イベント系列マイニングを目的とする 新聞記事からの時間情報に基づく単語抽出

A Word Extraction Method from Newspaper Articles Based on Time Information for Event Sequence Mining

多田 知道
Tada Tomomichi

山梨大学大学院 医学工学総合教育部 コンピュータ・メディア工学専攻*1
Computer Science and Media Engineering, Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi

岩沼 宏治
Iwanuma Koji

山梨大学大学院 医学工学総合研究部
Interdisciplinary Graduate School of Medical and Engineering, University of Yamanashi
iwanuma@yamanashi.ac.jp

鍋島 英知
Nabeshima Hidetomo

(同上)
nabesima@yamanashi.ac.jp

keywords: word extraction, event sequence mining, TF-IDF, newspaper article.

Summary

This paper shows a new method of extracting important words from newspaper articles based on time-sequence information. This word extraction method plays an important role in event sequence mining. TF-IDF is a well-known method to rank word's importance in a document. However, the TF-IDF method never consider the time information embedded in sequential textual data, which is peculiar to newspapers. In this research, we will propose a new word-extraction method, called the TF-IDayF method, which considers time-sequence information, and can extract important/characteristic words expressing sequential events. The TF-IDayF method never use so-called burst phenomenon of topic word occurrences, which has been studied by lots of researchers. The TF-IDayF method is quite simple, but effective and easy to compute in sequential textual mining. We evaluate the proposed method from three points of view, i.e., a semantic viewpoint, a statistical one and a data mining viewpoint through several experiments.

1. ま え が き

本論文では、イベントの時系列的関係を表す辞書（イベント系列コーパス）の自動生成を目的として、新聞記事からの時系列情報に基づくイベント単語抽出法について提案する。また、実データを用いた3種類の評価実験を行い、良好な結果を得たので報告する。

近年、Web ニュース、掲示板、ブログなど膨大な量の時系列テキストデータが氾濫している。その中で話題の自動追跡は大変困難な課題であるが、時系列上のイベントを関連付けてあるコーパスがあれば大変役に立つ。イベント時系列コーパスを自動構築するために、[広瀬 04, 高野 04] は系列データマイニング手法を用いて新聞記事コーパスに頻出するイベント系列の抽出を試みている。ここではイベントを表現する重要単語の選定には TF-IDF 法 [北 02] を採用している。TF-IDF 法は情報検索に幅広く使われている標準的手法であるが、時系列上のイベントを表現する単語の抽出には十分ではない。その原因とし

て、新聞記事を文書の集合とみなして、時間情報を考慮していないことが挙げられる。本研究では、この時間情報を考慮するように TF-IDF 法を修正し、新聞記事からイベントをよく表現する単語を抽出する TF-IDayF 法を提案する。また抽出された単語集合に対して、意味論的評価、統計的評価、およびデータマイニング応用上の評価の3種類の観点からの評価実験を行い、良好な結果を得たので報告する。

テキスト時系列からの重要単語の抽出は多くの研究があるが、その多くは、ある期間の単語のバースト出現現象に着目した研究 [白倉 96, Kleinberg 02, 藤木 04] である。本研究の特色は、TF-IDF 法を直接的に改良し、バースト現象を全く利用せずに、簡便で計算も容易な重要単語抽出法を提案したことにある。また本論文で行う評価のうち、データマイニングの観点からの評価は、これまではあまり例が無い新しい試みと考えられる。また意味的評価においては、抽出した約 6000 の単語を全て人手で分類し、比較評価を行っている。

本論文の構成は以下の通りである。2 章では TF-IDayF

*1 現職：三菱電機株式会社

その他の定義を行う．3 章では前述の 3 つの評価実験と結果を示す．4 章はまとめと他研究との比較である．

2. 時間情報に基づく単語抽出

本論文で提案する単語抽出の概要は以下の通りである．

- (1) 前処理として，新聞記事の見出し^{*2}から，形態素解析 [松本 00] により，単語 (名詞と複合名詞) を抽出する．
- (2) 記事見出し中の各単語に対して重み付けを行い，評価値の上位 X 語を抜き出す．

