自 然 言 語 処 理 141－18
（2001. 1. 26）

Katz's K mixture による固有表現の異常出現の検出

仲村 大也 † 梅村 恭司 †

†豊橋技術科学大学 情報工学系
441-8580 豊橋市天伯町雲雀ヶ丘 1-1
TEL:(+81)532-47-0111(ex.5430)
mondo@ss.ics.tut.ac.jp, umemura@tutics.tut.ac.jp

概要

新聞記事中には多くの出来事 (イベント) を扱った記事が存在する. そのなかでも社会で話題となったイベント情報は有用であり, その話題性に応じて一定期間内に複数回取り上げられる. そこで, 本稿では新聞中のイベントを取り上げた記事と, その中に出現する固有表現との関係に着目し, 固有表現の統計検定によってイベント検出を行う手法を提案する. 統計検定にはポアソン分布と Katz's K mixture 分布を使用することとし, 両者の比較を行った. 後者の確率分布は実際のものに近く, 本稿で用いる場合には前者よりも適していることがわかった. また後者の確率分布を使用してイベント検出の実験を行った. その結果, 話題性があると判断できるイベントを検出することができ, イベント検出を効率的に行える可能性が確認できた.

Statistical Test for Irregular Occurrence of Names by Katz's K mixture

NAKAMURA Daiya†, UMEMURA Kyoji†

†Department of Information and Computer Sciences, Tyohashi University of Techonology
Tempaku, Toyohashi, Aichi, 441-8580, japan
TEL: (+81)532-47-0111(ex.5430)
mondo@ss.ics.tut.ac.jp, umemura@tutics.tut.ac.jp

Abstract

There are many events in newspaper articles. The event that gets attention, appears several times within certain period. This paper proposes a method to detect events by statistical test of the frequency of articles that contains a given proper name. This statistical test needs to assume some distribution. We compare Kat'z K mixture with poisson distribution for this statistical test. We have found that K mixture is closer to reality than poisson distribution and detect reasonable amount of articles. We also observed that the articles based on K mixture have more specific events than that of poisson. This suggests that statistical testing based on K mixture distribution is suitable to detect events.

1 はじめに

近年，情報のデジタル化によりテキストデータが大量に出回るようになった．それとともにその中から有用な情報を抽出するシステムへの関心が高まっている．このような状況の中，我々にもっとも身近なメディアの１つである新聞記事のテキスト化も進んでいる．新聞には日々起こったさまざまな事件や事故 (以後イベント) の記事が載っており数日間の新聞であってもかなりの量となる．さらにそれが月単位，年単位となるとそれらすべてに目を通すには相当な時間と労力を伴う．利用者がその中から興味のある情報を検索する場合，検索対象を特定するためのキーワードや，それに準じた検索要求を入力して行うのが一般的である [1]．しかし，これは利用者が明確な検索要求を持っていない場合には問題がある．例えば "過去１年間で話題となった事件について知りたい" といったような抽象的な要求しか持たない場合はキーワードによる検索を行うことが出来ない．しかし，このいわゆる"大事件"のようなある時期の新聞に頻繁に取り上げられ，社会で話題となった記事は，多くの人が興味を持ち，関心を寄せたものである [2][3]．このような情報を提示することは，明確な検索要求をもてない利用者にとって有用であり，より明確な検索イメージを持つきっかけになると考えた．このため本稿では新聞記事を対象として，その中で取り上げられたイベントの検出を行う．具体的には新聞記事から各固有表現の分布情報を抜き出し，その分布の統計検定を行うことによってイベントの検出を行う．この際の統計検定に使用する確率分布モデルとしてポアソン分布とkmixture分布を選択し，両者の比較を行った．そして，これを使用して実際にイベントの検出実験を行った．

2 イベント抽出

2.1 イベントと固有表現の関係

新聞記事中には世間で起きるさまざまなイベントを取り上げた記事 (以後イベント記事) が多数存在する，その中でも取り上げたイベントが世間で話題となったものであった場合，そのイベントの発生時期に関連して，一定期間集中的にそのイベント記事，関連記事等が出現すると考えられる．また，それぞれのイベント記事中には取り上げたイベントに関連した多くの固有表現が出現する．例えば，特定の事件を取り上げた記事中には，それが起こった場所，関係組織，被害者，加害者，といったその事件を特徴づけるようなさまざまな固有表現が記事中に出現し，同じイベントについて報じた記事や，そのイベントに関連した物事を取り上げた記事であれば同一，もしくは似通った固有表現が共通して出現すると考えられる．

