平成 26 年度　卒業論文

2015 年 2 月 3 日

Web ニュース頻出単語に関連する単語の時系列解析

法政大学　理工学部　経営システム工学科

経営数理工学研究室

11X4146　横山朔

指導教員　五島洋行　教授

学科名	経営システム工	学籍番号	11X4146
申請者氏名		横山　朔	
指導教員 氏名		五島　洋行	

論文要旨

論文題目	Web ニュース頻出単語に関連する単語の時系列解析

本研究では，Web 上のニュースの本文中に出現する，特定の単語とそれに関連する単語を用いて時系列解析を行い，評価をする．Web 上の情報は時間が経つにつれて変化するため，膨大な情報から正確に必要な情報を得ることが難しくなる．テキストデータだけに焦点を当てても，検索エンジンを用いて特定の単語を検索した際に，特定の単語を含む文書は，時間の経過とともに変化し内容を変えていく．よって，文書がどのように変化していくのかを調べることは重要である．

そこで本研究では，Web 上のニュースサイトが配信するニュースの文章をデータとして扱い，時間の経過によって変化する文書の内容を調べるために，時系列解析をする．具体的に，ニュースの本文中に出現する，特定の単語とそれに関連する単語を用いて時系列解析を行う．関連する単語には共起語を用いる．なぜなら文書の内容が変われば，関連する単語も変化していくからである．

検証の結果から特定の単語が同じとき，分析期間が変われば共起語は必ずしも同じにならないことがわかった．また，共起語が同じでも，関連の高さの順番が違うこともわかった．さらに，共起語は時間の経過によって，バースト度の順位が変わることと，特定の日の共起語のバースト度の順位は，関連の高さの順位に従わないことがわかった．

目次

第1章 はじめに

1.1 研究背景

近年 Web 上の情報量が，急激に増えてきている．テキストや画像，動画，音声など様々な形式の情報が日々増加している．その要因の一つが，インターネットの普及である．図 1 は総務省が提供している資料で，インターネット普及率の推移[1]を表している．1997 年末の時点では，世帯と個人ともに大抵がインターネットを利用したことがなく，主に企業が利用していた．しかし，2013 年末の時点では，企業ではほぼ 100%利用したことがあり，世帯，個人でも 80%以上がインターネットを利用したことがある．図 2 のインターネット利用人口の推移[2]を見ても，2013 年末のインターネット利用人口は 1 億人を超えており，日本人の 8 割以上がインターネットを利用している．また，図 3 は Web 上で検索エンジンが検索可能な情報量[3]を表しており，2003 年と 2009 年を比較すると，約 6 倍近く情報量が増えていることがわかる．これらの事から，近年インターネットの普及が進み，Web 上の情報量が急激に増加していることがわかる．さらに，スマートフォンなどの出現によって，インターネットの利用の仕方が変わってきている．Twitter や Facebook などの文章を中心とした情報や，YouTube といった動画を中心とする情報など，様々な形式で個人が発信する情報が増えている．このようなことから，今後ますます情報量が増えることが予想される．

このようにして，Web 上の情報が増加するほど，メリットとしてインターネットを利用する人が，情報を手に入れ易くなる．一方，情報は時間が経過するにつれて変化するため，デメリットとして膨大な情報から正確に必要な情報を得ることが難しくなる．テキストデータだけに焦点を当てても，検索エンジンを用いて特定の単語を検索した際に，特定の単語を含む文書は，時間の経過によって内容を変えていく．これを解決するために，時間の経過とともに変化する Web 上のテキストデータを考慮し，解析する研究は多くなされている．よって，文書がどのように変化していくのかを調べることは重要である．

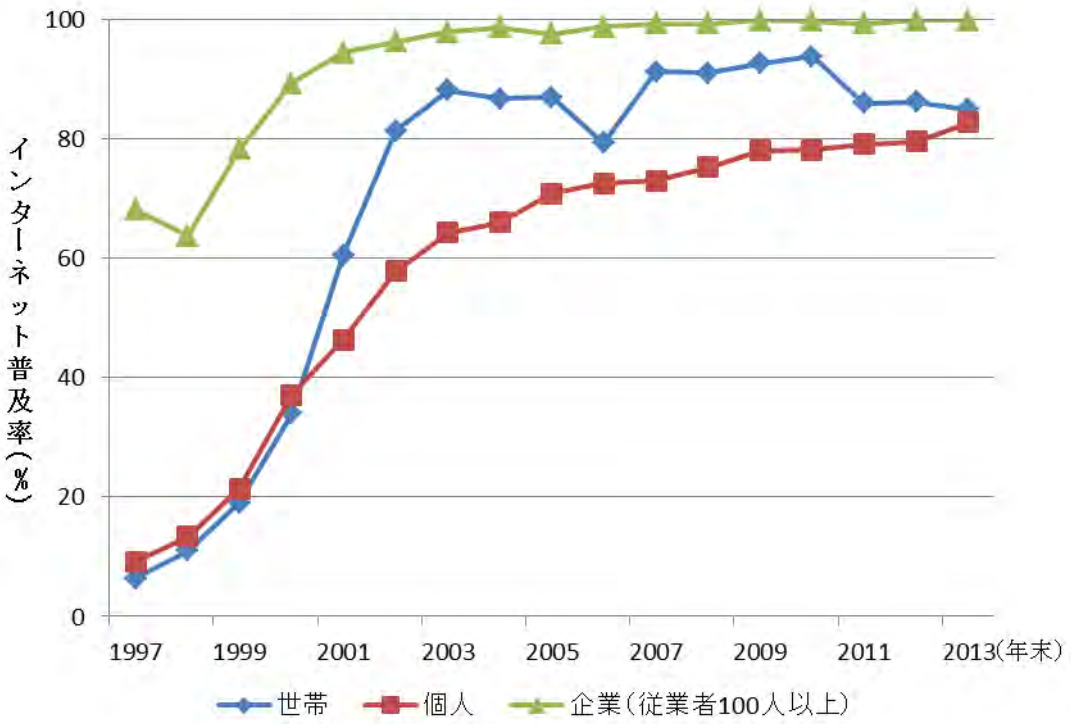

図 1. インターネット普及率の推移[1]

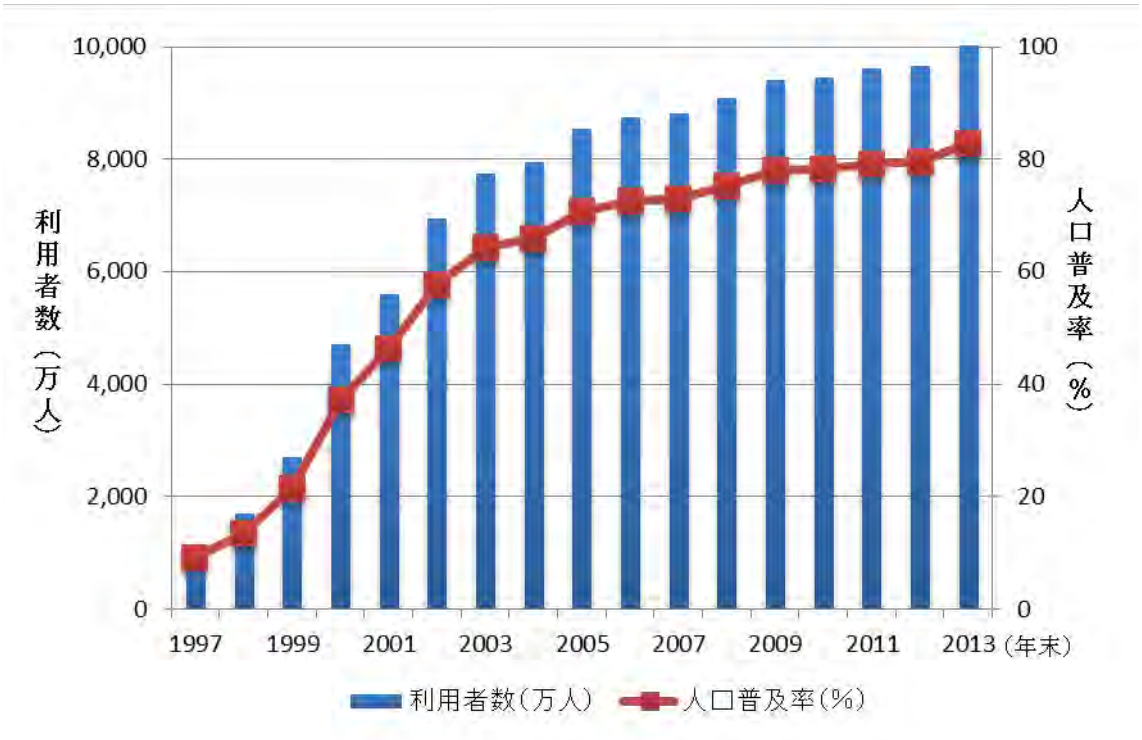

図 2. インターネット利用人口の推移[2]