文書中の単語の重み付けの標準的な手法としては TF-IDF 法 [北 02] があるが，本論文では時間情報を加味した，TF-IDayF 法を考える．

2.1 TF-IDF による単語の重み付け

新聞記事から重要単語を抽出する時，TF-IDF は以下の式で示される．

$$\text{TF-IDF}(w, a) = \text{TF}(w, a) \times \log\left(\frac{N}{\text{DF}(w)}\right)$$

TF(w, a): 新聞記事 a 中の単語 w の出現数．TF(w, a) 値が大きいほど， w はある記事 a を特徴付ける．

DF(w): 単語 w が含まれる新聞記事数．DF(w) 値が小さいほど w の重要度が大きい．

N : 全記事数

2.2 時間情報を考慮した TF-IDayF 法による単語の重み付け

時間情報を考慮するために，下記の TF-IDayF 法を導入する．

$$\text{TF-IDayF}(w, d) = \text{TF}(w, d) \times \log\left(\frac{N}{\text{DayF}(w)}\right)$$

TF(w, d): 日付 d の記事における単語 w の出現回数．値が大きいほど， w は d を特徴付ける．

DayF(w): w を含む記事が出現する日数 (day frequency)．値が小さいほど w の重要度が大きい．

N : 全日数

TF-IDayF 法では，単語の出現を日別で区別しており，TF-IDF 法とは違って，時間情報を加えた単語の重み付け^{*3}を行っている．

2.3 他の研究との比較

テキスト時系列データからの重要単語の抽出は数多くの研究があるが，その多くは，ある期間における単語のバースト的な出現現象に着目した研究 [白倉 96, Kleinberg 02, 藤木 04] である．基本的に，ある期間に集中して出

現した語を重要なトピック単語として抽出する．例えば [Kleinberg 02, 藤木 04] では，平常時と集中 (バースト) 時の単語の出現間隔に異なる確率分布を仮定して，今着目している出現間隔の状況がどちらの分布で良く近似できるかを判定して，着目している単語のバースト出現の有無やその期間，バースト強度などを算出している．

本論文で提案する TF-IDayF 法は，時間情報を利用した TF-IDF 法の直接的な改良となっている．単語のバースト的な出現現象は全く利用せず，その代わりに「語の蔓延的な出現現象」に着目している点に特徴がある．即ち，TF-IDayF 法は，時系列上に蔓延して (= 万遍なく) 出現する単語を排除し，節目節目に出現した単語のみを重要単語として抽出する手法と考えられる．簡便で計算も容易な手法であるも関わらず，有効に働くことを次節以降の評価実験で明らかにする．

一般にバースト出現した単語は，非常に重要なイベントを表すと考えられるが，そのような語だけを抽出・利用したのでは，語彙が不足して，通常のイベント系列の抽出や，イベント系列コーパスの作成は困難と考えられる．より多くの (明確なバースト現象は示さない) 「それなりに重要なイベント」を表現する語を抽出する技術が必要と考えられる．TF-IDayF 法はそのような用途に適した手法と考えられる．

3. 評価実験と考察

毎日新聞 2004 年上半期，社会面 181 日 (13499 記事) 分を入力データとし，単語 (名詞と複合名詞) を抽出する．181 日分の記事見出し中の単語出現総数は 24850 であり，1 日当りの平均出現数は 137 となる．これらの単語を TF-IDF 法と TF-IDayF 法で重み付けし，1 日当り上位 20 単語を重要単語として抽出^{*4}し，比較した．その結果 TF-IDF 法で 3579 単語，TF-IDayF 法で 3278 単語 (共通単語 420 語) ^{*5}を抽出した．抽出した単語の品質を以下のように意味的観点，統計学的観点，及び実際にデータマイニングを行った結果から評価を行う．

3.1 意味的分類

抽出単語を事件，行動・動作，人物・役職，場所，組織，その他，意味不明語の 7 つに分類し，手法間の傾向の違いを示す．以下，7 つの分類について説明する．

事件：イベントを想起することができる複合名詞．イベント表現としては非常に重要である．

例：鳥インフルエンザ拡大，小泉首相訪朝 etc...

