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

2007－NL－181 (14)
2007／9／26

単語出現の意外性に基づく話題性評価方法

佐藤 吉秀　坂井 俊之　川島 晴美　奥田 英範
日本電信電話株式会社 NTT サイバーソリューション研究所

ある期間における単語の出現頻度が，同単語の過去の出現頻度に対して高く，意外性のある出現と呼べる場合，その期間の出現を話題性が高い出現とみなし，時間的な出現頻度の変化から単語の話題性を評価する手法を提案する．単語の出現が記憶に残る強度を指数関数による減衰曲線で表し，2種類の異なる減衰曲線の差に基づくスコアをインパクト値とすることで，文書増加時の更新が容易で，特に話題発生の際の即応性に優れる手法を実現した．また，実験により，98.5%以上のバースト状態をカバーできることを確認した．

Detection of topic words based on their temporal unexpectedness

Yoshihide SATO, Toshiyuki SAKAI, Harumi KAWASHIMA, Hidenori OKUDA
NTT Cyber Solutions Laboratories, NTT Corporation

Impact value is a filter that detects the increase of word frequency in many documents. The filter works based on the temporal unexpectedness of words, and it enables immediate detection of topics occurrences. Our method has two features, such rapidity of response and ease of adding documents. Experiments shows that the method covered more than 98.5% of bursts in document stream.

1 はじめに

話題は，人々の関心が1ヶ所に集中した状態と定義することができる．情報の氾濫，洪水と言われて久しい現在，とりわけインターネットの世界では次から次へと新しい話題が現れては消えている．検索エンジンの精度向上や，検索・閲覧履歴に基づくリコメンデーション等により，初心者でも，話題や所望の情報を発見しやすい環境は整いつつあるが，その一方で話題のスピードは加速化しつつあり，話題が話題として持続する，すなわち話題が注目を集め続ける期間は短期化傾向にある．

こうした状況で，膨大な情報の中から話題を検出する技術は，情報を積極的に破棄，もしくは無視しない限り，瞬く間に情報を取り逃がしてしまうような現在の状況下では，不可欠な技術の一つである．

はてなブックマーク [1] 等のソーシャルブックマークサイト，gooトレンドランキング [2] 等の blog 記事中での引用数ランキングサイトは，人々が Web ページに対してブックマークやリンクを張る行動を，関心の明示的表現と捉え，多数決の原理に基づいて人気の情報を話題として検出，提示するサービスである．一定期間中に新たに張られたリンクやブックマークを単純に集計するだけでも，ある程度，人々の最新の関心対象を発見することができる．

しかし，ここでのリンクやブックマークは，インターネット上に存在する文書に対してのみ張られるものであるため，話題の種となり得る情報もインターネット上の文書に限定される．より広い範囲から話題を発見するためには，blog をはじめとする CGM など，人々が自由に記述した文書を対象とする必要がある．自由記述の文書には，現実の世界も含めた幅広い情報に対する興味や関心が直接的に表れるためである．

そこで本研究では，人々が自然文で記述した文書群からの最新の話題検出を行うための，単語の話題性評価を目的とする．単語の話題性の高さを数値化することができれば，話題を把握するために必要な単語のみを選択したり，クラスタリング等の文書整理に応用することができる．

ここでの話題性評価に求めるのは，話題の発生に対する即応性である．話題の発生をいち早く捉えて

評価値に反映させることができなければ，短期化する話題の流れに追従することができない．したがって，話題発生過程の初期を捉える能力のある手法をめざす．また，膨大な数の文書を扱い，さらに文書が逐次増加する状況においては，文書が増加した場合の更新処理が，容易に行える手法であることも要求される．

以下，第2章で関連研究について述べ，第3章で提案手法を説明する．続いて，第4章でblog記事を対象に行った実験とその結果を示し，第5章で考察を述べる．最後に，第6章でまとめる．

2 関連研究

Kleinburgの手法[3]は，時系列文書におけるバースト状態の検出手法である．電子メールなどの文書が均等な時間間隔で到着する状態を定常状態とし，それより短い間隔で集中的に文書が到着する状態，つまり時間軸上で文書の密度が高い区間をバースト状態として検出する．この手法では，バースト/非バースト状態間の状態遷移に一定のコストを設けることで，密度の変化に対する過剰反応を防ぎ，異常な集中状態のみを検出できるようにしている．ある時期の文書密度が高いか否かは，正確には全ての文書が揃って初めて判断できることであるため，話題発生の即時検出に対する親和性は低い．