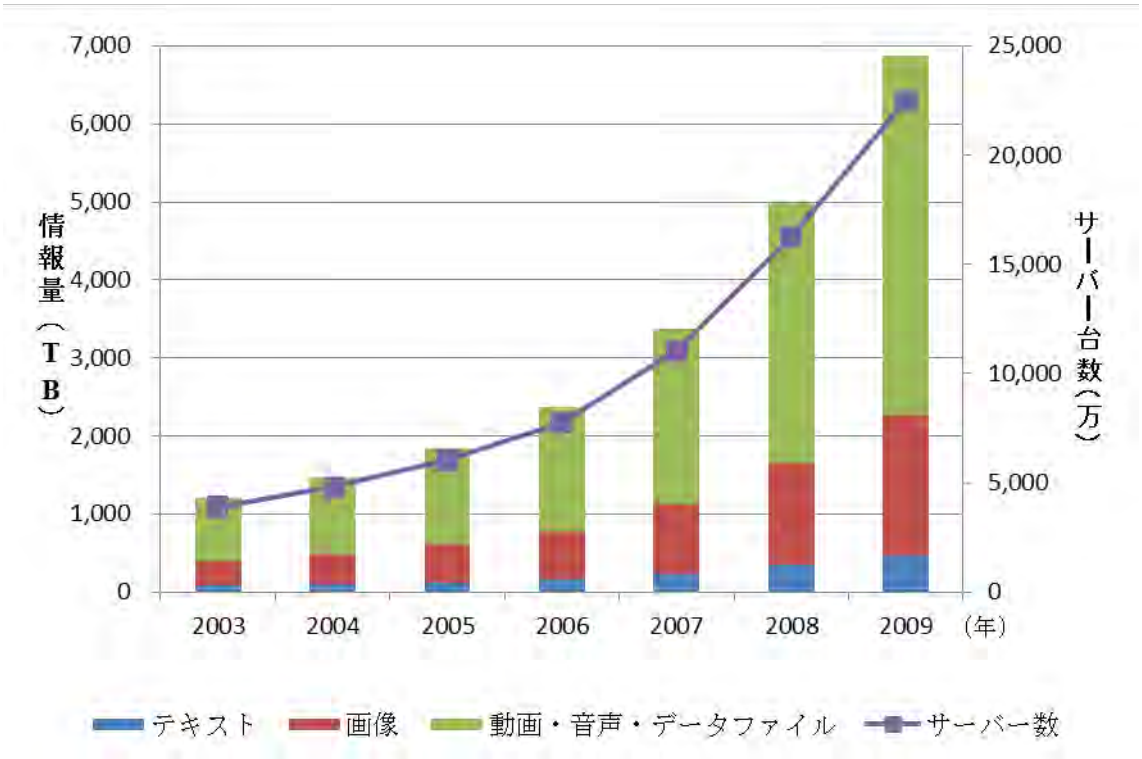

図 3. Web 上で検索エンジンが検索可能な情報量[3]

1.2 研究動機

1.1 では，近年の我が国のインターネット普及率と，情報量について着目し，時間の経過とともに変化する情報の解析の重要性について述べている．Web 上には様々な形式のデータが存在するが，時間の経過とともに変化する情報として，Web 上に存在するニュースサイトが配信するニュースを挙げる．ニュースは時間の経過とともに内容を変え，時事を知ることができる．このようなニュースを時系列で解析することは，正確に時事を捉え，特定のキーワードを用いれば，キーワードの変化を知ることができる．

そこで本研究では，Web 上に存在するニュースサイトが配信する，ニュースの文章をデータとして用い，時間の経過によって変化する文書の内容を調べるために，時系列解析をする．具体的に，ニュースの本文中に出現する，特定の単語とそれに関連する単語を用いて時系列解析を行い，評価することを目指す．なぜなら，特定の単語は変化しなくても，ニュースの内容が変われば，文書内の関連する単語が変化すると考え，特定の単語と関連する単語を時系列解析することで，時間の経過による文書の変化を捉えられると仮定したからである．

第2章 関連研究

本章では，本研究に関連のある研究を紹介する．本研究で用いるニューステキストデータについては，第 3 章 3.1 で詳しく述べるが，時間の経過によってニュースの内容が変化するという特徴がある．よって，ニューステキストデータを分析する上で，時間の経過による内容の変化を考慮した研究方法は非常に重要である．以下に，本研究で必要な考え方になる時間の経過による内容の変化を考慮した高橋佑介らの研究[4]と，福原らの研究[5]を挙げる．

2.1 関連研究の概要と特徴

高橋らの研究[4]

高橋らは，多種多様な情報が氾濫していることに，注目している．中でも Web 上の情報は爆発的に増えているため，情報の集約や俯瞰を行うことが重要だと考えている．そこで高橋らの研究[4]では，情報集約を行うために，Web 上のニュース記事にバースト解析とトピックモデルの 2 つの手法を組み合わせることによって，トピック集合のバーストを検出する方式を提案している．高橋らは，バーストの検出を行うために，J.Kleinberg のバースト解析アルゴリズムを用いている．これによって，文章中のキーワードのバースト期間と，非バースト期間とを自動で切り替えることを可能にしている．さらに各キーワードに対して，バースト度を付けることを可能にし，一定期間に関連の高い単語の出現が急増する瞬間を，求めることができる．トピックモデルには，分布を推定するために DTM を用いている．これによって，ニュース記事集合全体を，いくつかのトピック集合に分類し，トピック集合ごとに関連の高い単語を割り出すことを可能にしている．本研究では，トピック集合とそれに関連の高い単語の時系列解析するために，ニュース記事集合を DTM によって分類し，トピック集合とそれに関連の高い単語を，J.Kleinberg のバースト解析アルゴリズムを用いて，時系列解析を行うことを先行研究の特徴として捉える．

福原らの研究[5]

福原らは，時間の経過によって変化する情報に注目している．また多言語，実世界データとの関連，共起語を用いた焦点変化に関心がある．そこで福原らの研究[5]では，時系列テキストである Web 上の blog 記事, 新聞記事, メールマガジン等の本文中の文章を用いて，社会的関心の分析を行っている．分析方法として，(1)言語横断型関心分析，(2)感情表現を用いる分析，(3)共起語を用いる焦点変化に関する分析，(4)実世界データとの関連，(5)キーパーソンの関心分析の 5 つの分析方法を提案している．本研究では，特定の期間において Web 上の blog 記事対して, (3)の方法を用いて分析することを先行研究の特徴として捉える．

福原らは，問題の焦点は時間の経過とともに移り変わり，この問題に対する焦点変化を把握することが重要である，と述べている．そこで，焦点変化を把握するために(3)を用いて，日中韓 Web 上の blog 記事を対象に分析を行っている．共起語を検索するために，Disc 係数を用いて記事内における共起を計算している．分析対象期間を 2004 年 4 月 5 日から 5 月 30 日までとし，本文の中で期間において“イラク”と共起した単語の推移をグラフに表している．グラフは，y 軸を出現頻度とし，x 軸を時間としている．“イラク”に対し，最初は“人質”，“日本人”，“自衛隊”に焦点が当てられているが，時間が経過するにつれて“解放”，“自己責任”，“虐待”と焦点が変化していく．

2.2 関連研究の課題と本研究の方向性

2.1 では，関連研究として高橋らの研究[4]と，福原らの研究[5]を挙げた．この二つの研究には，本研究にとって課題がある．高橋らの研究[4]では，バースト解析とトピックモデルの二つの方法を用いて，トピック集合とそれに関連の高い単語でバースト解析を行っている．しかし，高橋らによって「経済」や「芸能」といったトピック集合に文書を分類し，それに関連の高い単語のバースト解析では，関連の高い単語の時系列解析を行うことは可能だが，特定の単語とそれに関連の高い単語のみの解析ができない．福原らの研究[5]では，特定の期間における共起語が，一日に何回出現するか回数を数え，出現回数で焦点が変化していくことを述べている．しかし，時間の経過による文書の内容の変化を調べるためには，共起語の出現回数だけではなく，特定の単語とそれに共起する単語の出現回数を求め，時系列解析を行う必要がある．

そこで本研究では，先行研究の課題を考慮した分析方法を提案する．具体的に，特定の期間を一年間とし，特定の単語と関連の高い単語の時系列解析を行う．時系列解析には，J.Kleinberg のバースト解析アルゴリズムを用いる．また，特定の単語に一年間のニューステキストデータから最も出現回数が多い単語を選び，それに関連の高い単語を共起語として分析する．詳しい分析手順については，第 4 章で述べる．

第3章 関連知識

本章では，本研究で必要とする分析手順に関する，関連の知識について述べる．本研究では，ニュース記事の内容をデータとして用いており，文章を単語に分割し，単語と単語の関連の高さを調べる．データや分割方法や単語と単語の関連や特徴を節ごとに述べる．

3.1 ニューステキストデータ

本研究では，扱うテキストデータを Web 上に存在する，Yahoo!ニュースなどのニュースサイトが配信するニュースの文章に限定する．またテキストデータを MySQL 上に集め，これをニューステキストデータとする． MySQL とは Web サーバーのバックエンドとして，広く使用されている RDBMS，つまりリレーショナルデータベース管理システムのことである[6]．ニューステキストデータは，ニュースのタイトル，本文，配信された時間で構成されている．特徴は最新の出来事から，過去の事実までを文章にしてまとめているため，キーワードを文書中から検索した際に，時間ごとの情報を得ることができる．データは 2008 年 3 月 3 日から 2013 年 12 月 31 日までに，日本の各ニュースサイトが配信した 1,906,767 件のニュースである．表 1 は集めたニューステキストデータの配信元である．