^{*4} TF-IDF 法による単語重み付けに関しては，記事ごとに各単語の TF-IDF 値を出し，それらを日毎にまとめてソートする．まとめたものの中から日毎に上位 20 単語を重要単語として抽出している．

^{*5} 2 つの手法で抽出単語数が異なるのは，3.2.1 節で示すように，日毎に抽出される単語の重複度の違いによるものである．

^{*2} 見出しに限定したのは，見出しが新聞記事の究極の要約といえるからであり，冗長な語の抽出を極力抑制するためである．

^{*3} 本論文では，時間単位は便宜上「日」としている．必要に応じて適宜，変更していただいで問題はない．

表 1 分類結果

カテゴリ	TF-IDF	TF-IDayF
事件	85	242
行動・動作	818	450
人物・役職	641	942
組織	223	233
場所	321	287
その他の単語	1,370	1,062
意味不明語	121	62
合計	3,579	3,278

行動・動作：行動や動作を表す単語であり、重要なカテゴリである。

例：控訴，拉致，帰還，退職 etc...

人物・役職：固有名詞や社長，大臣といった役職名，主語や目的語などにあたる単語で，重要である。

例：山梨太郎，山梨大学学長 etc...

組織：組織や団体などを表す単語．主語や目的語などにあたる単語である。

例：毎日新聞，山梨大学，厚労省 etc...

場所：国や地名を表した単語．イベントの発生場所にあたる単語である。

例：甲府市，山梨県，日本 etc...

その他の単語：上記の 5 つのカテゴリ以外の単語で，例えば，物や出来事，賞，病気といった単語である。

意味不明語：その語単体では自立語になりえないような単語である。

例：さま，氏，お，新 etc...

2 つの手法において抽出された単語を，以上の 7 つの分類を行った結果を表 1 に示す。

大きな変化としては，TF-IDayF の方が重要なカテゴリである事件が多く，不必要なカテゴリである意味不明語が少ない．一般的に新聞記事はかなり短いので，TF-IDF 法による単語重みの妥当性はかなり怪しいと考えられる．これに対して，日単位で複数の記事をまとめて扱う TF-IDayF 法は，より大きな仮想記事を扱うことになるために妥当な重みを安定的に出力することが可能と思われる．意味不明語などの多くはランダムに近い形で出現すると考えられるので，TF-IDayF 法によって，より安定的に排除できたと考えられる。

3.2 抽出単語の統計的観点からの評価

統計的な観点からの評価項目を以下に示す。

- (1) 抽出された単語の出現頻度：TF-IDF と TF-IDayF によって抽出された単語の出現日数の分布。
- (2) 抽出された単語の文書や日付に対する特定性：単語の IDF 値と IDayF 値の分布。
- (3) 抽出された単語の周期性：TF-IDF と TF-IDayF によって抽出された単語の出現間隔の平均と出現間隔の標準偏差との関係。

図 1 単語別の出現日数の順位

表 2 出現日数別の単語数

出現日数	TF-IDF	TF-IDayF
1	3,442	1,955
2	11	419
3	4	188
4	10	138
5	7	96

§ 1 単語の出現日数の分布

出現日数が多い単語はデータマイニングにかけた際，その単語を中心に頻出パターンが生成されるため，影響が大きい．単語を出現日数により降順にソートし，グラフ化したものを図 1 に示す．図 1 を見ると，TF-IDF は単語による出現日数の差が TF-IDayF に比べて非常に大きいことがわかる。