2.2 固有表現の統計検定によるイベント検出

我々は上記のようなイベント記事と固有表現の関係，特徴に注目し，新聞記事中で，ある一定期間集中的に取り扱っているイベント，つまり社会で話題となったであろうイベントを検出する方法として，固有表現の統計検定を行う検出方法を提案する．その着眼点は，イベントが話題性のあるものであればそのイベントの発生時期に関連して，ある期間集中的にイベント記事が出現し，それとともにイベントに関連する固有表現の分布にもその期間での集中が観測されるとする考えである．この考えをもとに固有表現の出現分布からその出現の異常集中を捕らえることによって，逆にその固有表現に関連したイベントの存在を検出することが可能であると考えた．固有表現の出現集中を捕らえるための統計検定には，特定の分布モデルに従った検定を行うのが妥当である．出現が集中しているところは，確率分布モデルでは，その発生確率の低さから異常分布として検出されるはずである．そこで今回使用する分布モデルとして，ポアソン分布と K mixture 分布を候補として選択した．

3 固有表現の統計検定

3.1 統計検定の目的

我々は固有表現の出現集中の検出をモデルを使用した統計検定によって行うこととした．実際にはシステムが固有表現の出現集中を検出するための閾値算出を全体の出現情報を入力として与えたモデルによって行う．

図１は，1997年の毎日新聞における，固有表現 "第一勧業銀行" について日単位で集計した出現記

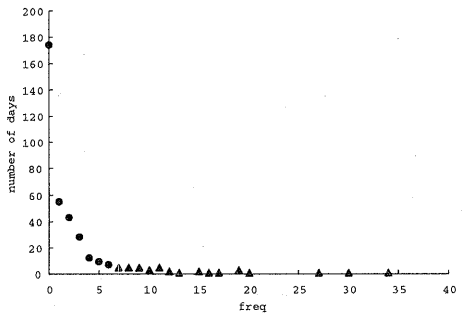

図 1: ”第一勧業銀行”の出現記事数分布

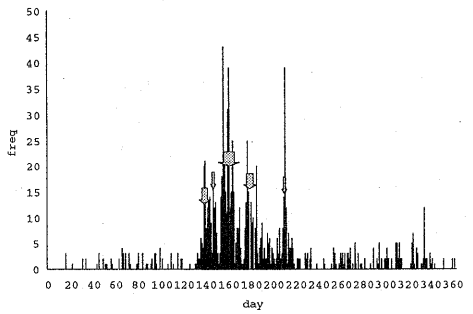

図 2: ”第一勧業銀行”の出現記数の年間推移

事数の分布を示している．この固有表現はこの年に発生したイベント ”第一勧銀の役員ら 4 人総会屋への利益供与容疑で逮捕” に関係がある．これを見ると全く出現しない日が 1 年の約半数であり 1, 2, 3 と出現記事数が少ない日もそれほど稀ではない．しかし図 1 中に▲で示した部分のような出現記事数が多い日は確率分布モデルから考えれば発生確率が低い異常な日であるといえる．イベント検出に用いる統計検定ではこのような日の検出を確率分布モデルに基づいて行うことを目的とする．その結果をもとに，このような日が連続して観測された部分を検出し，そこにイベントが出現したであろうことを推定し利用者に報告する．図 2 はさきほどの固有表現 ”第一勧業銀行” の年間出現記事数推移を示している．本稿ではこの図中で矢印で示したような，出現記事数の異常が連続して観測された部分にイベントの出現を推定する．

3.2 確率分布モデル

以下で選択した確率分布モデルであるポアソン分布と，K mixture 分布の特徴と，一般的な単語出現モデルとして用いる場合の確率密度関数を示す．それに加え，パラメータを ”日”，”出現記事数” として本稿のイベント検出に使用する場合の適用を示す．

3.2.1 ポアソン分布

この分布は，稀に起こる偶発的社会現象等がこの分布に従うものとして知られ，しばしば利用される．その確率密度関数は次のように表される．ある単語 W_i が N = 総ドキュメント数，cf = 総出現数である λ_i ，$\lambda_i = \frac{cf_i}{N}$ を取る場合，あるドキュメントに W_i が k 回出現する確率 $p_i(k)$ は poisson 分布により (1) 式で表される．

$$p_i(k;\lambda) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$

本稿のイベント検出に用いる場合は，N = 総観測日数，cf = 総出現記事数として適用する．

3.2.2 K mixture 分布

Katz's K mixture 分布は比較的シンプルな確率分布モデルであり，その確率密度関数は次のように表される [4]．ある単語 W_i が N = 総ドキュメント数，cf = 総出現数，df = 総出現ドキュメント数であるとき，あるドキュメントに W_i が k 回出現する確率 $p_i(k)$ は K mixture 分布により (2) 式で表される

$$p_i(k) = (1-\alpha)\,\delta_{k,0} + \frac{\alpha}{\beta+1}\left(\frac{\beta}{\beta+1}\right)^k \tag{2}$$

$$\begin{cases} \lambda & = \frac{cf}{N} \\ IDF & = \log_2 \frac{N}{df} \\ \beta & = \lambda \times 2^{IDF} - 1 = \frac{cf-df}{df} \\ \alpha & = \frac{\lambda}{\beta} \\ \delta_{k,0} & = 1, iff\ k = 0\ and\ \delta_{k,0} = 0 \end{cases}$$

本稿のイベント検出に用いる場合は，N = 総観測日数，cf = 総出現記事数，df = 総出現日数として適用する．この確率分布は式 [4] の $\delta_{k,0}$ からわかるように，$K = 0$ の場合は実値と同一で，特別である．それ以外の部分では $K = 1$ をピークとしてそれ以降は減少する分布を取る．

4 イベントの抽出手順

実際に新聞記事中のイベントを検出する際の具体的手順を以下に示す.

