Web 文書の時系列分析に基づく意見変化イベントの抽出

河合剛巨　　岡嶋穣　　中澤聡
NEC 情報・メディアプロセッシング研究所
{t-kawai@bx，y-okajima@bu，s-nakazawa@da }.jp.nec.com

1. はじめに

大量に存在する Web 文書には, しばしば不正確な記述や偏りのある意見，古びた情報が含まれており，意思決定に役立つ有用な情報を見極めるのは容易ではない. このような問題に対し，Web 情報の信頼性を分析する研究の取組みがある[1][2]. 筆者らは, Web 文書を深く分析し，多様な観点で俯瞰的な情報閲覧を可能とする情報信頼性判断の支援技術を開発している[3][4].

多様な観点の一つとして時間的な観点は有用と考えられる．時間的な観点を提示するため，我々は着目言明[1]に関する重要イベントについてのトピック表現を抽出する手法[5]を提案した．重要イベントは，世間の人々の判断を変えさせうる反響の大きい事件や行動，現象などの出来事である．本手法により，着目言明に関して Web 文書の出現数の時系列変化の大きい期間から，特徴的な部分木からなる言語表現を重要トピック表現として抽出できる.

しかしながら, 抽出した重要トピック表現は，キーワードの組み合わせや断片的なフレーズであり，イベントの内容を適切に表せるとは限らず，ユーザが重要なイベントを把握し判断するには分かり難いという問題がある.

そこで本稿では，Web 文書からユーザの着目する言明に関して，重要なイベントの内容を端的に表す代表文を抽出する手法を提案する.

2. 関連研究

意見の時間的な変化に関連する研究としては, blog と電子掲示板を対象に，その中で注目されている話題を示すトピック語を発見する藤木ら[6]の burst 検出手法がある.

また，数原ら[7]は，ひとつの話題語を入力として，blog 記事から実世界の実体になりうる固有表現に限定し，似た述語パターンを有する固有表現を関連語として抽出している.

しかし，いずれの研究も，ある時期に注目のキーワード群を得る方法である.

3. 時系列分析に基づくイベント抽出

重要イベントが発生すると Web 上での言及数も急増するため，特徴語等の表現は抽出可能である．しかし，それだけを読んでもどういったイベントや状況であるのかを判断・推測するのは難しい.

そこで，表現を単に出すのではなく，重要なイベントに関する文書を選別し，重要イベントの内容を端的に表した文を抽出する手法を提案する．最初に，時系列分析に基づいて重要なイベントに関する特徴表現を抽出する．次に，複数の話題を区別するために各特徴表現をまとめあげ，重要なイベント候補を含むクラスタを抽出する．最後に，抽出したクラスタの重要な順に，クラスタ内の文書から重要なイベントを端的に表す代表文を抽出しユーザに提示する.

以下に，提案手法の詳細な手順を説明する.

3.1. 変化区間検出と特徴表現抽出

最初に，重要イベント毎に特徴表現を得る．そのために，時系列分析に基づいて変化の大きな区間から特徴表現を抽出する.

具体的には，着目言明を表す入力文に対し，検索クエリを生成し発信日付つきの記事インデックスより Web 文書を時系列に検索する. クエリには着目言明の主題に合う自立語を使う．例えば入力文「アスベストには毒性があるそうだ」からは「アスベスト，毒性」の各語による検索クエリを生成する．一方,「ある」「そうだ」のような主題の限定に繋がらない表現は排除する.

次に，得られた時系列の検索結果を用い，大西ら[5]と同様に時系列分析に基づき変化区間を検出し，変化区間の重要トピックを示す特徴表現を抽出する.

変化区間は 1 日毎に過去 1 週での回帰直線による傾きを求め正の最小区間(7 日)を検出した.

[1] 着目言明とは,「石油は枯渇する」のような，ユーザが信頼性を判断したいトピックや内容を表すテキスト情報である.

