ニュースアーカイブを用いた話題変化と原因語の発見

森井　洸明†　　アダム　ヤトフト††　　田中　克己††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †{k.morii,adam,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本稿では，近年その量を大幅に増加させつつあるデジタル文書アーカイブの中から特にニュースアーカイブを対象とし，ユーザの与えたクエリ語に関する話題情報を機械的に抽出する手法を提案する．本手法では，まず対象期間の各年度について，話題を形成する原因となった語をニュース記事の本文中でクエリ語と共起する語の中から発見する．さらに話題の変化の検出と可視化を行うために，トレンド分析の考え方を利用する．トレンド分析とはパラメータの時間的な変化からデータ傾向やその変化の原因となった事象を分析しようとする手法のことである．本手法では様々なパラメータについてその変化をグラフとして可視化することで，クエリ語にまつわる話題の性質を分析することを目指す．

キーワード　時系列データマイニング, ニュースアーカイブ, トレンド分析

1. はじめに

Web の発達によるデジタル文書の増加や，文字認識技術の発達による書籍や新聞などといった既存の紙媒体のデジタル化などを背景に，電子的な文書メディアのアーカイブの量が近年急激に増加しつつある．またそれらを容易に Web 上で利用することも可能になってきており，例えば Google の提供する Google Books(注1)では，数百万冊の書籍データに対し全文検索や内容の閲覧を行うことができる。しかし簡単にアクセスすることができたとしてもデジタル文書アーカイブに長年にわたって蓄積された文書の数は膨大であり，ユーザが求める情報を自身で探し出すのは困難である．本研究は，こういったデジタル文書アーカイブを機械的に処理し，アーカイブに含まれている有益な歴史的情報を抽出してユーザに直接提供するための手法を提案する．

文書アーカイブには様々なものがあるが，本研究はその中でも特にニュースアーカイブを対象としている．これは，書籍や雑誌，ブログなどのアーカイブでは個々の文書に記述されている内容のコンセプトが様々に異なる場合があるのに対し，ニュース記事は大抵の場合過去に起こった特定のイベントについて端的に書き記されたものであり，個々の文書が記述する内容にある程度の一貫性が期待できるためである．そして，本研究ではそういったニュース記事の性質を考慮して，ユーザが入力したクエリ語に対し，その語に関する「話題」を対象とした知識の抽出を考える．

ここで，話題とはニュース記事の記述の対象となった事象を指す．ニュースアーカイブには過去に社会で起こった様々な出来事を話題とする記事が数多く集積されている．これに関連する情報要求の例として,『どのような「災害」が過去に話題になったのか』や『「日本」に関して過去にどのような話題があったのか』といったものが考えられる．こうした要求に対する情報は Wikipedia など既存の Web サイトからもある程度得ることはできるが，Wikipedia などは特に重要な部分だけを後からまとめただけのものであり，ある話題がどれぐらいの間報道されていたか，その報道のされ方にどのような変化があったかといった情報は失われていることがほとんどである．それに対し，過去に起こったイベントについてその発生当時からの報道内容が時系列的にそのまま保存されているニュースアーカイブからは，現在の Web にはないより有益な知識が得られると考えられる．一方で，ニュースアーカイブに長い年月にわたって集積された大量の記事を人手で読み解くのは非常に手間がかかる．

そこで，本稿ではニュースアーカイブからこのような情報を抽出し，ユーザが記事を読むことなくクエリ語にまつわる過去の話題を一望できるようなシステムの提案を行う．本手法ではまず対象期間内の各年度について，クエリ語に関する話題を形成する原因となった語をニュース記事の本文から抽出する．さらに，得られた話題の原因語とトレンド分析の考え方を利用して，話題に関する様々な情報の検出と可視化を行う．トレンド分析とはパラメータの時間的な変化に注目し，データ傾向やその変化の原因となった事象を分析しようとする手法のことである．パラメータとして何を用いるかによって様々な応用がなされており，典型的には株価や為替レートをパラメータに用いて経済学の分野で利用されていることが多い．本手法では様々なパラメータについてその変化をグラフとして可視化することで，クエリ語にまつわる話題の性質を分析することを目指す．

本稿の構成は以下の通りである. まず第 2 章では関連研究について述べ, 第 3 章で話題形成の原因となった語の発見, 第 4 章で話題変化の検出・可視化手法についてそれぞれ述べる．第 5 章で実装したシステムを用いた実験及び考察を行い，第 6 章で本稿のまとめと今後の展望について述べる．

(注1)：http://books.google.com

2. 関連研究

2.1 文書アーカイブマイニングとトレンド分析

いくつかの文書アーカイブはそのデータがWeb上に公開されており，さらにサイト内で検索インタフェースを提供していることも多い．Corpus of Historical American English(COHA)[(注2)]には1810年から2009年までに出版された合計4億語以上の様々な書籍がアーカイブされており，サイト内の検索インタフェースでクエリ語を含む文書を検索したり，時間経過によるクエリ語の出現頻度の変化をグラフで見ることが可能である．これはクエリ語の出現頻度をパラメータとしたトレンド分析と考えることができる．また，Google books Ngram Viewer[(注3)]ではより大規模な文書アーカイブを使って同様のグラフを見ることができる．これらのインタフェースから得られるのは単なる語の出現頻度の変化のみであり，その変化の原因を知ることはできないが，こういった情報をもとに言語進化に関する様々な知見を得ようとする研究が行われている [1] [2] [3].