3.2 形態素解析

形態素とは，意味を持つ最小の文字列の単位のことである．形態素解析とは，文を単語ごとに分割し，品詞情報などを付け加える作業である[7]．本研究で，形態素解析環境である MeCab を用いて，対象とするニューステキストデータに形態素解析を行い，意味をなす最小単位の単語に文章を分割し，単語を抽出することで分析を行う．

MeCab とは，形態素解析を行うために，京都大学情報学研究科と NTT コミュニケーション科学基礎研究所の共同プロジェクトを通じて，工藤拓氏が開発した環境である．大きな特徴は，辞書とコーパスに依存しない凡庸的な設計と，同じ形態素解析環境である茶筌，JUMAN より処理速度が高速な点である[7]．MeCab を用いて形態素解析を行うと，品詞，品詞細分類 1~3，活用形，活用型，原形，読み，発音の順で単語に情報を与えられる[9]．本研究では品詞，品詞細分類 1~3 のみの情報を用いて，研究を行う．表 1 は形態素解析を行い，単語に品詞情報を付ける場合の一覧である．ただし，品詞情報は No.0~68 の 69 通りあるため，本研究では必要な部分だけを表 2 であらわす．

3.3 共起語

単語と単語の関連の高さを表す指標として，共起語がある．共起語とは，文章中の特定の単語に対して，同じページ内で別のある単語が頻繁に共起して出現する単語のことである．例えば，政治家の選挙演説で，「国民」の後に「の」を挟んで「皆さま」が続くパターンが多いが，この場合「国民」と「皆さま」は共起している[8]．このように，「国民」を特定の単語とした時，「皆さま」は共起語である．

表 1.ニューステキストデータの配信元一覧

配信元
ロイター
YOMIURI ONLINE
Yahoo!ニュース
gooニュース
NHKニュース
朝日新聞デジタル
毎日新聞
産経ニュース
J-CASTニュース
CNET Japan

表 2. 形態素解析を行い，品詞に情報を付けた場合の一覧[9]

No.	品詞	品詞細分類1	品詞細分類2	品詞細分類3
36	名詞	サ変接続	*	*
37	名詞	ナイ形容詞語幹	*	*
38	名詞	一般	*	*
39	名詞	引用文字列	*	*
40	名詞	形容動詞語幹	*	*
41	名詞	固有名詞	一般	*
42	名詞	固有名詞	人名	一般
43	名詞	固有名詞	人名	姓
44	名詞	固有名詞	人名	名
45	名詞	固有名詞	組織	*
46	名詞	固有名詞	地域	一般
47	名詞	固有名詞	地域	国
48	名詞	数	*	*
49	名詞	接続詞的	*	*
50	名詞	接尾	サ変接続	*
51	名詞	接尾	一般	*
52	名詞	接尾	形容動詞語幹	*
53	名詞	接尾	助数詞	*
54	名詞	接尾	助動詞語幹	*
55	名詞	接尾	人名	*
56	名詞	接尾	地域	*
57	名詞	接尾	特殊	*
58	名詞	接尾	副詞可能	*
59	名詞	代名詞	一般	*
60	名詞	代名詞	縮約	*
61	名詞	動詞非自立的	*	*
62	名詞	特殊	助動詞語幹	*
63	名詞	非自立	一般	*
64	名詞	非自立	形容動詞語幹	*
65	名詞	非自立	助動詞語幹	*
66	名詞	非自立	副詞可能	*
67	名詞	副詞可能	*	*

第4章 分析手順

本章では，研究の方向性から必要な分析の手順を述べた後に，主な分析の手法などを説明する．まず，3.1 で述べたニューステキストデータを，1 月 1 日から 12 月 31 日までの一年ごとに分け，一年ごとのデータに形態素解析を行った後に，出現頻度分析を行う．ただし 2008 年は，3 月 3 日から 12 月 31 日までとする．次に一年ごとのテキストデータに対して，クリーニングを行う．出現頻度分析で得られる結果から特定の単語に最頻出単語を選び，選んだ単語に対する共起語分析を行う．共起語分析で得られる結果から上位 3 単語を選び，共起語の時系列解析を行う．この手順によって得られる結果については次の章で述べる．

4.1 出現頻度分析

出現頻度分析とは，ニューステキストデータに形態素解析を行い，文章を単語ごとに分割し，単語の出現頻度を分析することである．本研究で出現頻度分析を行う際には，単語を固有名詞に限定し，最も出現する単語を調べ，これを最頻出単語とする．限定した理由は，最頻出単語を選ぶ際に，句読点いった記号や助詞などの単語が多く，単語だけでは意味が分かりにくいからである．また分析する上で，単語の意味をわかりやすくするために固有名詞を選ぶ．ただし，MeCab の品詞情報 No.46，47 の固有名詞と地域は除く[9]．除いた理由は，上位単語に国名ばかりが出現し，その後の分析で様々なニュースと関連してしまい，単語がどのニュースと関連しているのか，わからなくなるからである．

4.2 テキストのクリーニング

テキストのクリーニングとは，テキストを統計的に分析するために，分析する対象を電子化し，データの形式をそろえたり，不必要なものを削除したりする作業のことである[7]．本研究では，共起語分析の際に，共起する単語に句読点や（）といった記号の出現を防ぐために，テキストのクリーニングを行う．これによって特定の単語に対する，共起語の上位に出現する単語の意味を捉えやすくする．

4.3 共起語分析

共起語の分析方法として，コロケーションがある．コロケーションとは，単語と単語の結びつきの強さに着目することである[8]．文章中の特定の単語を中心に，前後何語かの範囲を指定することで，特定の単語に対しての共起語を特定することができる．特定の単語と共起語の結びつきの強さを表すために，*T* と *MI* を用いる．*T* は式(1)で表すことができ，共起する頻度を総語数や期待値度数で調整した値である．実測値とは，特定の単語に対しての共起語の頻度である．

$$T = \frac{\text{実測値} - \text{期待値}}{\text{実測値の平方根}} \tag{1}$$

MI は式(2)で表し，共起の頻度を共起語の期待値で割り，2 を底にした対数を取った値である．

$$MI = \log_2 \frac{\text{共起回数}}{\text{共起語の期待値}} \tag{2}$$

T の取る値は，全体の頻度や期待値を考慮しているため，頻度が多くても値が大きくなるとは限らない．*T* は絶対値 2 を超えると，有意に出現回数が偏っていると判断できる．一方 *MI* の取る値は，低頻度の単語であっても共起関係を抽出できるという特徴がある．

4.4 時系列解析

本研究で時系列解析とは，2008 年 3 月 3 日から 2013 年 12 月 31 日までの時間の経過によって，ニューステキストデータの内容の変化を分析することである．ニューステキストデータに時系列解析を行うために，J.Kleinberg のバースト解析アルゴリズムを用いる[10]．バーストとは，一定期間に特定の単語の出現が，急激に増加する現象のことである．このアルゴリズムの特徴は，特定の期間のニューステキストデータに対して，単語のバースト状態と非バースト状態を求めることができ，単語に対してバーストの度合を付けること可能にすることである．ニューステキストデータは，離散時間でデータを収集するため，enumerating バーストを用いる[10]．解析期間において各時刻$t = 1,2,\cdots,n$としたとき，特定の単語を含むすべての文書集合を D とし，$D = \sum_{t=1}^{n} d_t$で表す．また共起語を含む関連の文書集合を R とし，$R = \sum_{t=1}^{n} r_t$で表す．式(3)は非バースト状態を表している．

$$p_0 = \frac{R}{D} \tag{3}$$

式(4)は式(3)にパラメータsをかけた，バースト状態を表している．ただし$s > 1$を満たすパラメータであり，$p_1 \leq 1$でなければならない．

$$p_1 = p_0 s \tag{4}$$

式(5)はコスト関数を表す．コスト関数は，二項分布に従い，

$$\sigma(i, r_t, d_t) = -\log\left[\binom{d_t}{r_t} p_i^{r_t}(1-p_i)^{d_t-r_t}\right] \tag{5}$$

と表すことができる．式(6)はバースト度の式である．式(6)に，式(4)を代入することで，共起語のバーストの度合である，バースト度を求めることができる

$$\sum_{t=t_1}^{t_2}\left(\sigma(0, r_t, d_t) - \sigma(1, r_t, d_t)\right)$$

$$= -\log\left[\left(\frac{d_t}{r_t}\right) {p_0}^{r_t}(1-p_0)^{d_t-r_t}\right] - \left(-\log\left[\left(\frac{d_t}{r_t}\right) {p_1}^{r_t}(1-p_1)^{d_t-r_t}\right]\right)$$

$$= -\log\left[\left(\frac{d_t}{r_t}\right) \times \left(\frac{R}{D}\right)^{r_t} \times \left(1-\frac{R}{D}\right)^{d_t-r_t}\right] - \left(-\log\left[\left(\frac{d_t}{r_t}\right) \times \left(\frac{R}{D}s\right)^{r_t} \times \left(1-\frac{R}{D}s\right)^{d_t-r_t}\right]\right) \tag{6}$$