Ishikawaら[4]は，文書が時間と共に忘却され，古い文書が他のどの文書とも類似しにくくなるモデルを導入し，オンラインクラスタリングを行った．忘却のモデルに指数関数を用いるため，DF(Document Frequency)等の統計量は，直前に算出した値を保持しておけば，新たに文書が到着する度に，前回の値を用いて容易に更新できる．更新の容易性に対する着眼は本研究と類似し，提案手法でも同じ指数関数の忘却モデルを用いているが，話題性の評価を行わない点が本研究とは異なる．

3 提案手法

単語の話題性の高さを評価するために，まず，以下に例示するように，意外性のある出現を考える．

- 過去にほとんど出現しないが，最近よく出現するようになった
- 一定頻度での出現が継続していたが，最近の出現頻度が急上昇した
- 以前は高頻度で出現していたが，低頻度状態が長期間継続し，最近再度よく出現するようになった

これらの意外性は，いずれも過去の出現頻度と，最近の出現頻度との相対関係で定義されるものである．このような意外性の高さは，ある単語に注目が集中することによって，相対的に以前よりも高頻度で出現することによると考えることができるため，意外性を以て話題性の高さを評価することができる．

上記をふまえ，話題発生に対する即応性と，更新の容易性と兼ね備えた話題性評価手法として，頻度変化に対する微分作用を持つインパクト曲線と，インパクト曲線を用いて算出される話題性評価値，インパクト値とを定義する．

3.1 インパクト曲線

単語の出現が記憶の中に残る強度は，時間の経過にしたがって薄れる．出現からtだけ時間が経過した後の強度を，Ishikawaらの手法と同様，指数関数を用いて次式で表すことにする．

$$P_T(t) = e^{-t/T} \tag{1}$$

ここで，定数Tは時間経過にともなう強度の減衰の速度を決定するパラメータであり，Tの値が大きいほど，減衰が速い．

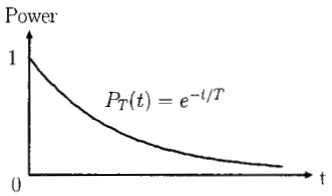

図 1: 時間と記憶の強度

$P_T(t)$に基づき，ある単語が出現する毎に蓄積される記憶の強度の，現在までの総和を求める．

単語wが出現した時刻を，$t_1, t_2, \cdots, t_n$とし，現在時刻をt_nとすると，図2のように，時刻t_1での出現の，現在時刻において残存する強度$P_T(t_n - t_1)$が最も小さく，後の出現ほど残存する強度は高くなり，現在時刻t_nでの出現による強度が最大値1となる．ただし，図1は時間経過に対する強度の減衰を示すグラフであったが，図2では，理解を容易にするため，経過時間ではなく時刻を横軸にとっている．

このとき，単語wの，時刻t_nにおける記憶の強度の総和$M_T(w, t_n)$は，式(2)で表すことができ

る．以後，この総和 $M_T(w,t_n)$ を記憶量と呼ぶ．

$$M_T(w,t_n) = \Sigma_{i=1}^{n} e^{-(t_n-t_i)/T} \qquad (2)$$

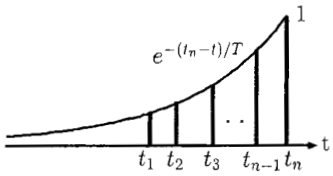

図 2: 記憶量の概念

なお，記憶量 $M_T(w,t_n)$ は，文書が次々と入力される状況において，直前の文書入力時に算出した値を保持しておけば，それを用いて容易に更新することができる．