一方で，Loreyらの開発したBlack Swan [4] [(注4)]は，GDPや平均寿命などの様々な統計データの変化に対し，あらかじめ蓄積しておいた日付や場所，カテゴリなどのメタデータが付与されたイベント情報をルールマイニングにより結びつけることで説明付けを行っている．クエリ語が検索された回数の変化をグラフ化してユーザに提示するGoogle Trends[(注5)]では，グラフ中のピークのような特徴的な変化に対し関連するニュース記事を推薦することで，ユーザがその変化の原因をある程度調べることができるようになっている．本研究では，語の集合を用いてグラフに対する説明付けを行うことを考える．

2.2 話題に注目したニュースアーカイブ分析

文書アーカイブの中でもニュースアーカイブを対象とした研究は数多く行われているが，特にニュース記事が持つ話題に注目したものの中で代表的なものとしては，ニュース記事から話題 (topic) を自動的に検出しそれに基づいた分類などを行うことを目的としたTDT(The Topic Detection and Tracking) プロジェクト [5] が挙げられる．

また，ニュースアーカイブの時系列的側面を利用した研究にも様々なものがある．Ohshimaら [6] は入力語と特定のパターンのもとで頻繁に共起する語が時間経過とともにどのように変化するかを調べることで，入力語を取り巻く同位語関係の変遷を知ろうとする研究を行っている．張ら [7] は内閣支持率などの時系列データに対し，それに関連した話題を持つ記事の出現頻度の変化を用いて入力データの変動に影響を与えた話題の発見を行う手法を提案している．

本研究では [6] [7] と同じくニュースアーカイブを時系列データとして利用し，入力されたクエリ語に関する話題が時間変化とともにどのように変化したかを分析する．なお, [5] や [7] ではニュース記事そのものが持つ話題に注目しているが，本研究ではそれは考慮せず，単純にクエリ語を本文中に含む記事の中からクエリ語に関する話題を抽出することを考える．

(注2)：http://corpus.byu.edu/coha
(注3)：http://books.google.com/ngrams/
(注4)：http://blackswanevents.org
(注5)：http://google.com/trends

3. 話題の原因語の発見

本章では，ユーザが入力したクエリ語に対し，対象期間中の各年度に書かれた記事の集合から，その年度におけるクエリ語に関する話題の原因語を抽出する手法について述べる．ここで原因語とは，ある年度においてクエリ語に関する話題を形成，あるいは説明・象徴するような語を指す．例えば1995年において，クエリ語 “Japan” に対して “Kobe”, “earthquake” といった語は，1995年に日本に関して起こった代表的な出来事である阪神大震災を象徴する，話題の原因語と考えることができる．

3.1 共起語集合の生成

まず仮定として，ある年度における話題の原因語は，その年度に書かれた記事の本文中でクエリ語と頻繁に共起した語の中に含まれていると考える．ここで,「共起する」とはクエリ語の前後 k 語以内に出現することとみなす．クエリ語 q に対し期間 $(start, start+1, \ldots, end)$ 中のある年度 i における共起語集合 $C(q,i)$ を，以下のように表す．

$$C(q,i) = \{(t_1, f(t_1,i)), (t_2, f(t_2,i)), \ldots\} \quad (1)$$

ただし，

t_k：共起語 (各共起語は複数の共起語集合に含まれうる)
$f(t_k,i)$：語 t_k の年度 i における共起回数

である．各年度について記事本文中におけるすべてのクエリ語の出現に対し，その共起語をカウントしこのような共起語集合を生成する．ただし，共起語を収集する際には，各語に対してステミングを行い，また冠詞や前置詞など一部の単語をストップワードに設定している．

ここで，各共起語の共起回数はその年度の記事の総数やクエリ語の総出現回数の影響を受けるので，$C(q,i)$ に含まれる各共起語に対しその共起回数を年度 i におけるクエリ語 q の総出現回数 f_{qi} で割った，以下のような正規化済み共起語集合 $C'(q,i)$ をつくる．

$$C'(q,i) = \{(t_1, f'(t_1,i)), (t_2, f'(t_2,i)), \ldots\} \quad (2)$$

ただし，$f'(t_k,i) = f(t_k,i)/f_{qi}$

3.2 共起語へのスコア付け

3.1節で得られた共起語集合 $C'(q,i)$ に含まれる語の中でも特に共起回数が上位にあるものは，その年度においてクエリ語との関連が非常に強い語であると考えることができる．しかし，これらの集合には “people”, “Mr”, “year” のような一般語，クエリ語 “disaster” に対して “damage”, “victim” のように年度にかかわらず常によく共起するような語を多く含んでいる．

そこで，より厳密な仮定として，共起語集合 $C'(q,i)$ に含まれる語の中で，特にその年度に特徴的な語，つまり他の年度にはあまり出現せずその年度にだけ共起回数が多い語が，その年度における話題の原因語とみなせるのではないかと考える．したがって，本手法では共起語集合 $C'(q,i)$ に含まれる各語に，どれだけその年度に特徴的であるかを表すスコア付けを行い，

高いスコアのついた語を原因語として検出する．

共起語集合 $C'(q,i)$ から，スコア付けされた共起語集合

$$S(q,i) = \{(t_1, s(t_1,i)), (t_2, s(t_2,i)), \ldots\} \quad (3)$$

ただし，

$s(t_k,i)$：語 t_k の年度 i におけるスコア

を生成することを考える．スコアの計算方法の基本的なアイデアとしては，各年度の共起語集合を仮想的な文書と考え，多くの文書に出現する語の重要度を下げる tf-idf の考え方を利用する．最も典型的な tf-idf ではある語の tf-idf 値を計算する際の逆文書頻度としてその語を含む文書の数，つまり本手法の場合その語を共起語集合に含む年度数を用いるが，共起語集合には多くの語が共通に含まれているため，単純な年度数ではなく各年度における語の共起頻度 $f'(t_k,i)$ を用いる．よって，語 t_k の年度 i における単純なスコアの計算式を以下のように定義する．

$$s_{simple}(t_k,i) = f'(t_k,i) \cdot \frac{1}{\sum_{j=start}^{end} tf(t_k,j) + 1} \quad (4)$$