第5章 分析結果

本章では，第 4 章の分析手順で得られるそれぞれの結果について述べ，結果を図や表で表す．また，最頻出単語と共起語の時系列解析によって得られる結果を比較し，共起語が時間の経過によってどのように変化するのか，検証を行う．

5.1 頻度分析結果

表 3 は，一年ごとのニューステキストデータから得た，単語の出現頻度分析結果である．2008 年から 2013 年までの，一年毎のニューステキストデータのタイトルから，単語の出現頻度の高い順に表している．表 3 から，2012 年と 2011 年はともに「東電」という単語が最も出現している．また，2009 年と 2008 年は「大リーグ」という単語が最も出現していることがわかる．最頻出単語では出ないが，「自民」という単語が 2008 年を除いて毎年出現している．表 4 は，一年ごとの最頻出単語の出現回数である．表 4 から，2010 年の「iPhone」が 4,794 回で，6 年間で最も出現回数が多いことがわかる．また，2009 年の「大リーグ」が 1,279 回で，6 年間で最も出現回数が少ない．6 つの最頻出単語は，全て MeCab の品詞情報 No.45 に分類される．

5.2 共起語分析結果

表 5 は，一年ごとの出現頻度分析から得た，最頻出単語に対しての共起語分析結果である．最頻出単語に対する共起語を求め，関連の高さを上から順に表している．2013 年を例に，Before，After，Span，Total，*T*，*MI* について説明する．Before と After は，文章中の「日経」を中心にしたとき，共起語が前後どちらで出現したかを回数で表している．Span は，文章中の「日経」を中心にして特定の範囲内で，出現した回数を表している．よって，Before と After の出現回数を足すと，Span の出現回数となる．本研究では，範囲を文章中の最頻出単語を中心に，前後二単語とした．また Total は，文章中の特定の範囲外を含めた出現回数である．これらの Before，After，Span，Total をもとに，4.2 で述べたコロケーションを用いて *T* と *MI* を求め，関連の高さの順を求めた．最頻出単語が「日経」である 2013 年の結果を見ると，共起語である「平均」が全体で 4,179 回出現しており，そのうち 3,680 回は範囲内で出現しているため，*T* と *MI* がともに高い値を取っている．2012 年と 2011 年の共起語を比較すると，最頻出単語が同じ「東電」であっても，共起する単語が違うことがわかる．また，2009 年と 2008 年を比較すると，最頻出単語と共起語は同じだが，関連の高さの順番が異なることがわかる．

表 3. 一年ごとの出現頻度分析結果

年 / 順位	2013	2012	2011	2010	2009	2008
1	日経	東電	東電	iPhone	大リーグ	大リーグ
2	参院	自民	小沢	小沢	自民	Google
3	日銀	小沢	iPhone	W杯	鳩山	ドコモ
4	自民	日経	JR	菅	W杯	巨人
5	JR	衆院	自民	自民	イチロー	トヨタ

表 4. 最頻出単語と出現回数

年	2013	2012	2011	2010	2009	2008
単語	日経	東電	東電	iPhone	大リーグ	大リーグ
出現回数	3,715	2,773	3,523	4,794	1,279	2,310

表 5. 一年ごとの共起語分析結果

年	共起語	Before	After	Span	Total	*T*	*MI*
2013	平均	17	3,663	3,680	4,179	60.34	7.552
	円	1,143	126	1,269	23,472	32.53	3.526
	終値	3	726	729	1,944	26.66	6.321
2012	値上げ	78	324	402	1,008	19.82	6.444
	賠償	70	111	181	1,115	13.07	5.147
	殺害	1	182	183	2,347	12.73	4.089
2011	賠償	52	275	327	1,279	17.46	4.856
	停電	165	149	314	1,958	16.74	4.183
	社長	8	235	243	1,844	14.54	3.900
2010	アプリ	18	129	147	573	12.02	6.817
	向け	8	77	85	1,918	8.746	4.283
	アップル	74	6	80	846	8.729	5.377
2009	安打	152	40	192	856	13.61	5.829
	松井	6	172	178	926	13.07	5.606
	イチロー	23	152	175	1,109	12.90	5.321
2008	安打	358	79	437	1,866	20.63	6.256
	イチロー	33	256	289	1,274	16.77	6.210
	松井	2	210	212	1,099	14.33	5.976

5.3 時系列解析結果

図 4 から図 12 は，バースト解析アルゴリズムを用いて，最頻出単語とそれぞれの共起語の時系列解析を行った結果である．全ての期間において，パラメータsは$s = 1.01$とした．理由は 2013 年のパラメータは，$s = 1.01$が最適であったので，他の期間の時系列解析結果と比較するために，パラメータの値を統一したからである．図 4 は，2013 年「日経」と共起語の時系列解析を行った結果である.「日経」をすべて含む全体の文書は 3,724 件あった．それに対し「平均」,「円」,「終値」を含む関連文書はそれぞれ 3,662 件， 2,609 件， 839 件であった．図 5 は，2013 年 7 月 1 日から 7 月 31 日までの，時系列解析結果である．図 5 から最もバーストしなかったのは，2013 年 7 月 3 日に「平均」を共起語に選んだときに，約-5.010 の値であったことがわかる．同日の「円」を共起語に選ぶときは 0.006917 で,「終値」を共起語に選んだときは 0.01540 の値であった．この結果から，全体の文書集合に対して，関連の文書集合が多いと，頻繁に「日経」と「平均」が現れるので，正の値を取るバーストをすることがなく，負の値を取るバーストが起きることがわかった．