時刻 t_n から Δt だけ経過した時刻 t_{n+1} に同じ単語 w が出現したとすると，時刻 t_{n+1} における記憶量 $M_T(w,t_{n+1})$ は，式 (3) に示すように計算できる．新たな記憶量 $M_T(w,t_{n+1})$ は，前回までの値 $M_T(w,t_n)$ に対し，経過時間 Δt に相当する減衰 $e^{-\Delta t/T}(<1)$ を乗じ，時刻 t_{n+1} での出現が与える記憶の強度としてさらに 1 を加算した値となる．

$$\begin{aligned} M_T(w,t_{n+1}) &= \Sigma_{i=1}^{n+1} e^{-(t_{n+1}-t_i)/T} \\ &= \Sigma_{i=1}^{n} e^{-(t_n+\Delta t-t_i)/T} + 1 \\ &= e^{-\Delta t/T}\Sigma_{i=1}^{n} e^{-(t_n-t_i)/T} + 1 \\ &= e^{-\Delta t/T} M_T(w,t_n) + 1 \end{aligned} \qquad (3)$$

続いて式 (4) で，2 種類の異なる T の値 $T_s, T_l (T_s < T_l)$ に対して計算した，時刻 t_n における 2 種類の記憶量の差，インパクト値を求める．α, β はそれぞれ定数である．

$$I(w,t_n) = \alpha M_{T_s}(w,t_n) - \beta M_{T_l}(w,t_n) \qquad (4)$$

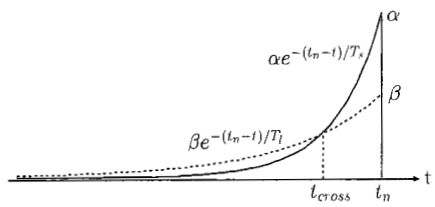

図 3: 2 種類の記憶量

インパクト値 $I(w,t_n)$ の算出を，図 3 を用いて考える．図 3 は，図 2 と同じく横軸に時刻をとり，時刻 t と，$\alpha e^{-(t_n-t)/T_s}$ および $\beta e^{-(t_n-t)/T_l}$ との関係を示したグラフである．無限の過去から時刻 t_n までの 2 曲線の積分値が互いに等しくなるように定数 α および β の値を設定すれば，ある時刻 t_{cross} で互いに交わる 2 曲線となる．このとき，$\alpha e^{-(t_n-t)/T_s}$ と $\beta e^{-(t_n-t)/T_l}$ の差を表すグラフは，図 4 のように，過去からある時刻 t_{cross} までは負値，それ以降の時刻で正値を取る曲線になる．これを，インパクト曲線と呼ぶ．また，インパクト曲線において，無限の過去から交差時刻 t_{cross} までを負区間，t_{cross} から現在時刻 t_n までを正区間と呼ぶ．

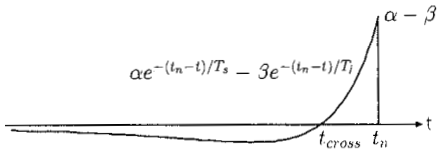

図 4: インパクト曲線

2 種類の記憶量の差であるインパクト値 $I(w,t_n)$ を算出することは，単語 w の各出現時刻に対し，インパクト曲線に基づいて重みを決定し，それらを全て加算するのに等しい．さらに，図 3 の 2 曲線の積分値を等しくしていることから，インパクト曲線の正区間と負区間の面積が等しくなる．したがって，無限の過去から時刻 t_n に至るまで，一定の頻度で出現する単語の場合，インパクト値は 0 になり，頻度が増加傾向にあれば正値，減少傾向にあれば負値となるような，頻度変化に対する微分作用を有する曲線と言える．

言い方を変えれば，インパクト値は，$T_l > T_s$ なる 2 定数に対して，T_s を用いて算出する短期的な記憶量から，T_l を用いて算出する長期的な記憶量を減算し，過去の長期的な出現状況に対する，現在時刻付近での短期的な出現状況の意外性を数値化するものである．T_s, T_l の組み合わせによって正区間の長さが変わるため，想定する話題の周期によって，これらの値を変えればよい．

なお，実際のインパクト値の算出では，2 種類の記憶量をそれぞれ別に計算して保持しておき，それらは文書増加の度に，式 (3) により更新する．インパクト値は，式 (4) にしたがえば，容易に算出できる．1 単語に対して 2 種類の記憶量を保持するだけ