特徴表現の抽出は，Morinaga らの特徴部分木抽出手法[8]を用いる．正例期間を変化区間とし，負例期間を変化区間以前の 3 ヶ月分として，正例期間と負例期間の検索結果を比較し，正例期間（変化区間）に特徴的に含まれる言語表現を特徴度の大きい順に重要トピックの特徴表現として抽出する．

3.2. 重要イベントクラスタの抽出

前節の処理により，何らかの理由により着目言明に関する言及数に変化が生じた区間と，その区間の特徴表現が得られる．しかし，変化区間毎に抽出された特徴表現は，複数話題の特徴表現が混ざり合っていると考えられる．

そこで，特徴表現を関連する話題ごとに分けるためにクラスタリングを行ない，重要なイベントを想定して各クラスタをランキングする．

まず，クラスタリングは，各特徴表現k_iが出現する変化区間内の文書集合$\{d_1,\cdots,d_m\}$中の文書有無による式(1)を Ward 法に適用する．

$$\mathbf{M}_{ij}=\begin{cases}1/\sqrt{m} & (\text{特徴表現}k_i\text{が文書}d_j\text{に出現})\\ 0 & (otherwise)\end{cases} \quad (1)$$

クラスタリング結果の各クラスタを重要イベント候補とする．

次に，重要イベント候補 e の集合より重要イベントとするクラスタをランキングし選別する．選別は，イベントの前に比べてイベント中の方が文書数が増加すると想定し，式(2)の重要イベントスコアEScore(e)を求め，正の候補をスコア順に選ぶ．

$$EScore(e)=\frac{D_{k\in K+}/T_+}{D_{k\in K-}/T_-}-A_1 \quad (2)$$

第一項の分子は正例期間の平均文書数を表し，分母は負例期間の平均文書数を表す．T+は正例期間の長さで，T-は負例期間の長さである．$D_{k\in K+}$は，正例期間中の文書数，$D_{k\in K-}$は，負例期間中の文書数を表す．A_1は閾値とするパラメータである．

3.3. 重要イベントクラスタの代表文抽出

抽出された重要イベントクラスタには複数の文書を含むため，ユーザに提示するには不便である．また，特徴表現だけを見てもユーザはイベントの内容を把握し難い．

そこで，イベントの代表文として出力することを考える．そのため，クラスタの特徴表現をよく含む文を抽出し出力する．

クラスタ内の特徴表現 k を含む各文 s について，3.1 節で求めた特徴表現の特徴度 G(k)を用いて式(3)の代表文スコア SenScore を求め，最上位 1 件を得る．

$$SenScore(s)=\sum_{k\in K_s}G(k)*P_{len}(s)$$
$$P_{len}(s)=\begin{cases}1 & (length(s)\le 60)\\ 0.5^{(length(s)/60-1)} & (length(s)>60)\end{cases} \quad (3)$$

ここで，$P_{len}(s)$は，あまりに長い文がイベント文として抽出されるのを避けるためのペナルティである．K_sは文 s 内の特徴表現集合である．

3.4. 重要イベントクラスタのリランキング

重要イベントの代表文を選出した後に，この代表文を用いて重要イベントをリランキングする．この代表文には，世の中の話題変化が大きな特徴的記述が含まれうるが，イベントとは無関係の話題記述も含まることがある．そのような記述の優先度を下げるため，ユーザに提示する重要イベントとして相応しい記述を選択するよう，着目言明に関して関連性を有するほど重要とする．また，イベントとしての記述性も考慮する．出来事の記述に現れやすい「よう･だ」「て･いる」「発表」「報道」，などの事態性を表すモダリティ表現を持つ代表文を優先する．

具体的には，式(4)の要点記述スコア GScore を算出し，正の候補をスコア順に選ぶ．

$$GScore(e)=\begin{cases}Escore(e)\sum_{n=1}^{N_q}1/n & (g(e)>0)\\ \beta\,Escore(e) & (otherwise)\end{cases} \quad (4)$$

$g(e)$は，対象イベント e の代表文に事態性を表すモダリティ表現が含まれる場合に正となる関数である．事態性を表すモダリティ表現は，モダリティ表現辞書を事前に設定する．

N_qは，代表文中に含まれる着目言明中の検索クエリ語の数である．βは 1 以下の重みであり，本稿の評価時では 0 を用いた．

4. 意見変化イベントの抽出

3 章までの手法は基本的に Web 文書の時系列変化に基づいているが，人々の意見に影響を及ぼしたかどうかは間接的である．本章では，さ

らに，前章までの処理に加えて意見の変化を取り入れた．重要イベントに対して人々の意見に与えた影響を考慮できれば，より信頼性判断に役立つと考えられるためである．そこで，着目言明に関する肯定意見や否定意見の増加に基づきイベントの反響を測定し，反響の大きいイベントを重要な意見変化イベントとして抽出する．