ただし，

$$tf(tk,j) = \begin{cases} f'(t_k,j) & \text{if } t_k \in C'(q,j) \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

さらに本手法では，ニュースアーカイブが時系列データであることを考慮し，式 (4) の総和内の各項に年度 i と j の差を用いた以下のような重み付けを行っている．

$$s(t_k,i) = f'(t) \cdot \frac{1}{\sum_{j=start}^{end} \omega_{ij} \cdot tf(t_k,j) + 1} \quad (5)$$

ただし，

$\omega_{ij} = \mathrm{e}^{\lambda|i-j|}$ (λは重みの強さを決定づける値)

である．年度 i と年度 j の差が大きくなるにつれて重み ω_{ij} も大きくなり，年度 j の tf 値は増加し語 t_k のスコアは減少することになる．つまりこの重みは離れた年度の共起語集合に共通に現れる語のスコアを下げる方向に働くが，これは特定のイベントの発生に関係しない共起語は様々な年度の共起語集合の中に無差別に現れるのに対し，特定のイベントの発生によって共起頻度が上がった語はイベントが発生した年度の近くに固まって現れると考えられるので，より原因語である可能性の高い後者のスコアを増加させるというヒューリスティクスに基づいている．また，式 (4) は式 (5) において $\lambda = 0$ とおいたものと言える．

このようにして，クエリ語 q に対し，対象期間中の各年度においてスコア付けされた共起語集合の集合

$$Scores(q) = \{S(q,start), S(q,start+1), \ldots, S(q,end)\} \quad (6)$$

が得られ，この中の各集合 $S(q,i)$ において特に高いスコアを持つ共起語が，その年度におけるクエリ語に関する話題の原因語であるとみなすことができる．

4. トレンド分析による話題変化の検出と可視化

第 3 章で述べた手法により各年度における話題の原因語が得られたことで，ユーザは任意の年度についてその年度におけるクエリ語に関する話題を形成する語を知ることができる．しかし，クエリ語に関する知識を持たないユーザには期間中のどの年度が，なにか大きな話題のあった注目すべき年度であるのかを判断することはできない．また，時系列を持ったデータソースであるニュースアーカイブの性質を考慮すると，注目した年度における主要な話題が，注目した 1 年だけではなく他の年度においてどれだけの話題性を持っていたかといった情報についても知ることができるとより有益である．

そこで本章では，大きな話題の盛り上がりがあった注目すべき年度や，注目した年度で起こった話題がどれほど継続したか，またどれほどその予兆があったかといった情報を，トレンド分析の考え方を用いて検出・可視化する手法について述べる．トレンド分析とはパラメータの時間的な変化からデータ傾向やその変化の原因を分析しようとする手法のことである．以降の節で利用したパラメータとそれにより得られる情報について述べる．

4.1 クエリ語を本文中に含む記事の数を用いた話題性の可視化

トレンド分析により話題の変化を検出するための最も直感的かつ主要なパラメータとして，対象期間中の各年度においてクエリ語が本文中に出現した記事の数を利用することが考えられる．ここで，パラメータとして記事数を用いる理由として,「記事数の変化は，クエリ語が持つ話題性の変化に対応する」という仮定を与える．この仮定によれば，例えばある年度にクエリ語に関する記事の数が急激に増加したということはその語が話題になったということを表し，このことからその語に関して何か大きなイベントがその年度に発生したのではないかという推測を行うことができる．ただし，記事数はその年度に書かれた記事の総数の影響を受けることが考えられるので，各年度ごとにその年度に書かれた記事の総数に対する割合を実際のパラメータとして用いる．

このようにクエリ語を含む記事の数をその語が持つ話題性に対応する指標とみなし，その変化をグラフ化することで，クエリ語に関して知識を持たないユーザに対しても注目すべき年度を指し示すことができると考える．そして注目すべき年度が分かれば，第 3 章に述べた手法で得たその年度における原因語を提示することで，ユーザはクエリ語に関して過去にどのような大きな話題が起こったのかを知ることができる．もしこのシステムをアプリケーション化するとすれば，ユーザが入力したクエリ語に対しまずこのグラフを見せ，その後ユーザが興味を持った年度を例えばクリックすることで，その年度における話題の原因語や，次節以降で述べるような話題の追跡の結果を表示するというようなモデルが考えられる．

4.2 話題の追跡

本章のはじめに述べたように，時系列データであるニュースアーカイブからはある話題についてそれが起こった年度だけで

はなく，その前後の年度における扱われ方に関する情報も得ることができると考えられる．そこで本節では以下に述べる 2 つの指標を用いて，注目した年度における話題に関してその話題が注目年度以降どれぐらいの間報道されていたか，あるいは注目年度以前にその話題が発生する予兆のようなものなかったかといった情報を可視化する手法について述べる．

4.2.1 原因語のスコア平均

3.2 節では，ある注目年度についてその年度のスコア付けされた共起語集合に含まれる語の中で特に高いスコアを持つ語は，その年度における話題の原因語とみなすことができるという議論を行った．そこで本手法では注目年度における共起語集合の中で特に高いスコアを持つ語上位 10 語をその年度における話題を形成する語とみなし，その各語が持つスコアの平均値を話題性の指標として用いることを考える．するとクエリ語 q に関して注目年度 i における話題を形成する語の集合 $\{t_1, \ldots, t_n\}$ に対し，任意の年度 j について話題性スコア $s_{avg}(q,i,j)$ が以下のように計算できる．

$$s_{avg}(q,i,j) = \frac{s_1 + s_2 + \ldots + s_n}{n} \tag{7}$$

ただし，

$$s_k = \begin{cases} s(t_k, j) & \text{if } t_k \in S(q,j) \\ 0 & \text{otherwise} \end{cases}$$

