様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 ２８ 日現在

機関番号：３２６２２
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５５８００９３

研究課題名（和文）自己相関関数を用いた言語統計的手法の確立に関する研究

研究課題名（英文）Calculating Autocorrelation Function for Word Occurrences in Texts and Its Modeling with Stochastic Processes

研究代表者

小倉　浩（OGURA, HIROSHI）

昭和大学・教養部・准教授

研究者番号：４０２１４１００

交付決定額（研究期間全体）：（直接経費）　　1,700,000円

研究成果の概要（和文）：テキストデータを動的な時系列データととらえる方向性での研究は少なく，従ってテキストデータの時間的な相関を問題とした研究もほとんど行われていない．本研究では，テキストデータを動的な時系列データととらえ，考えている語の文書中での出現状況の相関を表す自己相関関数を計算する方法を提案する．提案手法では，適切な自己相関関数を定義するために，語の出現過程を記述するための基本時間単位を，文書中の個々の文に設定する．文書全体の主題と密接に関連した「概念語」と，文書全体の主題と密接な関連を持たない「非概念語」に対して，それらの自己相関関数が全く異なる特徴的な振る舞いをすることを示す．

研究成果の概要（英文）：In this study,we attempt to offer a new analyzing point of view for texts in which occurrences of words are considered as dynamical time series. Based on this interpretation of texts, we propose a method for calculating autocorrelation function (ACF) which represents the correlation between occurrences of a considered word. In our method, the basic time unit of the stochastic process of word occurrence is taken to be one sentence and this allows us a suitable definition of ACF. The examples of ACF obtained through our method for 'conceptual words'and those for 'nonconceptual words' are given and their characteristic behaviors are discussed. Here, the term 'conceptual word' means the word which is deeply related with the central concepts or themes of text, and the 'nonconceptual word' represents the word which is not related with themes of text. It was found that the ACFs for 'conceptual words' and those for 'nonconceptual words' show entirely different characteristic behaviors.

研究分野：統計的機械学習，計量言語学

キーワード：自己相関関数　拡張指数型関数　確率過程　ポアソン過程　非定常ポアソン課程　Cox過程　言語統計　テキストマイニング

１．研究開始当初の背景

研究代表者は自動文書分類のための特徴語抽出分野における研究を行ってきた．ここで，自動文書分類とは，与えられた文書をあらかじめ規定された複数のカテゴリのうちのどのカテゴリに分類するかを自動的に決定するアルゴリズムのことである．自動文書分類においては，各カテゴリの文書を特徴づけるような単語，すなわち特徴語をいかに精度よく抽出するかが分類精度の向上に大きな影響を与えることが知られている．各語ごとに，それぞれのカテゴリに分類するために使用することがどの程度ふさわしいかの目安を与える特徴語抽出の指標を計算し，その値に基づいて特徴語抽出が行われる．研究代表者は，それまでによく使用されていた情報利得（Information Gain）やカイ 2 乗分布指標とは異なる，ポアソン分布からのずれを用いた特徴語抽出指標を提案し，その有効性を明らかにしてきた①～④．本研究は，この考え方を援用し，テキストを動的な，時系列にしたがって生成されたものであると考えたときに，文書を特徴づける単語をどのように選択するのか，また文書中での単語の重要度をどのように推測するのかという問題意識に基づいて開始されたものである．

２．研究の目的

電子化された文書の蓄積量の増加に伴い，情報科学分野におけるテキストマイニング技術や，その基礎を与える言語統計分野（計量文体学，計量言語学，コーパス言語学等の言語学諸分野）の発展に大きな期待が寄せられている．これらの諸分野では多変量解析法を含む統計学の各手法を援用してテキストデータの解析を行うが，筆者の知る限りその多くが静的な統計指標を用いた解析に限定されている．上述の自動文書分類のための特徴語抽出指標も静的な統計手法である．テキストデータを，動的な時間発展の概念を伴った時系列データであると見なした上でその特徴を解析した例も存在するが，その解析手法はやはり静的な統計指標を用いたものが多く，動的なデータ本来の特徴を明らかにするための時系列解析の手法を直接適用した研究例は少ない⑤．