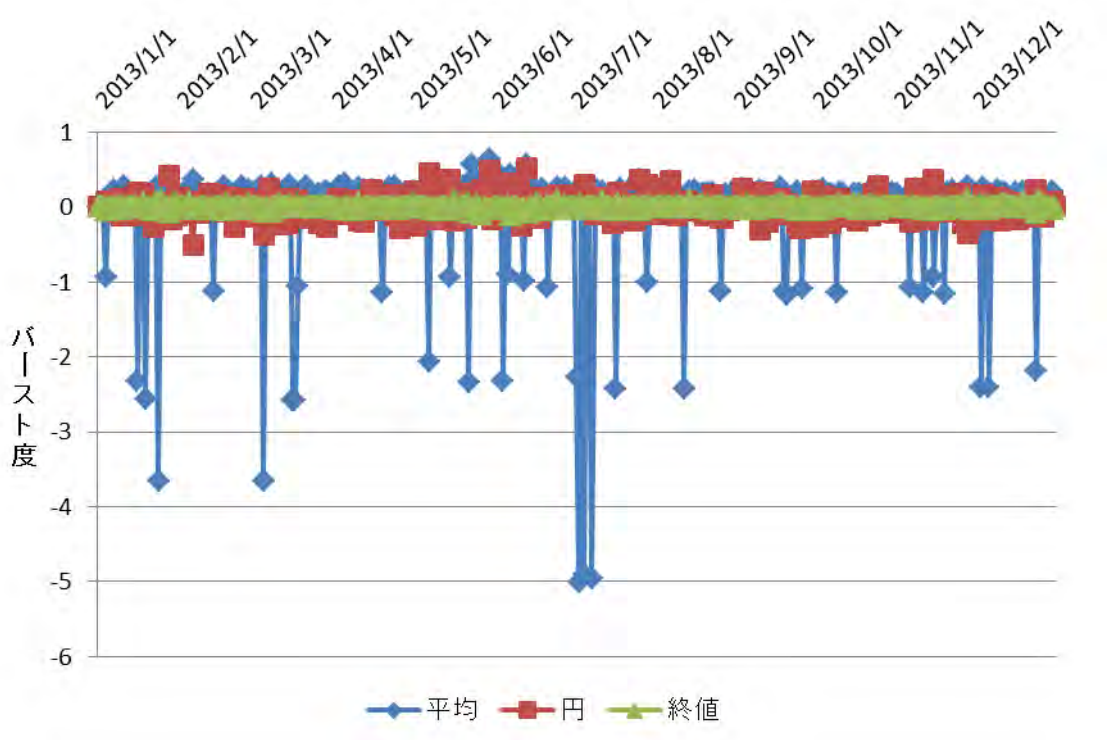

図 4. 2013 年「日経」と共起語の時系列解析結果

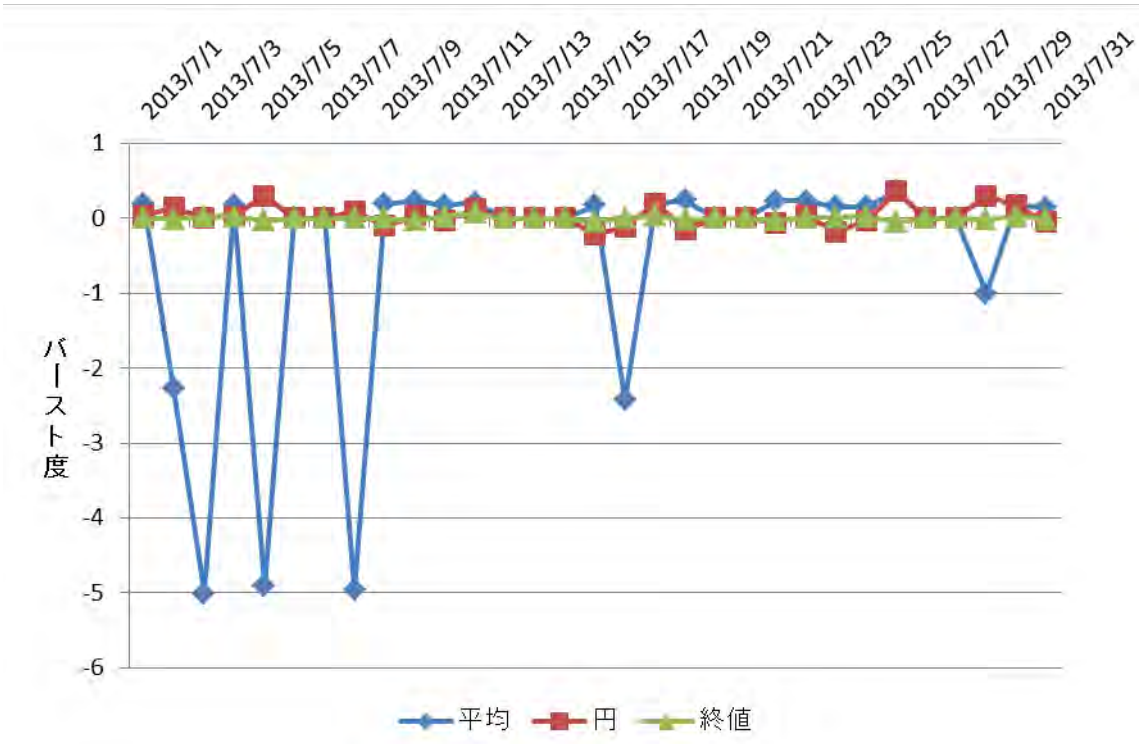

図 5. 2013 年 7 月 1 日から 7 月 31 日までの時系列解析結果

2013 年同様に，2012 年の結果を評価する．図 6 は，2012 年「東電」に対する共起語の時系列解析を行った結果である．「東電」をすべて含む全体の文書は 2770 件あった．それに対し「値上げ」，「賠償」，「殺害」を含む関連文書はそれぞれ 481 件， 262 件， 182 件であった．図 7 は，2012 年 6 月 1 日から 8 月 31 日までの，時系列解析結果である．図 7 から，最もバーストしたのは，2012 年 6 月 7 日に「殺害」を共起語に選んだときに，約 0.4636 の値であったことがわかる．同日の「値上げ」を共起語に選んだときは 0.008624 で，「賠償」を共起語に選んだ時は-0.06402 の値であった．反対に，最もバーストしなかったのは，2012 年 8 月 6 日に「値上げ」を共起語に選んだときに，約-0.1426 の値であったことがわかる．同日の「賠償」を共起語に選んだ時は-0.02244 で，「殺害」を共起語に選んだ時は-0.001590 の値であった．

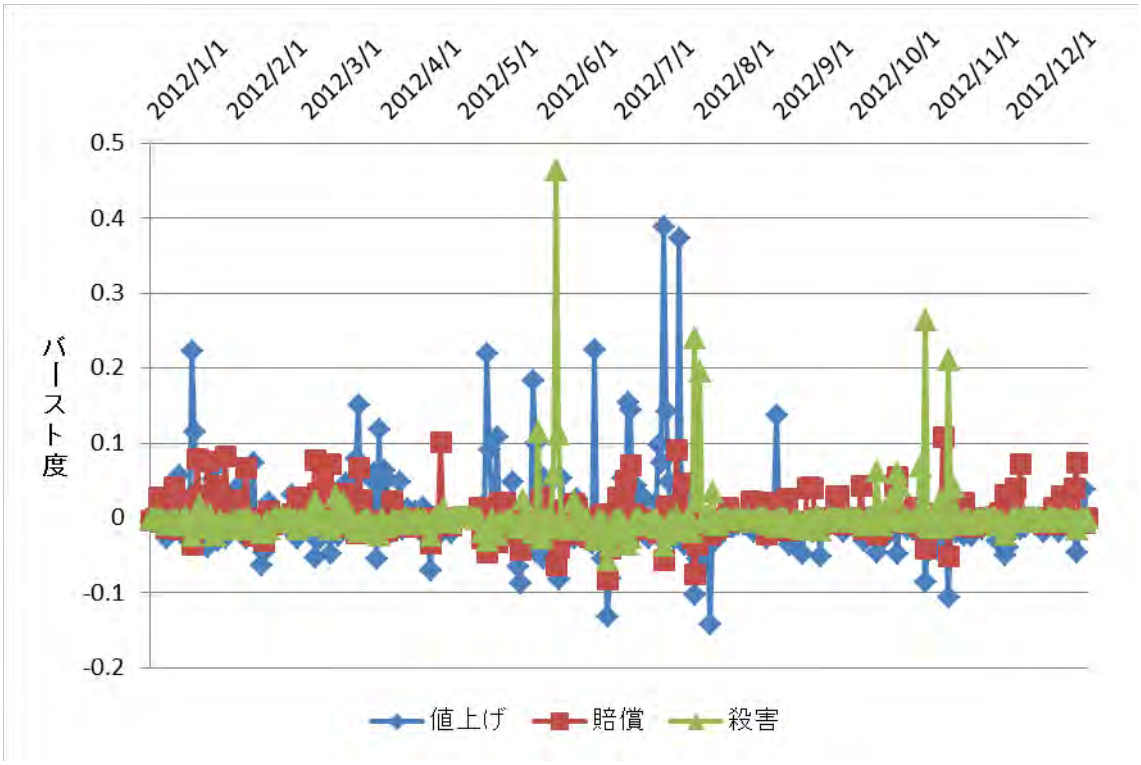

図 6. 2012 年「東電」と共起語の時系列解析結果

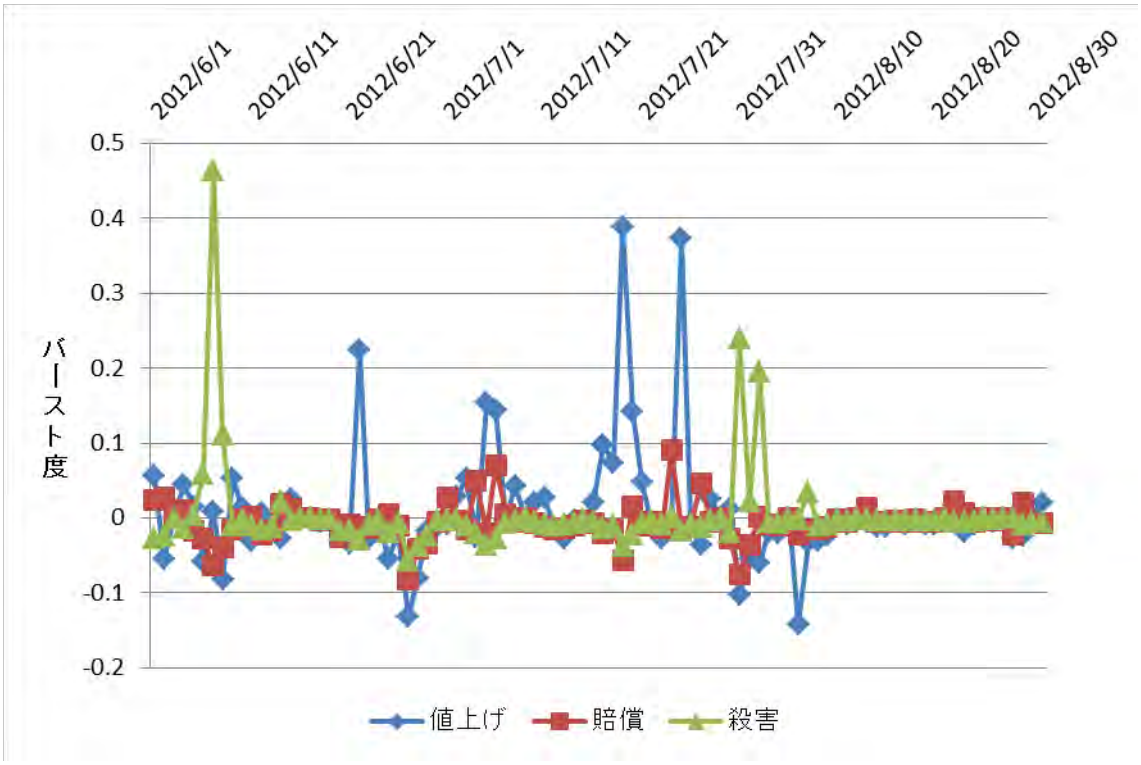

図 7. 2012 年 6 月 1 日から 8 月 31 日までの時系列解析結果

次に，2011 年の結果を評価する．図 8 は，2011 年「東電」に対する共起語の時系列解析を行った結果である．「東電」をすべて含む全体の文書は 3,518 件あった．それに対し「賠償」，「停電」，「社長」を含む関連文書はそれぞれ 399 件， 378 件， 251 件であった．図 9 は，2011 年 3 月 1 日から 3 月 31 日までの，時系列解析結果である．図 9 から，最もバーストしたのは，2011 年 3 月 16 日に「停電」を共起語に選んだときに，約 0.8065 の値であったことがわかる．同日の「賠償」を共起語に選んだときは-0.1348 で，「社長」を共起語に選んだときは-0.08094 の値であった．2012 年と 2011 年の最頻出単語と，共起語である「賠償」は同じであった．しかし，「賠償」の最も高いバースト度は，2012 年では約 0.1 の値を取ったが，2011 年では約 0.2 の値を取っている．このことから，時間によってバーストする度合いが違うことがわかる．