出現日数別の単語数を表 2 に示す．これを見ると，出現日数が 1 日の単語数は，TF-IDF の方が格段に多い．実に 96% の単語の出現日数が 1 となった．これは，IDF 値が単語の選定に大きく影響しているためだと思われる．そこで，候補単語をそれぞれ IDF 値と IDayF 値によって降順にソートし，グラフ化したものを図 2 に示す．IDF 値の最高値は 1 記事（即ち 1 日）だけに出現する単語の値である．1 記事だけに出現する単語が非常に多い事が分る。

出現日数が多い単語を表 3 に，出現頻度が多い単語のカテゴリ分類を表 4 に示す．表 3 より「新」や「後」といっ

図 2 単語別の IDF 値と IDayF 値の順位

表 3 出現日数上位の単語

TF-IDF		TF-IDayF	
単語	出現日数	単語	出現日数
年	153	北朝鮮	84
氏	92	鳥インフルエンザ	72
火災	86	茨城	59
後	73	強盗	58
新	69	村	58
駅	69	事件	53
車	69	拉致	52
会長	63	囲碁	45
お	61	中国	44
村	53	厚労省	44

表 4 出現日数上位 50 単語の分類

カテゴリ	TF-IDF	TF-IDayF
事件	0	2
行動・動作	8	14
人物・役職	1	2
組織	7	5
場所	7	8
その他の単語	11	14
意味不明語	16	5

た意味不明語の占める割合が，TF-IDayF の方が TF-IDF よりも小さいことがわかる．TF-IDayF では，「北朝鮮」や「鳥インフルエンザ」といった 2004 年上半期に話題になった単語を多く抽出できている．以上の結果より，TF-IDayF の方が TF-IDF に比べて，意味不明語をより多く削除することができ，イベントを想起させる語が多く出てきたことがわかる．

§2 単語の出現間隔の平均と標準偏差の関係

単語の出現間隔とは，ある単語が出現してから，次に出現するまでの日数を指す．出現間隔の平均と標準偏差の関係を図 3 と図 4 に示す．

出現日数が 1 日の単語は，出現間隔の平均と標準偏差は 0 となり，出現日数が 2 日の単語は出現間隔の標準偏差の値が 0 となる．TF-IDF によって絞り込まれた単語の出現日数の 96% が 1 日であるため (表 2)，座標 (0, 0) に集中している (図 3)．また，TF-IDayF に関しても単語の出現日数の 60% が 1 日なので，(0, 0) に集中している (図 4)．さらに，TF-IDF は 11 単語，TF-IDayF では 419 単語の出現日数が 2 日であるため，横軸上に無数の標準偏差の値が 0 の点が存在する．標準偏差が小さい単語は周期性が強く，自己相関性などの時系列データ解析技術等で抽出して別途取り扱うことも可能であるから，ここでは重要視しない．また，出現間隔の平均は小さいが標準偏差が大きい単語は，ランダムに出現する場合や，同じ表記で異なる意味を持つ場合が多い．さらに出現間隔の平均と標準偏差が共に大きい単語は，他の単語との因果関係は殆んどないと考えられる．以上から，イベント系列の要素として望ましい単語 (イベント) は出現間隔の

図 3 TF-IDF によって抽出された単語の分布

図 4 TF-IDayF によって抽出された単語の分布

平均と標準偏差が中程度の値をもつ単語であると考えられ，TF-IDayF によって抽出された単語群の方が TF-IDF により抽出されたものより望ましいと言える．出現間隔の平均と標準偏差が中間の単語に絞って抽出していけば，より理想的なイベント表現用の単語が抽出できる可能性もあると思われる．

なお抽出対象を，1 日当り，TF-IDF および TF-IDayF 値の上位 50 単語にして同様な評価実験を行った結果，抽出単語の分布はあまり変わらず，殆ど同じ結果となった．更に，抽出する単語を上位 100 単語に増やした場合は，共通して抽出される単語が大幅に増え，TF-IDF および TF-IDayF の間で分布の違いは薄れてくる．