この一因として，物理学（統計力学・物性物理学），工学（信号解析），経済学などにおいて，時系列データを解析する際に最も基本的かつ不可欠な量である自己相関関数（autocorrelation function：ACF）を，テキストデータに対してどのように定義したらよいのかが明らかでないことが挙げられる．Sarkar ら⑥および Altmann ら⑦は，注目している語が文書で一度使用されてから再び使用されるまでの間の語数（ギャップ長：gaps between term または再起時間：recurrence time）の分布を調べ，不完全ながらも時系列解析の概念を導入した．本研究は，これらの研究に触発されたものであるが，ギャップ長の分布と比較してより本質的かつ応用範囲の広い自己相関関数を用いて，動的時系列データとしてのテキストデータの解析手法の確立を目指すものである．

３．研究の方法

Sarkar ら⑥および Altmann ら⑦による既存のテキストデータの動的解析では，すでに述べたようにもっとも基本となる時間ステップとして語のカウント数を用いている．すなわち，既存の方法では，現在の時刻と n 時間ステップ後の時刻との相関を，現在の位置の語と n 語離れた位置の語との相関と考える．しかし，この時間単位の取り方を用いて自己相関関数を求めた場合，通常のテキストデータにおいて同一の単語が 2 度続けて記述される確率はほぼ 0 であるから，1 時間ステップ後の自己相関関数の値はほぼ 0 となってしまうことになる．すなわち，（時間ステップ数）=（2 つの単語の間の語数）とする従来の方法では，通常の意味での自己関数関数と同様の振る舞いをする自己相関関数を定義することはできない．そこで，本研究では語数を時間ステップ数にとるのではなく，（1 文）=（1 時間ステップ数）と考える．すなわち，ある語について現在と n 時間ステップ後の時刻との相関を，考えている語が現れる文とそれから n 個離れた文に現れる語との相関としてとらえる．二つの連続した文に同一の語が使用される確率は 0 ではなく，一度文中に出現した語がその後しばしば高い頻度で文書に出現するという語のバースト性から考えれば，むしろその確率は高くなる．この考え方が本研究における自己相関関数定義の基礎であり，これにより他の時系列解析を使用する分野における一般的な自己相関関数と同様の振る舞いをする自己相関関数を定義することが可能となる．

また上記 2 つの関連研究は，基本的な確率過程である通常のポアソン過程およびその拡張である非定常ポアソン過程の枠組みの範囲で，問題にしている語のテキスト中での出現パターンをモデル化しようとしたものであると考えることができる．しかしいずれのモデルにおいても，使用している確率過程が独立増分性の仮定を前提としているために，語出現のバースト性を適切にモデル化するには至っていない．本研究は，上記 2 つの関連研究⑥⑦と比較して，以下の点に新たな貢献がある．

- モデル化に使用する確率過程として，非定常ポアソン過程をさらに一般化した Cox 過程を使用する．非定常ポアソン過程における強度関数は時間を指定すれば一意に決定される確定関数であるが，Cox 過程では強度関数そのものが確率過程として与えられるため，記述できる確率過程の範囲がより広範なものとなる．本研究で Cox 過程を使用する意図は，過去の語出現の履歴情報を取り入れた

確率過程を強度関数として設定することにより，語出現のバースト性が記述できる可能性を追及するためである．

- 上記 2 つの関連研究では，語出現パターンの特徴を記述するためにギャップ長を使用している．本研究では，ギャップ長を使用する代わりに，文書中での語の出現パターンを特徴づける量として自己相関関数を使用することを提案する．自己相関関数は語の出現履歴の相関を直接記述することができるため，語出現のバースト性を計るために最もふさわしい量であると考えられる．自己相関関数を使用することにより，後述するように語の出現を計数した確率過程が，ポアソン過程のような独立増分を前提とする確率過程に従っているのか，あるいは本研究で新たに提案するような過去の語出現の履歴に依存する確率過程に従っているのかを明確に特徴づけることが可能となる．