で，インパクト値を算出できるため，スケーラビリティが高い．

3.2 インパクト値と単語の話題性

単語の話題性を正しく評価するために，インパクト値に対して以下の正規化を行い，さらに基準値との比較による有効性判定を行う．

3.2.1 記憶量に基づく正規化

出現頻度の高い単語ほど，インパクト値の振幅が大きくなる傾向があるため，インパクト値の大きさが，話題性の高さを表すのか，一般性の高さを表すのか区別が付かない．そこで，T_s や T_l に比べて極めて大きな T_d に基づいて算出した記憶量 $M_{T_d}(w,t_n)$ でインパクト値 $I(w,t_n)$ を割ることによる正規化を行う (式 (5))．T_d の値が，T_s および T_l に比べて充分に大きければ，減衰速度が極めて遅くなり，$M_{T_d}(w,t_n)$ は，単語 w を含む文書数と近似できる．

$$I_{norm}(w,t_n) = I(w,t_n)/M_{T_d}(w,t_n) \qquad (5)$$

単純に単語 w を含む文書数で割って正規化してもよいが，提案手法の記憶量のモデルでは，単語の出現から充分に時間が経過したときの影響力を 0 にするモデルであるため，正規化の分母もこれに合わせ，充分に大きな定数 T_d を用いて算出した記憶量 $M_{T_d}(w,t_n)$ とした．

さらに，時刻 t_n における $I(w,t_n)$ の値 $\alpha-\beta$ を 1 となるように α,β の値を決定しておくことで，単語 w が初めて出現した時の式 (5) の値が 1 となるようにしておく．

$$\alpha - \beta = 1 \qquad (6)$$

途中の計算は省略するが，式 (6) と，前節で述べたように 2 曲線の積分値が等しいことを表す式 (7) とを解くことで，式 (8) に示す α,β の値がそれぞれ得られる．

$$\int_0^\infty \alpha e^{-t/T_s}dt = \int_0^\infty \beta e^{-t/T_l}dt \qquad (7)$$

$$\alpha = \frac{T_l}{T_l - T_s}, \qquad \beta = \frac{T_s}{T_l - T_s} \qquad (8)$$

以後は，式 (5) の正規化を行い，式 (8) に示す α,β の値を適用した $I_{norm}(w,t_n)$ をインパクト値と呼び変える．

3.2.2 有効インパクト値

インパクト値には母集団の文書数変化が考慮されない．このため，例えば，ある日の母集団の文書数が前日の文書数の 2 倍であれば，単純に考えれば全ての単語の出現文書数も 2 倍になり，あらゆる単語のインパクト値が上昇する．そこで，母集団の文書数変化の影響を無くし，真の意外性を判断するため，基準となるインパクト値を定める．この基準値とは，単語ではなく，母集団の文書集合に対して算出するインパクト値である．仮に母集団の全ての文書に出現する単語があるとすれば，その単語のインパクト値がこの基準値と等しくなる．

基準値以下のインパクト値を持つ単語は，話題性が低いと判断する．さらに，インパクト値が 0 以下の場合も，出現頻度が増加傾向にないことを意味するため，同じく話題性は低いと判断する．上で求めた基準値を超え，かつ 0 よりも大きなインパクト値を，特に有効インパクト値と呼ぶことにする．

4 実験

インパクト値，および有効インパクト値の特性を調べるための実験を行った．使用した実験データは，goo ブログ検索 [5] を利用して継続的に収集した，2007 年 6 月 23 日〜8 月 22 日の 2ヶ月間のタイムスタンプを持つ blog 記事 18,322 件である．検索クエリには，期間中にニュース等で取り上げられる頻度の高かった話題に関する「年金」を使用した．

検索クエリの質が実験結果を左右する可能性があるにも関わらず，検索エンジンを利用して実験データを収集した理由は 2 点ある．有効インパクト値の効果を調べるためには，様々な話題が混在して平均化され，日毎の文書数の変化が小さいデータよりも，限定的な分野の話題を含み，文書数が日によって変動する実験データのほうが望ましいこと，および，分野に依存せずに網羅的に文書を収集することが困難であったことがその理由である．図 5 に，実験データの日別の文書数を示す．

実験は，まず最初に，各文書から人名，地名，組織名，人工物名の 4 タイプの固有表現を抽出 [6] し，全文書数の 0.1%以上に出現する固有表現をインパクト値算出の対象とした．次に，収集した文書を時刻の古い順に 1 文書ずつ入力しながら，文書中に出現する固有表現の記憶量を更新し，さらにインパクト値を算出した．