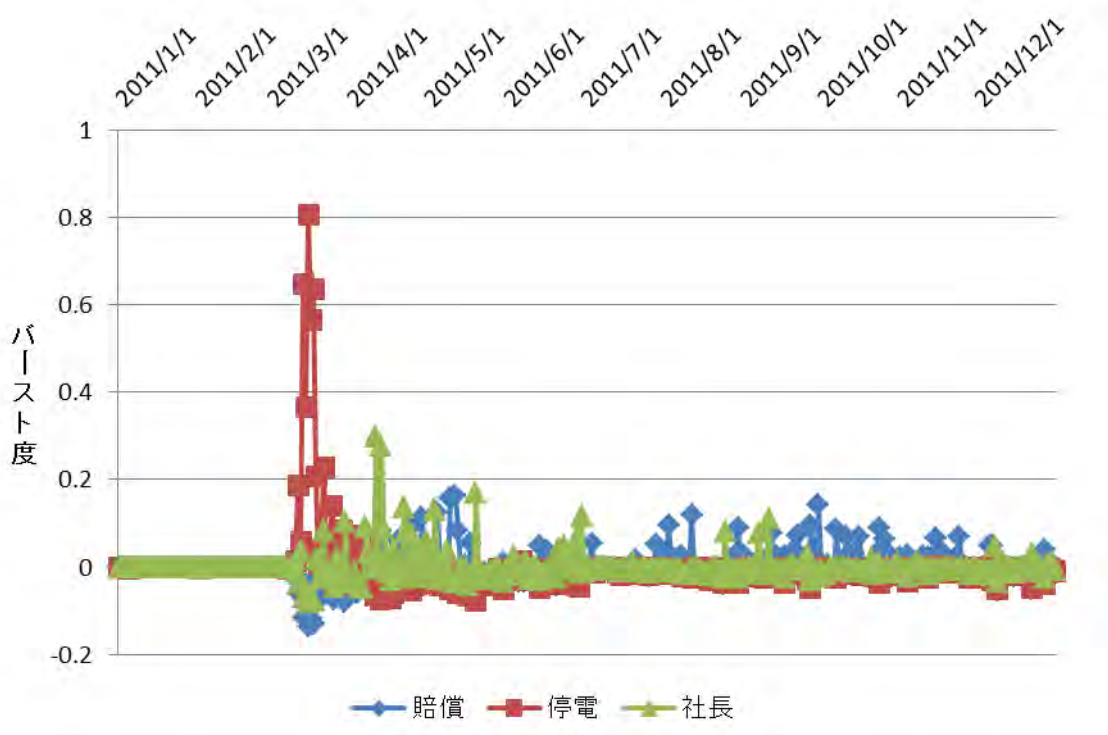

図 8. 2011 年「東電」と共起語の時系列解析結果

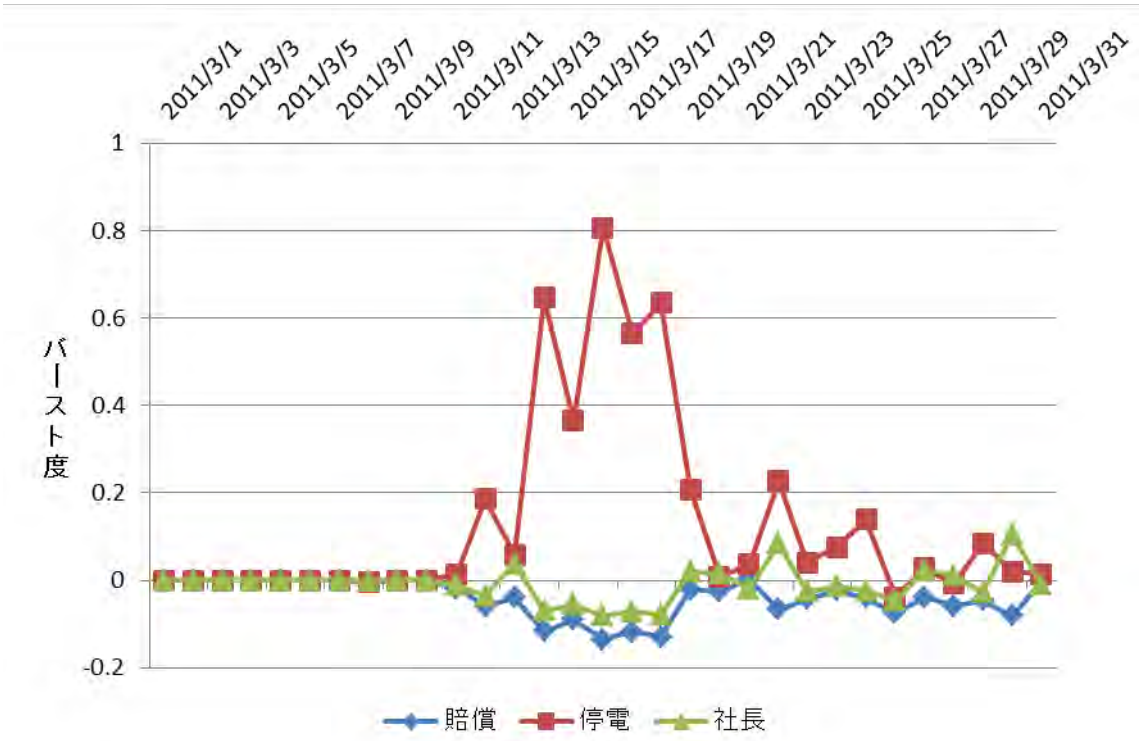

図 9. 2011 年 3 月 1 日から 3 月 31 日までの時系列解析結果

最後に，2010 年から 2008 年の結果を評価する．図 10 は「iPhone」と共起語「アップル」の時系列解析結果である．2010 年では「iPhone」を含む全体の文書 5,504 件に対して，「アップル」を含む関連文書 140 件であり，値は 0.1379 で最もバーストした．図 11 は「大リーグ」と共起語「松井」の時系列解析結果である．2009 年では「大リーグ」を含む全体の文書 1,276 件に対して，「松井」を含む関連文書 263 件であり，値は 0.1292 で最もバーストした．図 12 は「大リーグ」と共起語「安打」の時系列解析結果である．2008 年では「大リーグ」を含む全他の文書 2,308 件に対して，「安打」を含む関連文書は 622 件であり，値は 0.1669 で最もバーストした．値はどれも 0.2 を下回っており，2013 年から 2011 年の最もバーストした瞬間の値と比べると，値が小さいことがわかる．