４．研究成果

（1）時間とともに連続的に変化する場合の最も一般的な自己相関関数の定義式から出発し，（1 文）=（1 時間単位）と考えた場合の文書中の語の自己相関関数を効率的に計算するための計算式を提案し，その妥当性を検証した．複数の学術的書籍を選択し，それらの書籍中に出現する各語について，提案した計算式に基づいて自己相関関数を計算したところ，時刻 0 における規格化された自己相関関数の最大値 1 から，単調に減少するような，一般的な自己相関関数と同様の振る舞いが再現された．

（2）テキストデータに使用した学術的書籍中に 100 回以上出現する頻出語についてその自己相関関数を計算したところ，特徴的な振る舞いを示す 2 つの極端な頻出語グループの存在が明らかとなった．これらの極端な語の自己相関関数の代表例を図 1 および図 2 に示す．これらを仮に概念語（図 1）および非概念語（図 2）と呼ぶことにすると，概念語は一度文書中に出現すると，その後もしばしば文書中で使用され，徐々に使用されなくなる単語である．一方，非概念語は図 2 に見られるように，初期値 1 を除いて自己相関関数がほぼ一定の値をとり，文書中で使用される頻度は時間（文章経過）にかかわらず常に一定であることが示唆される．

概念語は，バースト性を伴って文書中に出現する語であり，その出現パターンから文書の主題と密接に関連する何らかの概念を説明するために使用される語であると推測される．一方，非概念語は文書の主題に関連する何らかの概念を説明するために使用される語ではなく，常に一定の確率で「偶然に」文書中に出現する語であると考えられる．