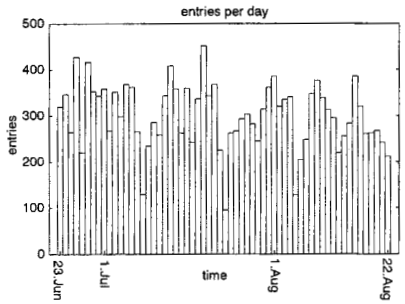

図 5: 実験データの日別文書数

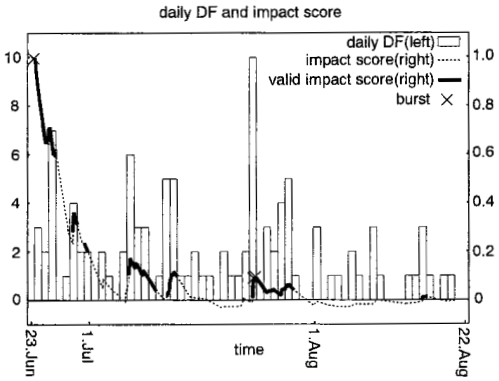

図 6: 日別文書数とインパクト値の一例

比較のため，Kleinburg の手法により，各固有表現に対して，出現の度にバースト/非バースト状態の判定を行った．本実験では，文書が逐次増加する中での話題の発生に対する即応性の観点で比較するため，時刻 t_n におけるバースト/非バーストの判定に，時刻 t_1〜t_n の文書のみを用いた．本来，Kleinburg の手法では，時刻列を入力し，各時刻でのバースト/非バースト状態を一度に判定するが，提案手法との条件を一致させるため，未来の文書を使用しないこととした．

特定の固有表現に注目した場合の，日別の文書数とインパクト値との関係の例を図 6 に示す．日別文書数を棒グラフで，インパクト値を破線で示し，有効インパクト値の区間を太い実線で示した．また，Kleinburg の手法に基づくバースト状態を×印で示した．なお，図 6 は $T_s = 3$(日), $T_l = 9$(日) の場合の例であり，正区間の長さは約 5 日である．

提案手法はインパクト曲線における正値と負値のバランスで話題性を評価する手法である．このため，最初の出現後しばらくの間は，インパクト曲線の負の効果が現れず，インパクト値は大きな値をとっているが，その後は安定し，文書数の増減に応じたインパクト値となっていることがわかる．また，バースト状態は計 3 回 (最初の出現直後に 1 回，最大ピークの付近にほぼ重なって 2 回) 発生しているのに対し，太線で示した有効インパクト値の区間は，最大ピークに比べれば規模が小さいが，中程度に大きなピークもカバーしていることがわかる．

続いて，Kleinburg の手法によるバースト状態を正解とした時の，提案手法の精度と再現率の関係を調べた．インパクト値を降順にソートし，各インパクト値に対応する時刻がバースト状態であれば正解とした．また，インパクト値が有効インパクト値でない場合は，たとえその時刻がバースト状態であっても不正解とした．

今回は，タイムリーな話題に関連するクエリを選んだことから，比較的短い期間で移り変わる話題が含まれていると考え，正区間がそれぞれ約 3 日，約 5 日，約 7 日となるよう，パラメータの組み合わせに，「$T_s = 2, T_l = 5$」「$T_s = 3, T_l = 9$」「$T_s = 4, T_l = 14$」の 3 通りを選んだ．ただし，最初の出現から，正区間の約 2 倍の期間 (それぞれ 6 日，10 日，14 日) のデータは，インパクト曲線の負の効果が現れない不安定な期間として，除外して計算した．

結果を図 7 に示す．提案手法は，話題の発生に対する即応性を重視しているため，文書数の変化に対して敏感であり，Kleinburg の手法でバースト状態と判定されない時刻であっても，頻度が増加傾向にあれば高いインパクト値をとる．正区間が 3 日，5 日，7 日と長くなるほど，文書数の変化に対する安定性が増し，バースト状態の判定結果に近付く結果となった．

表 1 に，対象とした全固有表現の出現毎に算出する，全てのインパクト値のうちで，有効インパクト値が占める割合 (有効判定率) と，有効インパクト値と判定された時刻が，Kleinburg の手法に基づくバースト状態であった割合 (再現率) とを示す．

