1990 年から 2008 年とした. 結果を図 8 に示す. ウィンドウズの主な発売時期は表 3 のようになっている. 1990 年・1993 年・2001 年・2007 年が検出されていない. 推測では,2005-6 年がピークになっており, 推測は成功していない. この例では, 検索の幅・生起間隔ともに短いため, 十分な検出ができていないと考えられる. 単位時間を年ではなく, 月や週, 日にすることで, 改善ができるであろう.

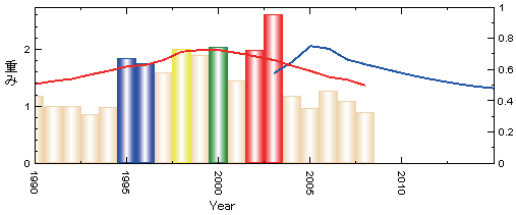


図 8　クエリ:Windows New Release

| 年 | 出来事 |
|---|---|
| 1990 | Windows 3.0 |
| 1993 | Windows NT3.1 |
| 1995 | Windows95 |
| 1996 | WindowsCE 1.01 |
| 1998 | Windows98 |
| 2000 | Windows2000 |
| 2000 | Windows Millennium Edition |
| 2001 | Windows XP |
| 2003 | Windows Server 2003 |
| 2007 | Windows Vista |

表 3　主な Windows 発売時期

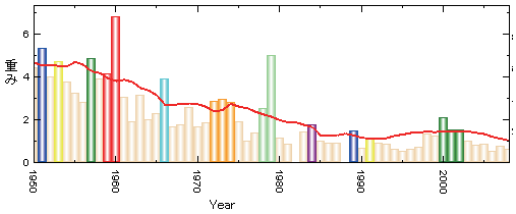


図 9　クエリ:U.S. President Visit Japan

### 5.4　U.S. President Visit Japan

イベントの検出ができなかった例として,『U.S. President Visit Japan』を示す. 米大統領の訪日は 1974 年が最初であり,1970 年以前に検出されたイベントは間違いである. これらの検出されているイベントは, 他の国の大統領の会談などで, 記事中に『U.S.』がでてきたためである. また,1996 年や 1998 年,2005 年,2008 年の米大統領訪日は検出できていない. これは正規化を行ったことによって, 近年の値が小さくなりすぎ, 古いイベントの重みが大きくなりすぎたため, 検出できなかったことが原因だと考える. イベントの判定が適切に行われていないため、推測は行っていない。

### 5.5　Batman Movie

検出できなかった次の例として, クエリ『Batman Movie』の結果を図 10 に示す. 表 4 に示した Batman の映像作品に関するイベントの情報と見比べると, 1995-97 で 2 個のイベントが 1 個と判定されている以外は, 過去のイベントの検出は成功している. 1995-97 が 1 個のイベントと判定されたように, 単位時間に比べて生起間隔の短い複数のイベントが同一のイベントとして判定されやすい.

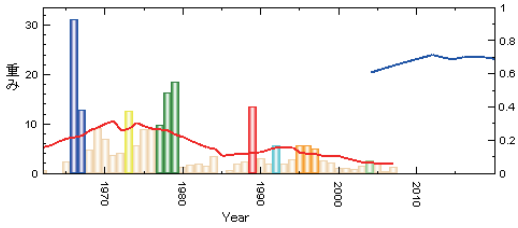


図 10　クエリ:Batman Movie

| 年 | 出来事 |
|---|---|
| 1966-8 | TV ドラマ『Batman』放送 |
| 1977-81 | アニメ『The New Adventures of Batman 』 |
| 1989 | 映画『Batman』公開 |
| 1992 | 映画『Batman Returns』公開 |
| 1995 | 映画『Batman Forever』公開 |
| 1997 | 映画『Batman & Robin』公開 |
| 2005 | 映画『Batman Begins』公開 |
| 2008 | 映画『Batman Dark Night』公開 |

表 4　Batman の映像作品に関するイベント

推測に関しては, 個々のイベントの生起に周期性がみられず, 推測は成功していない. しかし,1966〜1979 年中のイベント・1989〜1997 年中のイベント・2005 年以降のイベントをそれぞれ一つのイベントとしてみると, 生起に周期性がみられる. このようにイベントをクラスタリングすると, 周期性が現れるようなイベント群に対する推測の手法も必要となるであろう. また, 検出されたイベントには映画だけでなく,1966 年のイベントや 1970 年代のイベントなど,TV ドラマ・アニメなど複数の話題が入っていることも問題である.