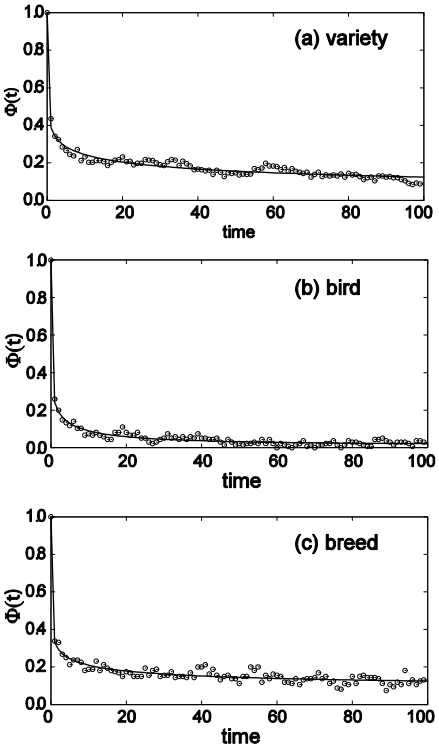

図 1　ダーウィンの「種の起源」中における典型的な概念語である(a)variety，(b)bird，(c)breed それぞれについての自己相関関数．

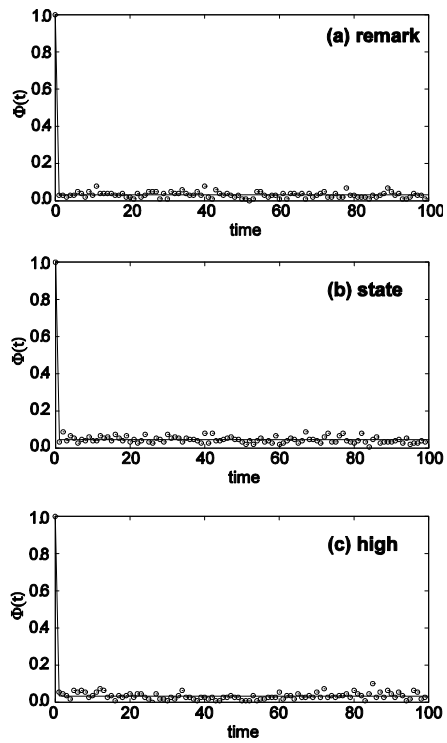

図 2　ダーウィンの「種の起源」中における典型的な非概念語である(a)remark，(b)state，(c)high それぞれについての自己相関関数．

（3）概念語グループ，非概念語グループ以

外のすべての語は，これらの 2 つの極端なグループの中間的な振る舞いを示すことが明らかとなった．そこで，概念語および非概念語を両極端として含み，その中間的な振る舞いを示す大多数の語も含めて，得られた自己相関関数を経験的にフィッティングできるようなモデル式を提案し，すべての語の自己相関関数がこのモデル式のフィッティングパラメータの調節により表現可能であることを確認した．ここで使用したモデル式は，拡張指数型関数⑧（物性物理の分野で KWW 関数と呼ばれるもの）に定数項を加えたものである．このフィッティングモデル式については，拡張指数型関数部分がバースト性を伴う概念語の出現パターン（過去の語の出現履歴が自己相関関数に与える影響を表している部分）であり，定数項が過去の語の出現履歴とは無関係にある一定の確率で文中に語が出現するという過程からの寄与を表すと解釈される．すなわち，定数項は語の出現回数を計数過程と考えると，ポアソン過程に対応する寄与である．したがって，すべての語の自己相関関数は過去の履歴に依存する拡張指数型のバースト項と，ポアソン過程から生じる定数項とが，その語独自の割合で足しあわされたものであると考えることができる．

（4）上記フィッティングモデル式における定数部分は，計数過程としてのポアソン過程をモデル化することにより導出可能であることが示された．すなわち，非概念語の自己相関関数は，確率過程としてポアソン過程を仮定することにより完全に再現可能であることが明らかとなった．

（5）上記フィッティングモデル式におけるバースト項，すなわち拡張指数型の自己相関関数部分は，確率過程として非定常ポアソン過程を拡張した Cox 過程において，特に強度関数を過去の語の出現履歴に依存する畳み込み積分で与えられることを仮定した確率過程のシミュレーションによって再現可能であることが示された．ただし，過去の語の出現履歴に依存した畳み込み積分は基本的に単調減少するため，強度関数がある閾値以下の値に減少した場合に，その値がある定数にリセットされるような仕組みを導入しないと，考えている語が文書中で複数回のバースト性を伴った出現パターンを示すことを再現することはできなかった．この仕組みを導入した語の出現パターンの確率過程シミュレーション結果およびその結果から求めた自己相関関数を図 3 に示す．この結果より，上記確率過程が拡張指数型自己相関関数で表現可能な自己相関関数を与えることが分かる．

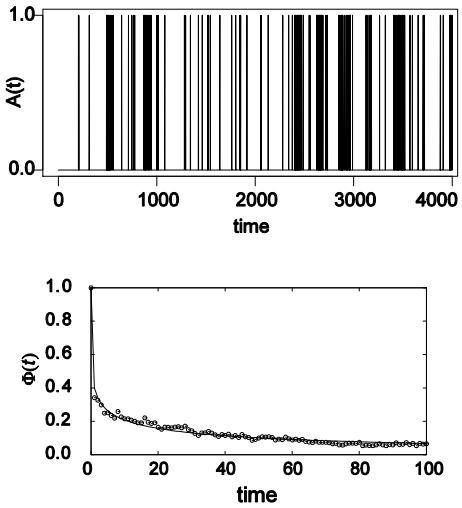

図 3　Cox 過程のシミュレーションにより生成された時系列データおよびその自己相関関数．ここで得られた自己相関関数は，図 1（c）の語 breed の自己相関関数とほぼ一致する．