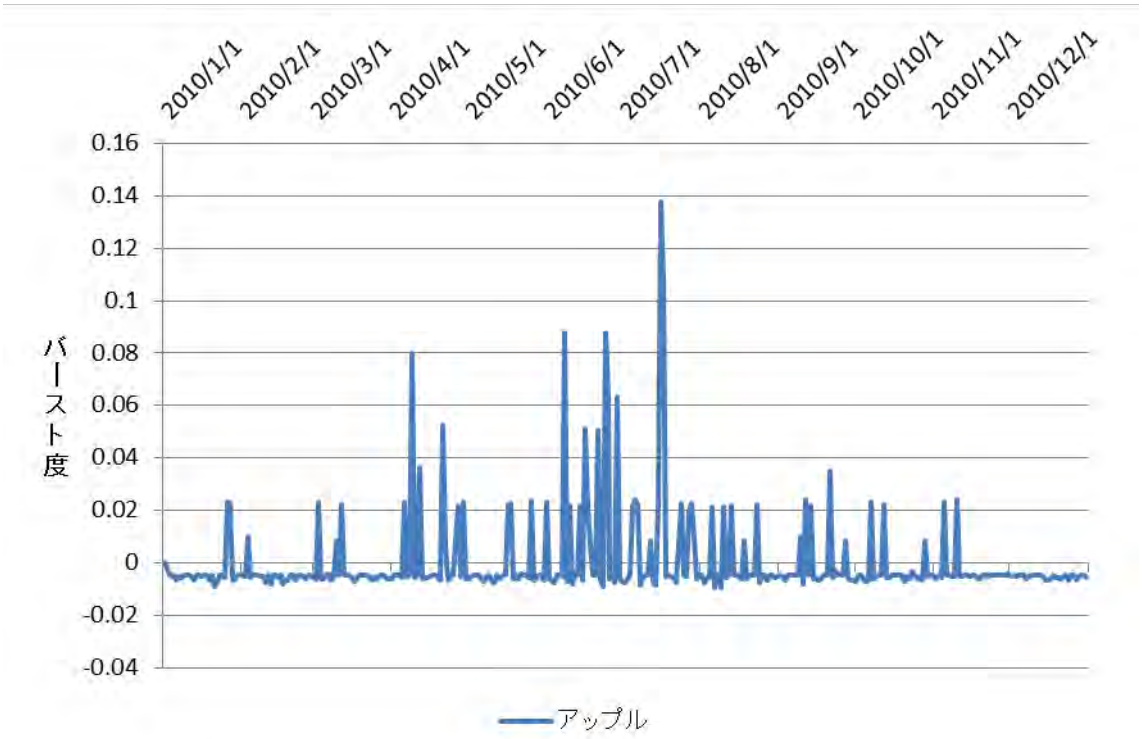

図 10. 2010 年「iPhone」と共起語「アップル」の時系列解析結果

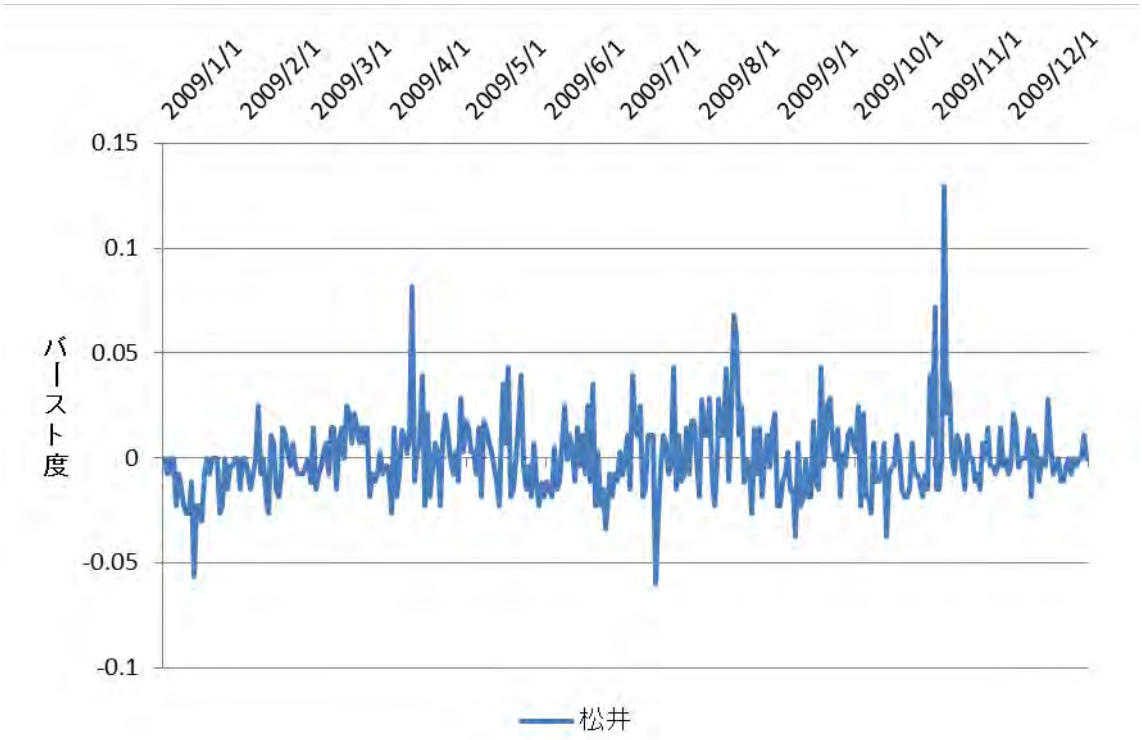

図 11. 2009 年「大リーグ」と共起語「松井」の時系列解析結果

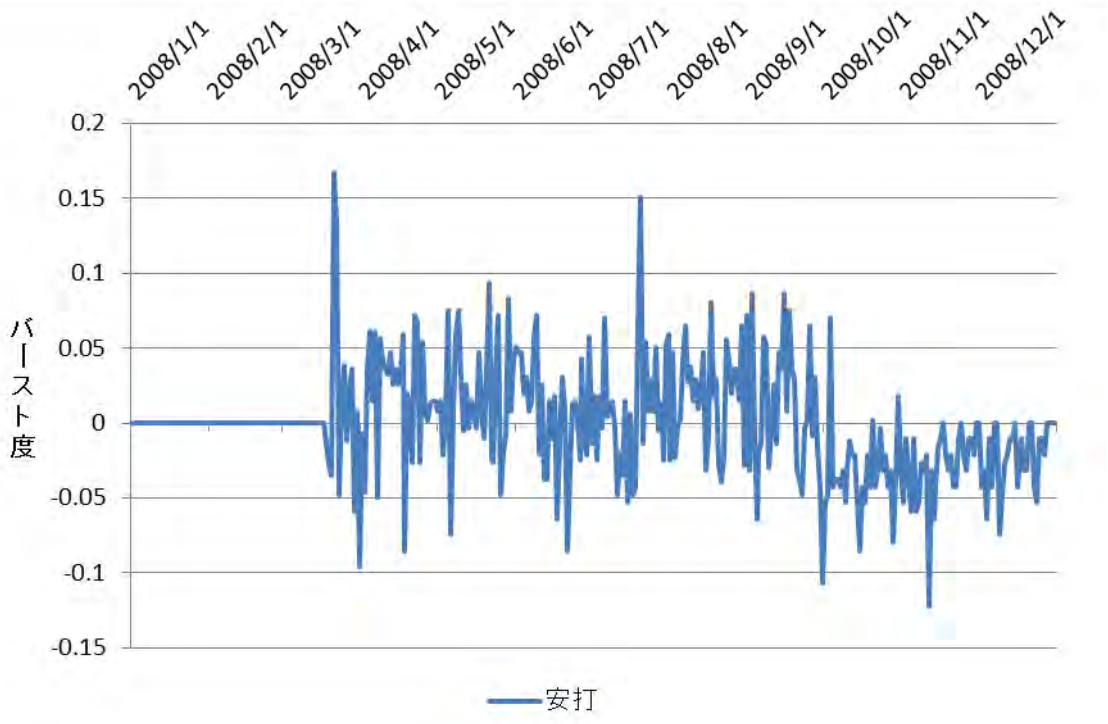

図 12. 2008 年「大リーグ」と共起語「安打」の時系列解析結果

これらのことから，共起語は時間の経過によって，バースト度の順位が変わることがわかった．また共起語分析結果と比較して，特定の日の共起語のバースト度の順位は，関連の高さの順位に従わないことがわかった．

第6章 おわりに

本研究では，Web 上のニュースサイトが配信するニュースの文章を用いて，特定の単語は変化しなくても，時間の経過によってニュースの内容が変われば，文章内の関連する単語が変化していくと仮説を立て，検証を行った．具体的に，特定の期間でニュースの本文中に出現する，特定の単語と関連する単語を用いて時系列解析を行うことで，検証を行った．まず，長期的な変化を調べるために特定の期間を一年にし，約 6 年分のニュースステキストデータを一年ごとに分割した．次に，一年ごとのニュースステキストデータに形態素解析を行うことで，文章に意味を持つ最小の単語に分割した．分割されたニュースステキストデータから最頻出単語を求め，これを特定の単語にした．次に，関連する単語に共起語を用い，共起語分析をコロケーションという手法を用いて，共起語を関連の高い順に求めた．これらの結果から，最頻出単語を含む全体の文書集合と，共起語を含む関連の文書集合にまとめ，J.Kleinberg のバースト解析アルゴリズムを用いて，共起語のバーストを求めた．

これらの手順によって得られた結果については，第 5 章で述べた．共起語分析結果から，関連の高さの順位を求めた．2012 年と 2011 年や 2009 年と 2008 年のように最頻出単語が同じとき，共起語は必ずしも同じにならないことがわかった．また，共起語が同じでも，関連の高さの順番が違うこともわかった．時系列解析結果からは，共起語は時間の経過によって，バースト度の順位が変わることがわかった．また共起語分析結果と比較して，特定の日の共起語のバースト度の順位は，関連の高さの順位に従わないことがわかった．

本研究では，仮説を立て検証を行ったが，情報は時間が経つにつれて変化するため，インターネットを利用する人が，正確な情報にたどり着くことができない，という課題の解消方法を提示したわけではない．今後の研究で，インターネットを利用する人が，時間の経過による情報の変化に関係なく，目的の情報にたどり着くための方法を提案する．また，分析では得手の単語の品詞を限定してしまい，他の品詞を考慮することができなかった．共起語を特定する手法もいくつかあるので，比較する必要がある．よって，これらを考慮した研究を今後行う必要がある．

参考文献

[1] 総務省:情報通信白書　インターネット普及率の推移　平成 26 年版,
http://www.soumu.go.jp/johotsusintokei/field/tsuushin01.html（2015 年 1 月 25 日確認）

[2] 総務省:情報通信白書　インターネット利用人口の推移　平成 26 年版,
http://www.soumu.go.jp/johotsusintokei/field/tsuushin01.html（2015 年 1 月 25 日確認）

[3] 総務省:インターネット検索エンジンの現状と市場規模等に関する調査　平成 21 年版,
http://www.soumu.go.jp/main_content/000035044.pdf（2015 年 1 月 25 日確認）

[4] 高橋祐介，横本大輔，宇津呂武仁，吉岡真治，河田容泳，神田典子，福原知広，中川裕志，清田陽司:時系列トピックモデルにおけるバーストの同定，第 4 回データ工学と情報マネジメントに関するフォーラム，DEIM フォーラム，F5–5(2012)

[5] 福原知広，中川裕志，西田豊明:時系列テキスト集合からの社会的関心の分析，第 16 回インテリジェントシステム・シンポジウム講演論文集，pp.51–56（2006）

[6] 遠藤俊裕，坂井恵，舘山聖司，鶴長鎮一，とみたまさひろ，班石悦夫，松信嘉範:
MySQL 徹底入門三版～5.5 新機能対応～，翔泳社（2013）.