インパクト値の大小を無視し，有効インパクト値であるとの判定のみをもって話題性があるとみなすならば，Kleinburg の手法が判定したバースト状態を，98.5%以上カバーしていることがわかった．

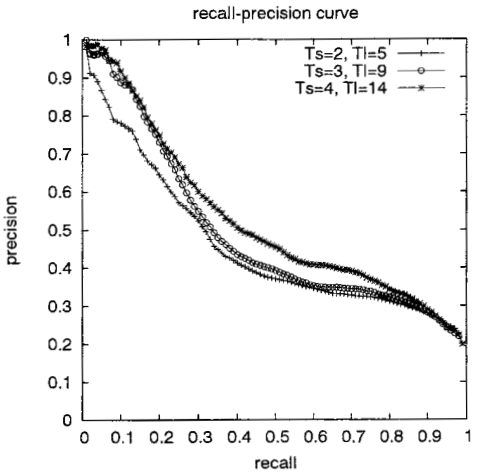

図 7: 提案手法でのバーストの検出精度と再現率

表 1: 有効判定率と再現率

T_s, T_l(日)	有効判定率 (%)	再現率 (%)
2, 5	69.3	98.5
3, 6	68.5	98.7
4, 14	69.4	99.0

5 考察

全体の 68.5%以上のインパクト値を有効インパクト値と判定したことを考慮すると，表 1 の高い再現率も当然であるが，突出したピークを含む，上位〜中位の出現集中状態を検出できたことが評価できる．

出現が記憶に残る強度を，時刻経過に対する減衰モデルで表し，出現の新鮮さを際だたせたため，提案手法は話題の立ち上がりの検出が素早い．一方，出現回数を記憶量として蓄積するため，出現が極度に集中したピークがあると，立ち下がり部分ではピークの余韻で，インパクト値の減少が遅れる傾向がある．しかし，今回，基準値に基づくインパクト値の有効性判定が，頻度の立ち下がり部分で効果を発揮し，突出したピーク後にインパクト値の有効状態が長く持続してしまうのを防いだ．このように，ピーク部分の立ち上がり，立ち下がりを敏感に捉えることができた．

ブログ記事の場合，作成時刻は書き手の生活リズムに依存するため，必ずしもある書き手の昼の記事が別の書き手の朝の記事より新しい内容を含んでいるとは限らない．今回の実験では 1 文書ずつ入力し，RSS から取得したブログ記事の作成時刻を用いて，厳密に秒単位での厳密な更新を行ったが，書き手によって異なる生活パターンを 1 日の単位で集約し，1 日分の収集記事を 1 日に 1 回入力する更新方法も理にかなっていたと言える．この場合は，記憶の強度の減衰が離散的になるだけで，計算の方法は変わらない．このことは，扱う文書集合の規模に応じて，時間軸のスケールを伸縮させる余地があることを意味する．

6 まとめ

長期的に見た頻度と短期的に見た頻度との対比に基づく単語の話題性評価尺度であるインパクト値の算出，更新方法を提案し，実験を行った．指数関数に基づく記憶量の差で評価することから，話題の発生に対する即応性と，文書増加時の更新の容易性を確保できた．また，実験により，Kleinburg の手法を網羅する，広義の話題に反応する手法であることを確認した．

今後は，周期の異なる話題を同時に検出可能な，統合的な手法の検討を行う予定である．

参考文献

[1] はてなブックマーク, http://b.hatena.ne.jp/

[2] goo トレンドランキング, http://blog.goo.ne.jp/portal/trend_ranking/

[3] Jon Kleinberg, "Bursty and Hierarchical Structure in Streams", *Proc. the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002

[4] Yoshiharu Ishikawa, Yibing Chen, and Hiroyuki Kitagawa, "An On-Line Document Clustering Method Based on Forgetting Factors", *Proc. of the 5th European Conf. on Research and Advanced Tech. for Digital Libraries (ECDL 2001)*, 2001

[5] boo ブログ, http://blog.goo.ne.jp/

[6] 齋藤邦子, 鈴木潤, 今村賢治, "CRF を用いたブログからの固有表現抽出", 言語処理学会第 13 回年次大会発表論文集, pp.107-110,2007