4.1. 評判情報の抽出

土田らの対象-属性-評価の 3 項関係の評判情報抽出手法[9]に基づき，事前に時間情報付き分析対象コーパスから，評判情報を抽出しておく．具体的には，各記事より評価表現または属性表現と評価表現のペアを抽出しておき，これらと対象物との関係性の同定を行ない，対象物に関する肯定･否定意見としてデータベース化する．

4.2. 意見変化イベントの抽出

重要イベントの人々への反響を考慮するため，評判情報を用いてイベント前後で意見割合が変化したものを意見変化イベントとして抽出する．

具体的には，各イベントに対して意見変化の対数尤度比を基にした式(5)の意見変化スコア $RScore$を算出し，閾値A_2以上のイベントを意見変化イベントとして抽出する．

$$RScore(k_{p_b}, k_{n_b}, k_{p_a}, k_{n_a}) = \sum_{ij} k_{ij} \, log\left(\frac{k_{ij}/(k_{pj}+k_{nj})}{(k_{ib}+k_{ia})/(k_{pb}+k_{nb}+k_{pa}+k_{na})}\right)$$

$$k_{p_b} = r_{p_b}(e) + c \quad (5)$$
$$k_{n_b} = r_{n_b}(e) + c$$
$$k_{p_a} = r_{p_a}(e) + c$$
$$k_{p_a} = r_{p_a}(e) + c$$

$r_{p_b}(e)$はイベントe時点前の肯定意見，$r_{n_b}(e)$はイベントe時点前の否定意見の出現数である．$r_{p_a}(e)$はイベントe時点後の肯定意見，$r_{p_a}(e)$はイベントe時点後の否定意見の出現数である．c は意見数が少ない場合のスムージング定数である．本稿では，c=1 を用いた．

5. 評価実験

提案手法の有効性を示すため，重要イベントスコア，要点記述スコア，意見変化スコアの 3 手法により選ばれた重要イベントの代表文をそれぞれ比較評価した．

5.1. 評価用データの作成

2005 年 1 月～2010 年 9 月の期間に発信日付が付与されインデックス化された Web 文書 2 千 120 万記事を対象にドメインの異なる 30 着目言明の個々に対して重要イベントを抽出した．

各抽出結果について 3 名の被験者により着目言明毎の重要イベント抽出結果を全て採点した．採点基準は表 1 の通りである．着目言明に対する適合性，有用性の 2 つの基準別に採点した．

表 1. 採点基準

適合性	2	着目言明に関する記述になっている
	1	着目言明に対して，間接的に分野が関係のある記述になっている
	0	着目言明と関係ない．記述として不適切．
有用性	2	着目言明の信頼性判断に役立つ
	1	参考として有用な情報や知識になる
	0	無用である

5.2. 評価実験

被験者による採点結果の多数決によりイベントの採点を決定して正解とし，各手法による重要イベントの抽出結果の精度を評価した．

図 1 に評価対象 30 着目言明に対する各手法別の全イベント代表文出力の適合率を示す．

ALL は重要イベント候補の各クラスタを全て使い，代表文を抽出した場合の結果である．

S は，ALL に対して重要イベントスコアを用いた，重要イベントの代表文の抽出結果である．

R は，ALL に対して意見変化スコアを用いた，重要イベントの代表文の抽出結果である．

M は，ALL に対して要点記述スコアを用いた，重要イベントの代表文を抽出結果である．

SRM は，S と R と M の全てのスコアを用いて重要イベントを選出した結果である．

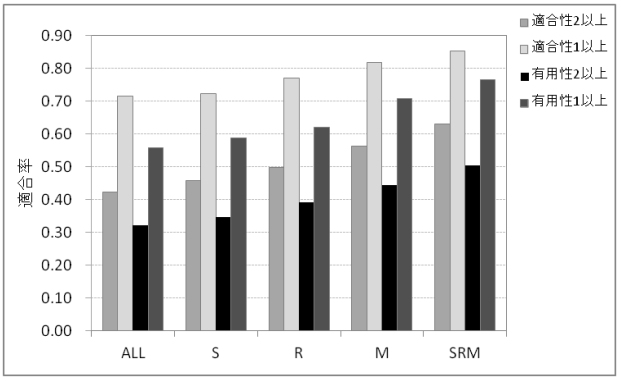

図 1. 各手法の適合率比較

表 2.「適合性」での各手法の再現率

Recall	S	R	M	SRM
2 以上	0.82	0.54	0.76	0.43
1 以上	0.77	0.49	0.65	0.35

表 3.「適合性 1 以上」での適合率比較

適合性 1 以上	S	R	M	SRM
全イベント	0.72	0.77	0.82	0.85
着目言明別	0.71	0.72	0.80	0.76

図 1 の ALL の重要イベント候補レベルでの適合率は，適合性 1 以上で 0.72，有用性 1 以上で 0.56 の精度を有している．つまり，変化区間検出および代表文抽出の基本性能として半数以上の価値ある情報を出力できることになる．