である．1981 年から 2010 年度までの全ての年度についてこの値を計算しその変化をグラフにすると，そのグラフは年度 i に最大値をとり i から遠ざかるにしたがって減少していくような形をとると考えられる．一般的に話題は時々刻々と変化していくのでこの値は急激に減少していくことが予想されるが，例えばもしある話題が発生した以後の年度におけるこの値の減少が緩やかであれば，その話題は発生した以後も継続的に報道され続けている話題なのではないかといった発見が得られるかもしれない．逆に話題の発生以前にもこの値が高くなっていることがあれば，その話題は発生する前に何らかの予兆のようなものがあったのではないかと推測することもできるだろう．このような話題性の追跡によって，注目した年度の話題に関する時系列的な情報の発見を目指す．

4.2.2 JS divergence を用いた共起語集合間の類似度

原因語のスコア平均を用いた話題性スコアの他に，話題の追跡を行うためのもう 1 つの指標として，共起語集合間の類似度を用いることを考える．前節で述べた指標は注目年度の話題がどれだけ他の年度に現れているかを求めるものであったが，本節では注目年度と他の年度の共起語集合同士を直接比較することで，注目年度の話題がどれだけ他の年度の話題と類似しているかを求める指標を提案する．

本手法では，集合の類似度の指標として，JS(Jensen-Shanon) divergence を利用する．JS divergence とは，2 つの分布の相違度を測る非対称な尺度である KL(Kullback-Leibler) divergence を対称化したものである．分布 P と分布 Q の間の JS divergence $D_{JS}(P||Q)$ は以下のように定義される．

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||R) + \frac{1}{2}D_{KL}(Q||R) \tag{8}$$

ただし，

$$D_{KL}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}, \quad R = \frac{1}{2}P + \frac{1}{2}Q$$

である．JS divergence は 0 から 1 までの値をとり，その値が大きいほど 2 つの分布はより異なっていると考えられる．よって，$1 - D_{JS}(P||Q)$ を分布 P と分布 Q の類似度とみなすことができる．

2 つのスコア付けされた共起語集合 $S(q,i)$ に JS divergence を適用することを考える．共起語集合が話題を形成すると考えると，2 つの共起語集合の類似度を求めることで 2 つの年度の話題がどれだけ似ているかを推測することができる．しかし，集合に含まれる共起語のうちスコアの低い語は話題の形成にほとんど寄与していないと考えられるため，本手法では共起語集合 $S(q,i)$ のうちスコア上位 m 件を取り出したものに対して JS divergence を適用することとした．また，JS divergence は分布の相違度を測る尺度であるが，共起語集合 $S(q,i)$ は分布ではないため，JS divergence を適用するに当たって各共起語のスコアを集合に含まれる全共起語のスコアの和で割ることで仮想的な分布として扱っている．以上より，クエリ語 q に関して注目年度 i における話題を形成する共起語集合 $S(q,i)$ に対し，任意の年度 j について年度 i との話題の類似度 $sim(q,i,j)$ が以下のように計算できる．

$$sim(q,i,j) = 1 - D_{JS}(S'(q,i)||S'(q,j)) \tag{9}$$

ただし，

$$S'(q,i) = \{(t_k, \frac{s(t_k,i)}{\sum_{t_l \in S} s(t_l,i)}) | (t_k, s(t_k,i)) \in S)\}$$

前節と同様に 1981 年から 2010 年までの全ての年度についてこの値を計算しその変化をグラフにすることで，注目年度と各年度の話題の類似度を可視化し，話題の類似関係の発見を行う．

5. 実　　験

5.1 使用したデータセット

Web 上にはニュースメディアが自社のサイトで過去の記事を公開しているものや，Google News(注6)のように様々な報道機関から集約したものなど多くのニュースアーカイブが公開されているが，本稿では API が整備されておりプログラムからの利用が容易な New York Times Article Archive (注7) (注8)を利用している．New York Times Article Archive には 1851 年から現在までの 1300 万件以上の記事が保存されているが，API で利用可能なのは 1981 年以降の記事のみなので，本稿では対象期間を 1981 年から 2010 年までの 30 年間とし，その期間に含まれる約 300 万件のニュース記事をデータセットとして以降の議論を行う．1981 年から 2010 年までの各年度における記事の総数は表 1 に挙げる通りである．4.1 節で述べたクエリ語を含む記事数の変化を調べる際には，これらの値を用いて記事数の正規化を行う．

(注6)：http://news.google.com
(注7)：http://www.nytimes.com/ref/membercenter/nytarchive.html
(注8)：http://developer.nytimes.com/

表 1　各年度ごとの記事の総数

年度	記事数	年度	記事数	年度	記事数
1981	95,090	1991	85,197	2001	93,552
1982	96,064	1992	89,798	2002	96,171
1983	101,344	1993	85,582	2003	94,482
1984	107,111	1994	82,582	2004	90,207
1985	103,844	1995	86,067	2005	88,527
1986	107,000	1996	80,972	2006	98,700
1987	104,186	1997	83,059	2007	104,054
1988	102,750	1998	87,668	2008	110,997
1989	101,363	1999	89,196	2009	107,063
1990	96,220	2000	91,816	2010	99,719
				計	2,860,381

表 2　実験により得られた共起語集合の例

query：Japan, year：1995		
$C'(q,i)$	$S(q,i)$ $\lambda = 0.0$	$S(q,i)$ $\lambda = 0.3$
japan	trade	**earthquak**
japanes	japanese	**kobe**
american	**earthquak**	auto
unit	car	troop
state	american	**deficit**
year	japan	expens
trade	peopl	dealer
market	dollar	presenc
mr	troop	side
bank	**deficit**	magazin