表 5 実験結果

TF-IDF		TF-IDayF	
抽出パターン数	100,804	抽出パターン数	681

表 6 TF-IDayF 法により抑制できた時系列パターン

系列長 4	
《(後)(%)(会長)(氏)》	《(後)(本)(駅)(台)》
《(後)(%)(会長)(氏 年)》	《(後)(本)(年)(台)》
《(後)(%)(駅)(氏)》	《(後)(本)(駅 年)(台)》
《(後)(地震)(年)(駅 年)》	《(後)(本)(車)(会長)》

3.3 系列データマイニングを用いた評価

TF-IDF 法と TF-IDayF 法によって絞り込まれた単語データ系列に対し、系列データマイニングを実行し、有用性を検証する。イベント系列コーパスの生成を目的とした頻出系列パターン抽出に、[高野 04, Iwanuma 05] の系列マイニング技術を利用する。新聞記事コーパスなどの単一で長大なデータ系列 S を考え、 S 中に繰り返し出現する部分系列パターンの高速抽出を行う。部分系列の発生期間であるウィンドウ幅 W と部分系列の最小出現頻度 MS をパラメータとして与え、単一の長大なデータ系列からウィンドウ幅 W 内に出現する部分系列のうち、最小出現頻度 MS 回以上出現する部分系列を抽出する。

実験の設定は最小出現頻度を 36 日 (全日数の 20%)、パターンの発生期間であるウィンドウ幅は 1 週間である。使用した新聞記事コーパスは、2004 年上半年期毎日新聞記事コーパス社会面 (13499 記事) である。表 5 より、抽出パターン数を見ると両手法間の間に、実に 150 倍弱もの差が出ている事がわかる。

TF-IDayF では抽出を抑制できたパターンの一部を表 6 に、表 7 と表 8 に TF-IDayF 法と TF-IDF 法によって生成された系列パターンの一部を表す。なお系列 $\langle s_1, \dots, s_n \rangle$ は、系列要素 $s_1, \dots, s_n$ がこの順に出現することを表し、各系列要素は同時に発生するイベントの集合 $\langle e_1 \dots e_k \rangle$ を表現している。

表 6 から、多くの無意味な系列パターンを抑制できたことがわかる。系列パターンの中で、1 つの括弧内に表記されている単語集合はある特定の日に同時に出現した単語を意味する。2004 年上半年期の主な事件として、「北朝鮮拉致被害者の帰国」や「鳥インフルエンザの発生」、長崎の佐世保での小 6 女児殺害事件」などがあり、それぞれ事件を表す単語が表 7 と表 8 より、いくつも抽出さ

表 7 TF-IDayF 法により得られた時系列パターン

系列長 2	
《(帰国 北朝鮮)》	《(被害者 強盗)》
《(事件 北朝鮮)》	《(小 6 女児殺害)(搜索)》
系列長 3	
《(小 6 女児殺害 佐世保 長崎)》	《(被害者 拉致 北朝鮮)》
《(拉致 北朝鮮)(帰国)》	《(事件)(茨城)(村)》
《(北朝鮮)(拉致)(被害者)》	《(判決)(病院)(厚労省)》
《(鳥インフルエンザ)(山口)(搜索)》	《(事件)(行方)(搜索)》
《(三菱ふそうタイヤ脱落)(事件)(搜索)》	

表 8 TF-IDF 法によって得られた時系列パターン

系列長 2	
《(少年 年)》	《(火災 職員)》
《(新 火災)》	
系列長 3	
《(火災 氏 年)》	《(後)(本)(割)》
《(車)(事件)(後 氏)》	《(火災)(職員)(全)》
《(新)(お 年)(川崎 年)》	《(お)(事件)(地震)》
《(氏 年)(台)(割)》	

れていることがわかる。また、「(拉致 北朝鮮)(帰国)」や「(事件)(行方)(搜索)」といった時系列的パターンも抽出されている。一方、TF-IDF に関しては、出現日数が多い語を含む系列パターンが非常に多く抽出された。「(車)(事件)」「(火災 職員)」といった時系列的パターンも抽出されたが、2004 年上半年期独自の事件に関連する系列パターンを得ることはできなかった。