S，M，R の結果では，それぞれ ALL よりも精度が向上していることから，重要イベントスコア，要点記述スコア，意見変化スコアの全てにランキング効果があることが示された．さらに，これらスコアを全て適用した SRM は最も精度が高かった．以上より各手法とも効果が認められ，提案手法の有効性が確認できた．このことから,各クラスタをユーザに提示する際に，必要な数を SRM>M>S の順でランキングすると，ユーザに優先度順に提示することが可能になると考えられる．なお，採点した重要イベントを正解母集団とした時の「適合性」基準での再現率を表 2 に示す．

表 3 に，各手法の 30 着目言明に対するイベント出力の適合率と，着目言明別でのイベント出力の適合率の平均値を示す．S や M は大差がないが，特に R と SRM で差が出ている．これは着目言明によっては評判情報が少ないのが原因であった．意見の数が多い着目言明ほど精度が向上していることを確認している．今後，評判情報の精度改善を行なうことでさらに向上できると考えられる．

6. まとめ

本稿では，Web 文書からユーザの着目する言明に関して，多くの人々の意見や判断に影響を与える変化の要因となるような重要イベントを抽出し，代表文をユーザに提示する手法を提案し，実験により有効性を確認した．

今後は，評価者および評価対象の着目言明を拡充した評価実験を行ない検証の確度を上げる．

また，事業応用に向けたユーザのニーズを分析する予定である．さらに，精度向上のために事実性やイベント性の考慮を行なう等の改善も引き続き検討する．

謝辞

本研究は，独立行政法人情報通信研究機構(NICT)の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」の一環として実施した．

参考文献

[1] S. Kurohashi, A. Akamine, D. Kawahara, Y. Kato, T. Nakagawa, K. Inui, and Y. Kidawara. Information Credibility Analysis of Web Contents. In Proc. of the Second ISUC 2008.

[2] 木俵豊, 赤峯享, 河原大輔, 加藤義清, 中川哲治, 黒橋禎夫, 中澤聡, 乾健太郎, 森辰則. Web コンテンツの信頼性分析. 言語処理学会 第 15 回年次大会, 2009.

[3] 中澤聡, 岡嶋穣, 大西貴士, 河合剛巨, 安藤真一. 時系列分析によるＷｅｂ文書の情報信頼性判断支援：全体概要. 言語処理学会 第 15 回年次大会, 2009.

[4] 岡嶋穣, 河合剛巨, 中澤聡, 村上浩司, 松吉俊, 水野淳太, エリック・ニコルズ, 渡邉陽太郎, 乾健太郎, 渋木英潔, 中野正寛, 宮崎林太郎, 石下円香, 森辰則. Web 文書の時間・論理関係分析に基づく情報信頼性判断支援システムの開発と実証実験. 言語処理学会 第 17 回年次大会, 2011.

[5] 大西貴士, 岡嶋穣, 河合剛巨, 中澤聡, 安藤真一. 時系列分析によるＷｅｂ文書の情報信頼性判断支援：時系列変化からの重要トピックの抽出. 言語処理学会 第 15 回年次大会, 2009.

[6] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学. document stream における burst の発見. 情報処理学会研究報告, 2004-NL-160, 2004.

[7] 数原良彦, 戸田浩之, 櫻井彰人. ブログにおけるイベントマイニングのための適切なキーワード抽出. 電子情報通信学会第 18 回データ工学ワークショップ(DEWS2007), 2007.

[8] S. Morinaga, H. Arimura, T. Ikeda, Y. Sakao, and S. Akamine. Key Semantics Extraction by Dependency Tree Mining, Proc. of KDD2005, pp.666-671, 2005.

[9] 土田正明, 水口弘紀, 久寿居大. ブログからの対象, 属性，評価のオンデマンド評判情報分析システム：eHyouban. 言語処理学会 第 14 回年次大会, 2008.