当然ながら文書アーカイブから得られる知識はそのアーカイブのドメインに大きく依存するので，本実験で提案手法により得られる情報は必然的にアメリカ合衆国でニュースになった話題に関するものになる．しかしながら，本手法に特定のデータソースに依存する箇所は存在しないので，別のニュースアーカイブをソースとして用いることで異なるドメインの知識も得られるであろうということをここで補足しておく．

5.2　原因語発見手法の評価

本節では第 3 章で述べたスコア付けによる話題の原因語発見手法の実行例を示し，その評価を行う．

5.2.1　実 験 概 要

“disaster”，“war”，“infection”，“Japan”，“Michael Jackson” の 5 件のクエリについて，各クエリごとに 3 つの注目年度を設定し，まずそれぞれの年度における共起語集合 $C(q,i)$ を生成する．本実験では，共起語はクエリ語の前後 20 語以内に出現する語とみなす．さらに各共起語集合の要素数は最大 200 件とし，共起回数が上位 200 件以内に含まれる共起語のみを集合に含める．また，処理にかかる時間を考慮して，共起語の抽出を行う記事数は 1 年につき最大 500 件とし，それを超える記事数が存在する場合には記事の中からランダムに 500 件を選択し抽出を行っている．共起語を抽出する際のストップワードのリストには，京都大学で開発されたプログラミングライブラリ SlothLib(注9)で使用されているものを用いている．

そして各年度について，$C'(q,i)$ における単純な共起頻度，$S(q,i)$ において式 (5) の λ の値を 0 とした重み付けを行わないスコア，以降の実験で実際に用いた値である 0.3 とした重み付けを行ったスコアがそれぞれ上位 5 件,10 件以内の共起語の中に，その年度の話題の原因語と考えられる語がどれだけ含まれているかを調べた．共起語が話題の原因語であるかどうかの判定に関しては，その年度におけるクエリ語に関する話題を容易に連想できるような語を選び，話題との関連が不明瞭なものや年度に関係なくクエリ語と共起していると考えられる語は選ばないこととした．なお，注目年度の設定に際しては，原因語の正解を分かりやすくするためにクエリ語に関して大きな話題が起こった年度を中心に選択している．

(注9)：http://www.dl.kuis.kyoto-u.ac.jp/slothlib

表 3　精度評価の結果

query：disaster year:1986, 2001, 2005		$C'(q,i)$	$S(q,i)$ $\lambda = 0.0$	$S(q,i)$ $\lambda = 0.3$
原因語の割合	上位 5 件	60%	86%	80%
	上位 10 件	36%	66%	60%
query：war year:1991, 1999, 2003		$C'(q,i)$	$S(q,i)$ $\lambda = 0.0$	$S(q,i)$ $\lambda = 0.3$
原因語の割合	上位 5 件	13%	66%	66%
	上位 10 件	13%	50%	56%
query：infection year:2001, 2003, 2005		$C'(q,i)$	$S(q,i)$ $\lambda = 0.0$	$S(q,i)$ $\lambda = 0.3$
原因語の割合	上位 5 件	0%	26%	40%
	上位 10 件	6%	13%	20%
query：Japan year:1990, 1995, 1998		$C'(q,i)$	$S(q,i)$ $\lambda = 0.0$	$S(q,i)$ $\lambda = 0.3$
原因語の割合	上位 5 件	0%	6%	20%
	上位 10 件	0%	3%	13%
query：Michel Jackson year:2003, 2005, 2009		$C'(q,i)$	$S(q,i)$ $\lambda = 0.0$	$S(q,i)$ $\lambda = 0.3$
原因語の割合	上位 5 件	26%	33%	46%
	上位 10 件	26%	36%	36%
全クエリ平均		$C'(q,i)$	$S(q,i)$ $\lambda = 0.0$	$S(q,i)$ $\lambda = 0.3$
原因語の割合	上位 5 件	20%	44%	51%
	上位 10 件	17%	34%	37%

5.2.2　実 験 結 果

まず，原因語判定に用いた共起語集合の例を表 2 に示す．この表にはクエリ語”Japan” の 1995 年における共起語集合の中から，共起頻度，式 (5) の λ を 0 としたときのスコア，λ を 0.3 としたときのスコアがそれぞれ上位 10 件以内であった語を記載し，原因語と判断された語を太字で表記している．この場合，共起頻度，重み付けを行わないスコア，重み付けを行ったスコアを用いたときの上位 5 件における原因語が占める割合はそれぞれ 0%, 20%, 60%，上位 10 件に原因語が占める割合はそれぞれ 0%, 20%, 30%と求められる．これらの値を各クエリに対し 3 つの年度についてそれぞれ求めその平均をとった結果と，

表 4　$Scores(disaster)$

1986	2001	2003	2005	2010
chernobyl	trade	columbia	tsunami	gulf
soviet	site	nasa	orlean	bp
rocket	manhattan	space	hurrican	oil
panel	center	shuttl	katrina	haiti
jan	sept	panic	natur	deepwat

表 5　$Scores(war)$

1990	1991	1998	1999	2002	2003
saudi	kuwait	movi	kosovo	hors	saddam
kuwait	saudi	bank	honor	terror	hussein
arabia	iraq’	ryan	nato	terrorist	iraq
loss	target	swiss	milosev	emblem	intellig
reflect	worker	kosovo	releas	queen	bush

全クエリの結果を平均した値を表 3 に示す.