4. む す び

本論文では、イベント系列コーパスの自動生成を目的とした、新聞記事からの時系列情報に基づく単語の抽出法を提案した。既存の TF-IDF 法では、一般的で意味のない単語を多く抽出するという欠点があった。そのため、TF-IDayF 法を新たに提案し、単語絞り込みを行った。単語の重み付けを記事ごとではなく、日毎に行うというシンプルなアイデアではあるが、不要な単語の抽出を減らし、その日の特徴づける意味のある単語を多く抽出することができた。TF-IDayF 法は、時系列上の「蔓延して出現する単語」を排除し、節目にのみ出現する単語を重要単語として抽出する手法である。簡便で計算も容易であるも関わらず、有効に働いた。特にデータマイニング応用の実証実験において、その効果が確認され、非常に良い結果が得られた。実用上、大変有効であると考えられる。

TF-IDayF 法による単語絞り込みによる有用系列パターンの見逃しなどは今後の課題である。この定量的な評価と改善策の立案は本質的な意味で難しい課題を持つと思われる。今後研究を進めて行きたいと考えている。

謝 辞

本研究の一部は科研費基盤研究 A(No.20240016) の援助を受けている。

◇ 参 考 文 献 ◇

- [藤木 04] 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, pp.85-92 (2004)
- [広瀬 04] 広瀬千夏, 高野洋, 岩沼宏治, 鍋島英知: 大規模時系列データからのイベントを象徴するキーワードの自動抽出. 人工知能学会研究会資料 SIG-KBS-A304, pp.239-244

(2004)

[Iwanuma 05] Iwanuma, K., Ishihara, R., Takano, Y, and Nabeshima, H: Extracting Frequent Subsequences from a Single Long Data Sequence: A Novel Anti-Monotonic Measure and a Simple On-Line Algorithm. *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM '05)*, pp.186-193 (2005)

[北 02] 北研二, 津田和彦, 獅々堀正幹: 情報検索アルゴリズム. 共立出版株式会社 (2002)

[Kleinberg 02] Kleinberg, J.: Bursty and hierarchical structure in streams. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.91-101 (2002)

[松本 00] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸: 日本語形態解析システム『茶筌』version 2.2.1 使用説明書. (2000)

[白倉 96] 白倉悟子, 梅村恭司, 小川貴英: 新聞記事における事件特定のための単語群の抽出 情報処理学会研究報告 1996-NL-113, pp.113-117 (1996)

[高野 04] 高野洋, 岩沼宏治, 鍋島英知: 単一の長大なデータ系列上の系列パターンの出現尺度とその逆単調性. 第3回情報科学技術レターズ, LF-012, pp.115-118 (2004)

〔担当委員: 浅井 達哉〕

2009年1月19日 受理

著者紹介

多田 知道

2007年山梨大学工学部コンピュータメディア工学科卒業.
2009年同大学大学院修士課程コンピュータメディア工学専攻在学修了. 現在, 三菱電機(株)勤務. 系列データマイニング等の研究と開発に従事.

岩沼 宏治(正会員)

1985年東北大学大学院・電子及通信工学専攻修士課程修了. 同年山形大学情報工学科助手. 現在, 山梨大学大学院・医学工学総合研究部教授. 工学博士. 人工知能の基礎, 特に非単調論理, 定理自動証明プログラム, WEB 知的処理, 系列データマイニング等の研究と開発に従事. 1987, 89, 90, 91年人工知能学会全国大会優秀論文賞受賞. 2004年FIT 優秀論文賞受賞

鍋島 英知(正会員)

2001年神戸大学大学院・自然科学研究科情報メディア科学専攻博士課程修了. 同年山梨大学工学部コンピュータ・メディア工学科助手. 現在, 同大学大学院・医学工学総合研究部准教授. 博士(工学). アクション理論, プランニング, 定理自動証明プログラム, WEB 知的処理などの研究に従事. 2004年FIT 優秀論文賞受賞.