[7] 金明哲:テキストデータの統計科学入門，岩波社（2009）

[8] 石田基広，小林雄一郎:R で学ぶ日本語テキストマイニング，ひつじ書房社（2013）

[9] MeCab:品詞 ID の定義,
http://mecab.googlecode.com/svn/trunk/mecab/doc/posid.html（2015 年 1 月 25 日確認）

[10] J.Kleinberg:“Bursty and Hierarchial Structure in Streams”, Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining，pp91–101(2002)

付録

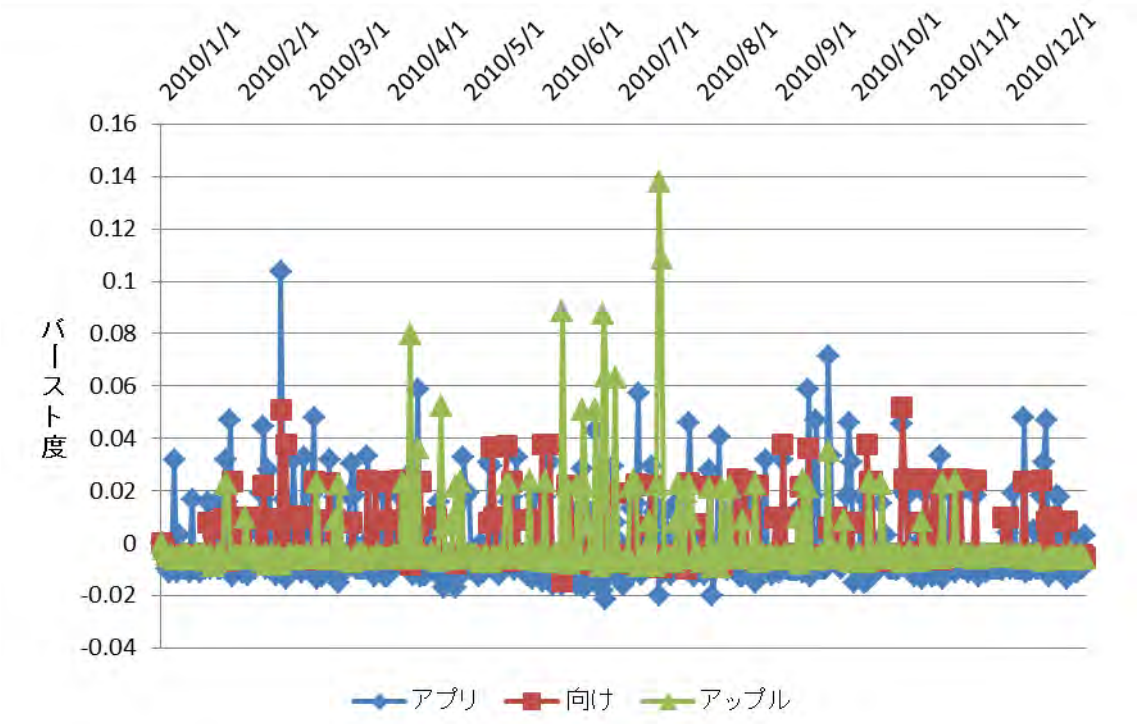

図 13. 2010 年「iPhone」と共起語の時系列解析結果

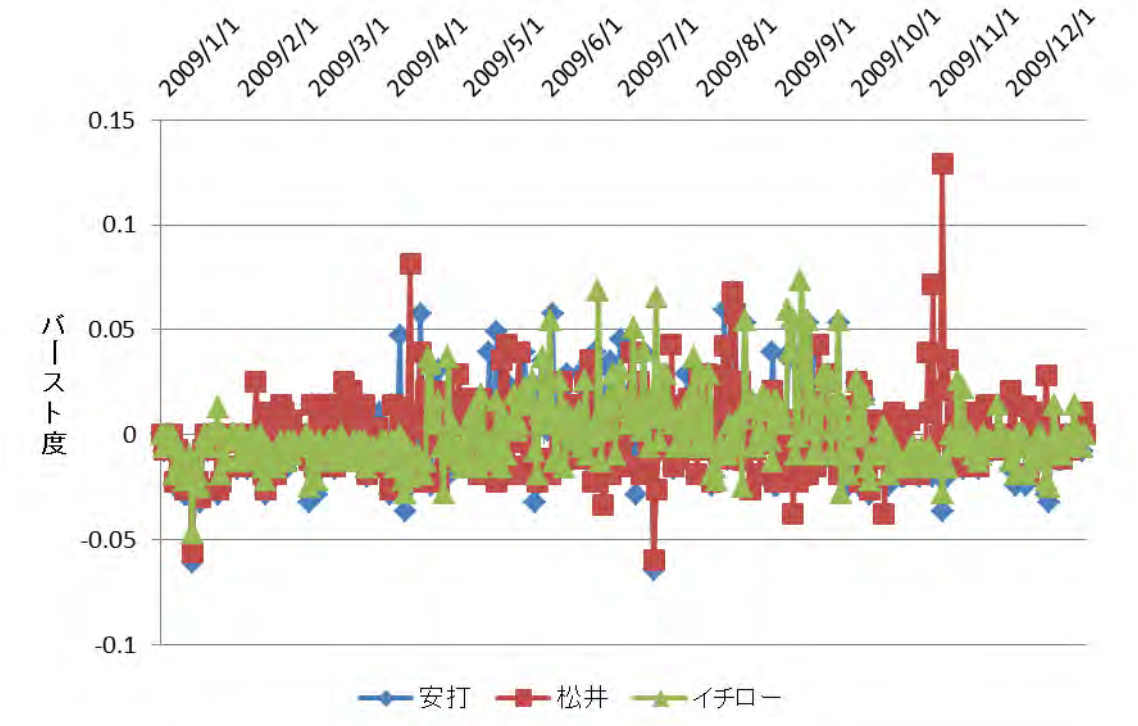

図 14. 2009 年「大リーグ」と共起語の時系列解析結果

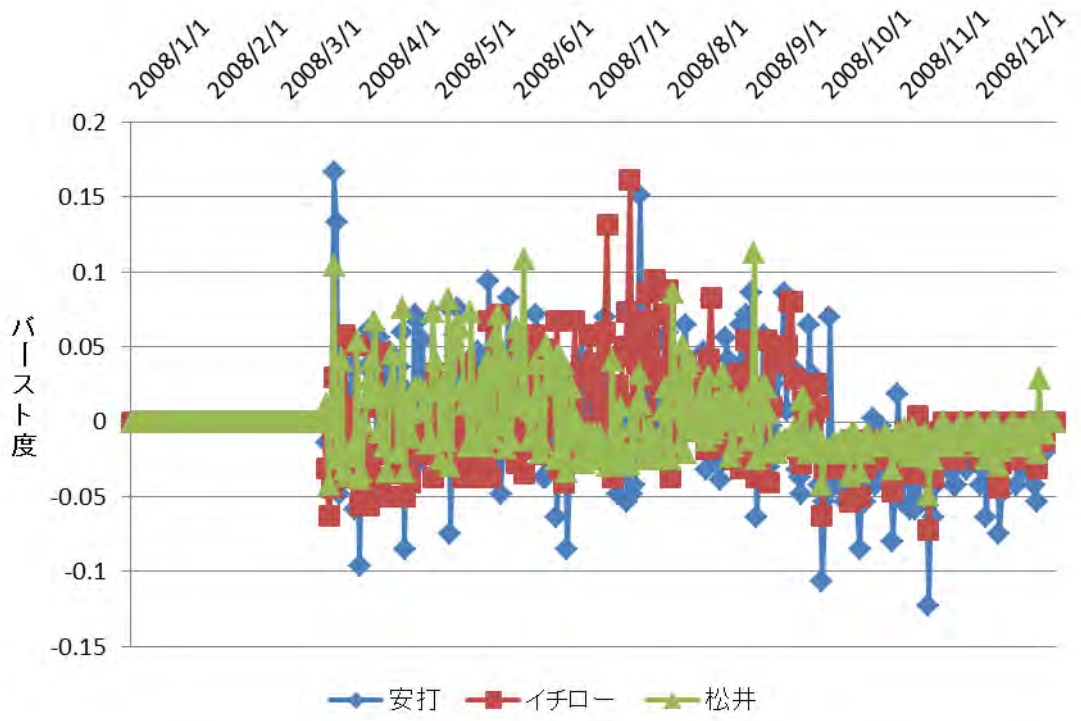

図 15. 2008 年「大リーグ」と共起語の時系列解析結果