## 6.　テキスト情報の利用

本研究では, 記事の数を基にした時間ごとの重みを用いて推測を行った. しかし,5 章で示したように, 推測は必ずしも成功していない. そこで, 精度を上げるための今後の課題として, テキスト情報の利用がある. 得られた記事には時刻印だけでなく, スニペットやタイトルなどのテキストの情報もある. これらを考慮すれば, よりよいイベントの分析・推測ができるであろう. 活用の方法として, 記事中の語の特徴ベクトルを用いたイベントの分析と, クエリ拡張について述べる.

### 6.1　語を考慮したイベントの分析

戸田らの研究では [5] では, 時系列を考慮したニュース記事のクラスタリングについて提案している. 文書間の語の特徴ベクトルの類似度と時系列の距離を考慮して, クラスタリングを行っている. 時刻印と語をともに考慮したイベントの判断をすることにより, 5.4 で示した『Batman Movie』などの複数の話題が混在している場合や生起時期の一部が重なるイベントがある場合の判断が可能になるであろう. このように時刻印と文書の特徴を考慮したバースト検出を考える必要がある.

### 6.2　クエリ拡張

提案手法は, ユーザの入力したクエリの結果に, 複数の別のイ

ベントが含まれている場合に正しいイベント分析・推測が行えない. たとえば『Wars』の結果をバースト検出した際には,『戦争』のイベントが検出されるが,『Star Wars』のイベントも検出されてしまう. そこで, 各イベントの語の特徴ベクトルを用いて, イベント間の類似度を比較し, 同種のイベントかどうかを判断することが考えられる. それにより, クエリを拡張して検索するイベントを絞り込み, イベントの分析・推測の精度の向上につなげる.

## 7. 本研究のまとめ

本研究では, 過去のイベント群の取得, 判別をおこなって, それらの情報より周期性を用いて将来のイベントの生起時間を推測する方法を示した. イベントの判定には, 移動平均線を基にしたバースト検出を用いた. イベントの推測には, 周期度を定義して, それを用いた. そして, 提案方法に対して例を示し評価を行い, これらの結果がユーザの推測を支援するものになっていることを示した.

今後の課題として, より実際のイベントに則した精度の高い推測が求められる. 本研究では, 時系列での記事の生成頻度のみを用いていたが,6 章で述べたスニペットなどのテキストの情報も考慮に入れたイベントの分析・推測を行っていく必要がある. また, 時間を記事の発行日としているが, 記事の発行日と記述されているイベントの時間が異なることがある. そこで, スニペットやタイトルの情報を用いて実際の記述されているイベントの時間を時刻とすることが必要である. さらに, ニュースなどのある程度構造化された文章ではなく,Web ページのような非構造な文書からイベントの抽出等をおこなっていくことで、より多くの情報が得られるであろう.

**謝辞**：本研究の一部は, 科学研究費補助金(課題番号 18049041, 18049073, 18700111, 19700091), 文部科学省「知的資産のための技術基盤」プロジェクト, 京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」, およびマイクロソフト産学連携研究機構 CORE 連携研究プロジェクトによるものです. ここに記して謝意を表します.

### 文　　献

[1] R. Baeza-Yates: "Searching the Future", In Proceedings of the ACM SIGIR WorkshopMF/IR 2005. ACM, 2005

[2] B. Wuthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, W. Lam: "Daily Prediction of Major Stock Indices from textual WWW Data", In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. AAAI, 364-368,1998

[3] M. Vlachos, C. Mee, Z. Vagena: "Identifying similarities, periodicities, and bursts for online search queries.", In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM, 131-142, 2004.

[4] M. D. Choudhury, H. Sundaram, A. John, D. D. Seligmann: "Can Blog Communication Dynamics be correlated with Stock Market Activity?", In Proceedings of the 19th ACM Conference on Hypertext and Hypermedia. ACM, 55-60, 2008.

[5] 戸田 浩之, 北川 博之, 藤村 考, 片岡 良治: "時間的近さを考慮した話題構造マイニング", 電子情報通信学会第 18 回データ工学ワークショップ (DEWS 2007), L6-4, 2007.

[6] 櫻井 茂明, 植野 研, 酢山 明弘, 折原 良平: "時系列イベントパターンマイニングにおける時間制約の導入", 電子情報通信学会第 16 回データ工学ワークショップ (DEWS 2005), 6C-o1, 2005.

[7] Z. Li, B. Wang, M. Li, W. Ma: "A Probabilistic Model for Retrospective News Event Detection", In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 106-113, 2005