5.2.3　考　　察

表 2 において，共起頻度が高い順番に並べただけの $C'(q,i)$ の上位に位置する語は “japan”，“american”，“year” といった一般的な語がほとんどであるが，重みのないスコア付けを行った時点でそういった他の年度にもよく出現する語に比べ，“earthquak” や “trade” などのクエリ語により関連の強い語が高い順位に位置するようになっていることが分かる．さらに重みのついたスコア付けを行うと，共起頻度が 1995 年の前後で高い語により高いスコアがつくようになり，“earthquak”，“kobe”，“deficit” などの 1995 年の日本に関する話題を象徴するような語が上位を占めるようになる (1995 年は阪神大震災の他に，史上空前の円高が起こった年でもある)．表 3 からも，ほとんどのクエリと全クエリ平均の両方において提案手法の重み付きスコアを用いたときが最も多くの原因語が得られていることが分かる．

また，クエリごとの結果に目を移すと，関連の強い話題として 2001 年にアメリカ同時多発テロ事件，2005 年にハリケーン・カトリーナの起こった “disaster”，1991 年に湾岸戦争，2003 年にイラク戦争の起こった “war” のように非常に大きな話題を持つクエリ語に関しては，高い割合で話題の原因語が得られている．一方で，“Japan” のように New York Times で大きく取り上げられるような話題があまり起こらないクエリ語については得られる共起語の割合は低くなっている．

5.3　話題変化の可視化実験

本節では第 4 章で述べた話題変化の可視化手法を実装し，いくつかのクエリ語に対し実行することで，提案手法の有効性を検証する．

5.3.1　実 験 概 要

クエリ語は “disaster”，“war”，“Japan” の 3 つを用い，それぞれのクエリ語についてまず第 3 章で述べたスコア付けされた共起語集合の集合 $Scores(q)$ を生成する．次に，第 4 章で述べた話題性の可視化と話題の追跡を行うため，各クエリ語を本文中に含む記事の数と，クエリ語ごとにいくつか設定した注目年度について 4.2 節で述べた 2 つの式 (7), (9) の値 $s_{avg}(q,i,j), sim(q,i,j)$ をそれぞれ求め，それらの変化をグラフ化した．式 (7) の計算の際には共起語集合の上位 10 件，式 (9) の計算の際には共起語集合の上位 50 件をそれぞれ使用している．

5.3.2　実 験 結 果

各クエリ語に対して得られた 1981 年度から 2010 年までの各年度における共起語集合 $S(q,i)$ の一部を，表 4 から表 6 までに示す．各年度について上位 5 件までを記載しているが，実際には各共起語集合は 200 件の語を含む．“disaster”，“war”，“Japan” の各クエリ語に対する共起語集合がそれぞれ表 4，表 5，表 6 に対応する．

表 6　$Scores(Japan)$

1987	1990	1995	1998	2003	2009
tariff	poll	earthquak	hong	iraq	korea
reagan	soviet	kobe	fell	north	obama
impos	latest	auto	neutrino	samurai	japan
treasuri	kaifu	troop	kong	koizumi	stimulu
retali	pressur	deficit	recess	send	south
1993	1994	1995	1996	1997	1998

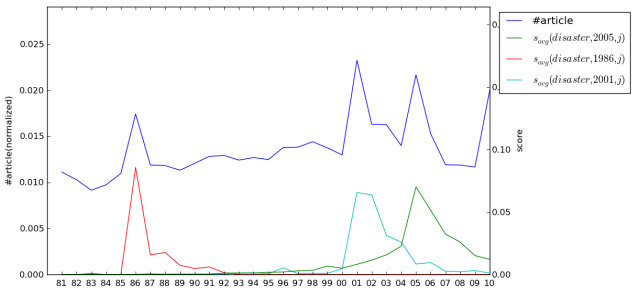

図 1　記事数，$s_{avg}(disaster,1986,j)$, $s_{avg}(disaster,2001,j)$, $s_{avg}(disaster,2005,j)$

そして各クエリ語について，クエリ語を本文中に含む記事数と注目年度に対する式 (7)，(9) の値 $s_{avg}(q,i,j), sim(q,i,j)$ の変化を表したグラフを，図 1 から図 6 までに示す．本実験では，“disaster” について 1986 年，2001 年，2005 年，“war” について 1991 年，1999 年，2003 年，“Japan” について 1990 年，1995 年，1998 年をそれぞれ注目年度に設定した．なお，記事数と式 (7) の値の変化は 1 つのグラフにまとめており，“disaster”，“war”，“Japan” の各クエリ語に対しそれぞれ図 1，図 3，図 5 が対応する．式 (9) の値の変化を表したグラフについては，“disaster”，“war”，“Japan” に対しそれぞれ図 2，図 4，図 6 が対応する．

5.3.3　考　　察

まずクエリ語 “disaster” について，図 1 のクエリ語を含む記事数の変化を表すグラフを見ると，1986 年, 2001 年,2005 年,2010 年に記事数の急激な増加が見られる．そこで例えば 1986 年における共起語集合の中で高いスコアのついた語を表 4 で調べると，“chernobyl”，“soviet” といった語が上位に位置していることがわかる．このことから 1986 年において “disaster” に関する大きな話題の盛り上がりが起こり，その原因はチェルノ