（6）上記結果，特に（3）で述べた知見は，医療分野学生が提出するポートフォリオ等の文書分析にも有効に適用可能であることが示された．

＜引用文献＞

① Hiroshi Ogura, Hiromi Amano, and Masato Kondo. Feature selection with a measure of deviations from poisson in text categorization. Expert Systems with Appli-cations, 36(3):6826-6832, April 2009.
② Hiroshi Ogura, Hiromi Amano, and Masato Kondo. Distinctive characteristics of a metric using deviations from poisson for feature selection. Expert Systems with Applications., 37(3):2273-2281, March 2010.
③ Hiroshi Ogura, Hiromi Amano, and Masato Kondo. Comparison of metrics for feature selection in imbalanced text classication. Expert Systems With Applications, 38(5):4978-4989, 2011.
④ Hiroshi Ogura, Hiromi Amano, and Masato Kondo. Gamma-poisson distribution model for text categorization. ISRN Articial Intelligence, Vol.2013 (Article ID 829630), 2013.
⑤ 金明哲．テキストデータの統計科学入門．岩波書店，2009
⑥ Avik Sarkar, Paul H Garthwaite, and Anne De Roeck. A bayesian mixture model for term re-occurrence and burstiness. In Proceedings of the 9th Conference on Computational Natural Language Learning, pages 48-55, 2005.
⑦ Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter.

Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. CoRR,abs/0901.2349, 2009.
⑧ Jean Laherrere and Didier Sornette. Stretched exponential distributions in nature and economy:\fat tails" with characteristic scales. The European Physical Journal B Condensed Matter and Complex Systems, 2:525-539, 1998.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　３件）

① 小倉 浩, 天野 弘美, 近藤 雅人, 文書中の語の出現に関する自己相関関数と確率過程モデル, 昭和大学富士吉田教育部紀要, 査読無, 第８巻, 2013, 1-10
② Hiroshi Ogura, Hiromi Amano, Masato Kondo, Classifying Document with Poisson Mixtures, Transactions on Machine Learning and Artifical Intelligence, 査読有, 2, 2014, 48-76
DOI：10.14738/tmlai.24.2014
③ Tatsuo Shirota, Takaaki Kamatani, Tetsutaro Yamaguchi, Hiroshi Ogura, Kotaro Makii Satoru Shintani, Effectiveness of piezoelectric surgery in reducing surgical complications after bilateral sagittal split osteotomy, British Journal of Oral and Maxillofacial Surgery, 査読有, 52, 2014, 219-222
DOI: 10.1016/j.bjoms.2013.11.015

〔学会発表〕（計　５件）
① 小倉 浩他, 初年次学部連携 PBL チュートリアルおよび初年次体験実習の教育効果, 日本医学教育学会, 2014 年 7 月 19 日, 和歌山県立医科大学（和歌山県・和歌山市）
② 榎田めぐみ他, 医・歯・薬・保健医療学部による学部連携病棟実習の教育効果, 日本医学教育学会, 2014 年 7 月 19 日, 和歌山県立医科大学（和歌山県・和歌山市）
③ 片岡竜太他, 学部連携 PBL・病棟実習によるチーム医療教育の効果～アンケートの因子分析とポートフォリオの質的解析結果～, 日本医学教育学会, 2014 年 7 月 19 日, 和歌山県立医科大学（和歌山県・和歌山市）
④ 小倉 浩他, 初年次学部連携 PBL チュートリアルおよび初年次体験実習の相互教育効果, 日本保健医療福祉連携教育学会, 2014 年 9 月 20 日, 学生総合プラザ（新潟県・新潟市）
⑤ 今福輪太郎他, 初年次学部連携教育における学習過程の縦断的調査：ポートフォリオの質的分析から, 日本保健医療福祉連携教育学会, 2014 年 9 月 20 日, 学生総合プラザ（新潟県・新潟市）

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　０件）
○取得状況（計　０件）

〔その他〕
ホームページ等
語の自己相関関数の計算手法に関する研究,
https://sites.google.com/site/autocorrelation2014/

６．研究組織
(1)研究代表者
小倉　浩（OGURA, Hiroshi）
昭和大学・富士吉田教育部・准教授
研究者番号：40214100

(2)研究分担者：該当者なし

(3)連携研究者：該当者なし