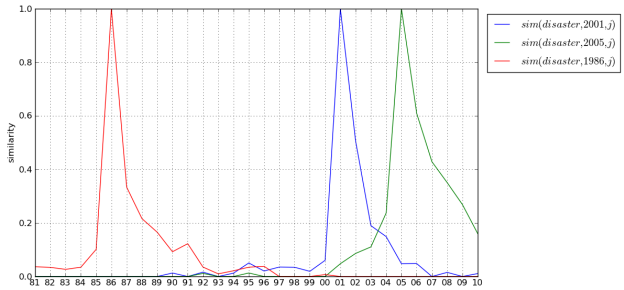

図 2　$sim(disaster, 1986, j), sim(disaster, 2001, j),$
$sim(disaster, 2005, j)$

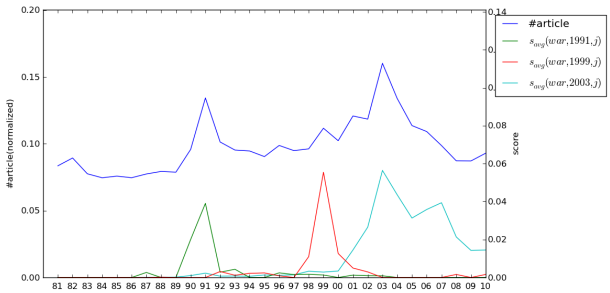

図 3　記事数, $s_{avg}(war, 1991, j),$
$s_{avg}(war, 1999, j), s_{avg}(war, 2003, j)$

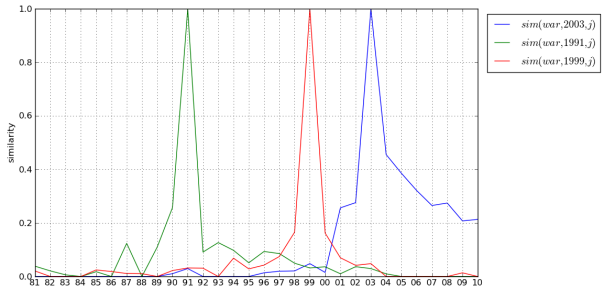

図 4　$sim(war, 1991, j), sim(war, 1999, j), sim(war, 2003, j)$

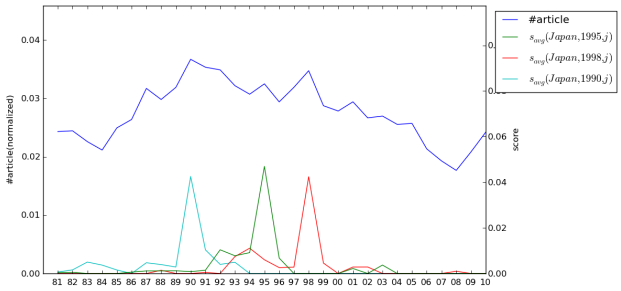

図 5　記事数, $s_{avg}(Japan, 1990, j),$
$s_{avg}(Japan, 1995, j), s_{avg}(Japan, 1998, j)$

ブイリ原子力発電所事故であると推測することができる．他の年度についても表 4 で高いスコアのついた語を調べると，2001 年に関して “trade”, “manhattan” などからアメリカ同時多発テロ事件，2005 年に関して “tsunami”, “hurrican”, “katrina”

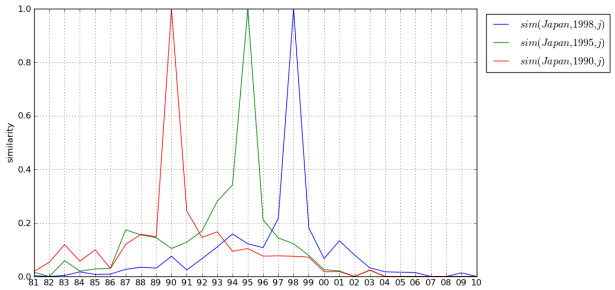

図 6　$sim(Japan, 1990, j), sim(Japan, 1995, j),$
$sim(Japan, 1998, j)$

からスマトラ沖地震とハリケーン・カトリーナ，2010 年に関して “gulf”, “bp”, “haiti” から英 BP 社のメキシコ湾原油流出事故とハイチ地震がそれぞれの話題の盛り上がりの原因であったと考えられる．また，記事数に大きな変化が見られないその他の年度に関しても，例えば 2003 年について “columbia”, “space”, “shuttle” といった共起語からコロンビア号空中分解事故というようにその年度に起こった話題の推測を行うことができる．このように，各年度における話題の原因語から，クエリ語に関してその年度に起こった出来事などを推測することが可能である．

次に，原因語のスコアと共起語集合の類似度を用いた話題の追跡に注目する．図 1 において 1986 年における話題語のスコア平均 $s_{avg}(disaster, 1986, j)$ が注目年度の翌年には急激に減少しているのに比べ，$s_{avg}(disaster, 2001, j)$ と $s_{avg}(disaster, 2005, j)$ は比較的緩やかに減少している．これはアメリカ国外の出来事であるチェルノブイリ原発事故に比べて，アメリカ国内で起こった非常に大きな出来事である同時多発テロや類似のイベントが何度も起こるハリケーンや地震のような出来事に関する話題の方が長い期間にわたって報道されているということを表しているのではないかと考えられる．また，注目年度以前の値を見ると，$s_{avg}(disaster, 1986, j)$ と $s_{avg}(disaster, 2001, j)$ の値は非常に小さくなっているのに対し，$s_{avg}(disaster, 2005, j)$ の値は注目年度以前の数年間においても比較的高い値を示している．これは，チェルノブイリ原発事故や同時多発テロが前触れなく起こる一回限りの出来事であるのに対し，ハリケーンや地震は注目年度以前にも類似の出来事が存在しているということに起因するのではないかと考えられる．図 2 の類似度のグラフに関しても同様の傾向が見られる．

このことに関連して，クエリ語 “war” の実験結果に目を移すと，図 3，図 4 のいずれの注目年度についても，$s_{avg}(disaster, 2005, j)$ と同様に注目年度以前の数年における各指標が比較的高い値をとっている．また，表 5 からも，注目年度とその前年度の共起語集合の間にいくつか共通の語がある．上位の共起語からもわかるように 1991 年,1999 年,2003 年のそれぞれの年度における主要な話題は湾岸戦争の開始，コソボ紛争，イラク戦争であると考えられるが，これらの話題に関してはいずれも注目年度以前からその予兆となるような出来事や戦争の継続の様子に関する報道が行われていたことが推測さ

れる．したがって，本稿で提案した話題の追跡を利用することで，前触れなく起こったのか予兆があったのか，一回限りの出来事なのか類似の出来事が存在するのかといった話題の性質を判別することができると考えられる．

一方で，クエリ語 “Japan” に関しては，図 5 の記事数のグラフには “disaster” や “war” のような目立ったピークが見られず，表 6 の共起語集合についても日本の総理大臣の名前と共に外国の国名や政治家の名前が現れていることが多く，日本自体に関してではなく他のものとの関係の中での報道が主であることが伺える．また，注目年度の 1 つである 1995 年は阪神大震災が起こった年であるが，阪神大震災が突発的であまり他に例のない出来事であるにもかかわらず図 6 中の $sim(Japan, 1995, j)$ の値は注目年度の前後においても高い値をとり，また他の注目年度についても同様の傾向が見られる．これは，アメリカの報道機関である New York Times では日本自体の話題に関する報道があまり行われていないために共起語がうまく 1 つの話題を形成せず，年度ごとの差が現れにくいためではないかと考えられる．ニュースアーカイブはあらゆる話題を網羅しているわけではないため，クエリ語のドメインによってはこのような問題が起こる可能性がある．

6. まとめと今後の展望

本稿では入力されたクエリ語とニュース記事の本文中で共起する語を利用してニュースアーカイブからクエリ語に関する話題を検出し，また検出された話題に対して様々な指標を用いてその性質を分析する手法を提案した．そして，いくつかのクエリ語について提案手法を適用し，実際にクエリ語に関する話題についての様々な情報を得る過程を示した．本手法は文脈を考慮しない単純な共起関係による単語ベースのアプローチにとどまっているため，得られる情報の精度をはじめとした様々な課題が残されている．

まず本手法の大きな問題点として挙げられるのが，高いスコアのついた共起語はその年度に起こった話題をうまく象徴することもあるが，その話題をはっきりと知るためにはユーザは得られた原因語と実際に起こった出来事を自分で結び付けなければならず，またそれが可能かどうかはユーザの持つ予備知識に大きく依存している点である．例えばクエリ語 “disaster” に対する 2001 年の話題の原因語として “trade”, “center” を提示されればほとんどのユーザにはそれが同時多発テロのことであると分かるが，1996 年の話題の原因語として “airline”, “flight” が提示されると，ユーザは 1996 年に何か航空事故が起こったのではないかと推測することはできても，具体的にどのような事故があったのかまでは知ることができない．このような場合にも原因語を手がかりに “1996 airline accident” のようなクエリで検索を行うことで実際の出来事を知ることができるかもしれないが，ユーザにそのような負担をかけるのは好ましくない．そこで，ユーザにただ原因語を提示するのではなく，原因語をクエリ拡張に用いて求める話題に特に関連の強いニュース記事をシステムがニュースアーカイブから検索し，記事の推薦やスニペットを用いた話題の説明を行うという方法が考えられる．

さらに，例えばハリケーン・カトリーナの発生した 2005 年の話題を追跡したとき，式 (7) や (9) が翌年以降も高い値をとっていても，単語レベルの比較ではそれがカトリーナの話題が継続していることによるものなのか，それとも別のハリケーンが発生したことによるものなのかを判別することができない．この問題を解決するためには，単語が使われている文脈を考慮したアプローチが必要になると考えられる．

また，本手法では共起語集合の中で高いスコアのついた語をその年度における話題の原因語としてひとまとまりに扱っていたが，クエリ語 “disaster” の 2005 年における原因語集合がカトリーナに関連する “hurricane” とスマトラ沖地震に関連する “tsunami” を含んでいたように，本来原因語は話題ごとに分けて扱われるべきである．これを実現するためには，Jaccard 係数などの共起尺度を用いて各原因語をクラスタリングするなどの方法が考えられる．

以上に述べたように本研究にはまだ不完全な部分が数多くあるが，手法をより洗練し高精度な話題の提示やその性質の分析が可能になれば，本研究はユーザが興味のある話題に関する様々な情報を一望できる有益なものになるであろうと考える．

謝辞 本研究の一部は，グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」（拠点リーダ：田中克己）によるものです．ここに記して謝意を表します．

文　献

[1] Jean-Baptiste, M., Kui, S. Y., Presser, A. A., Adrian, V., K., G. M., The Google Books Team, P., P. J., Dale, H., Dan, C., Peter, N., Jon, O., Steven, P., A., N. M. and Lieberman, A. E.: Quantitative Analysis of Culture Using Millions of Digitized Books, Science, Vol. 331, No. 6014, pp. 176-182 (2011).

[2] Nina, T., Kai, N., Thomas, T. and Thomas, R.: Using word sense discrimination on historic document collections, Proceedings of the 10th annual joint conference on Digital libraries (JCDL 2010), New York, NY, USA, ACM, pp. 89-98 (2010).

[3] Klaus, B.: Temporal Search in Web Archives, PhD Thesis, Universiat des Saarlandes, Saarbucken (2010).

[4] Lorey, J., Naumann, F., Forchhammer, B., Mascher, A., Retzla, P., ZamaniFarahani, A., Discher, S., Faehnrich, C., Lemme, S., Papenbrock, T., Peschel, R. C., Richter, S., Stening, T., Viehmeier, S. and Viehmeier, S.: Black swan: augmenting statistics with event data., Proceedings of the 20th ACM International Conference on Information and Knowledge Manage- ment (CIKM 2011), pp. 2517-2520 (2011).

[5] Allan, J.: Topic detection and tracking: event-based information organiza- tion, Kluwer Academic Publishers, Norwell (2002).

[6] Ohshima, H., Jatowt, A., Oyama, S. and Tanaka, K.: Seeing Past Rivals: Visualizing Evolution of Coodinate Terms over Time, Proceedings of the 10th Web Information Systems Engineering (WISE 2009), Poznan, Poland, pp. 167-180 (2009).

[7] 張一萌, 何書勉, 小山聡, 田島敬史, 田中克己: 時系列データに意味的に関連するニューストピックの発見, 日本データベース学会 Letters, Vol. 5, No. 1, pp. 133-136 (2006).