(i) 固有表現抽出
自作の固有表現抽出ツールを用い新聞記事中の固有表現を抽出する. 実際には新聞テキストから固有表現であると判断した部分に SGML タグを振る. このとき抽出対象となる固有表現は, 組織名, 地名, 人名, 固有物名である.

(ii) 出現情報の収集
抽出した各固有表現について記事単位での出現回数をカウントする. さらに出現集中を捕らえるための単位時間を日付とし, 日付けごとに出現記事数をカウントする.

(iii) 閾値の算出
確率分布モデルを使用した統計検定によって, 固有表現の出現集中を検出するための閾値を決定する. 閾値は, 1 日の出現記事数が異常に多く, モデルで算出した危険率が設定値以下となるような回数とする.

(iv) 異常出現集中の検出
決定した閾値を上回る出現数が複数日連続して現れた部分にその固有表現に関係したイベントを検出したとする. また, その情報をもとにそのイベント記事の抽出を行う. システムは, その日に掲載された記事から該当固有表の出現数, 掲載紙面をもとに, 適当な記事のドキュメント ID を出力する. 人間は, このドキュメント ID を持つ記事を見ることによってどんなイベントを取り扱っているかを判断する.

5 分布モデルの比較

実際にイベントを検出するににあたり, 1 年分の毎日新聞記事から抽出した年間の出現記事数が 50 記事以上の固有表現 1434 個の出現情報を用いて両モデルの比較を行った.

5.1 危険率の値の比較

ここでは各固有表現について両モデルが算出した確率分布を実際の分布を交えて比較を行った. 図

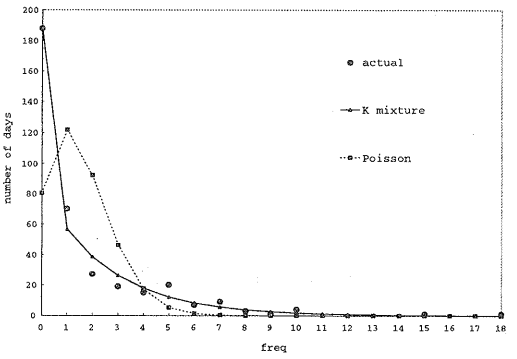

図 3: ”神戸市須磨区”の出現記事数分布とその確率分布

3 は固有表現 ”神戸市須磨区” における実際の分布と, 両モデルの確率分布をグラフ化したものである. このグラフをみると K mixture 分布により算出したものは実際の出現分布に近く, ポアソン分布による分布は実際のものとピークがずれており, 出現記事数が多い部分では急激に減少していることがわかる。ここで例として挙げた”神戸市須磨区”に限らず, ほとんどの固有表現で K mixture 分布は実際の出現分布に近い分布予想をしている. それに対してポアソン分布による確率分布は, 実際の出現回数よりも減衰が激しく, すぐに 0 に近い値となる. そのため統計検定にポアソン分布を用いた場合, 出現記事数の異常と判断する危険率を 1 %と低く設定した場合であっても人間の観測によって検出すべきではないと判断される記事に関連した固有表現を多数検出した. そのためポアソン分布で算出された危険率をそのまま用いると閾値が低くなりすぎるため, 何らかの数値を掛けるなどアドホックな加工をして用いる必要性が生じる. また, ポアソン分布はすべての語を総出現記事数だけで平等に扱ってしまうのに比べ, K mixture 分布はあらかじめ固有表現の総出現日数を入力として与えられているため Poissn 分布よりも各固有表現への適応性が高かった.

5.2 検出される単語の質の比較

次に各固有表現について実際の最多出現日における出現記事数の危険率を両モデルによってランク付けし, 比較を行った. その結果, 両モデルとも上位となる単語は図 4 に示した”ダイアナ”のような

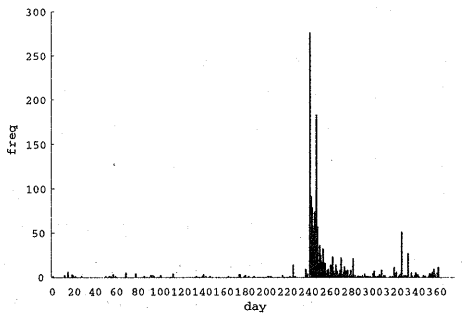

図 4: ”ダイアナ”の出現記数の年間推移

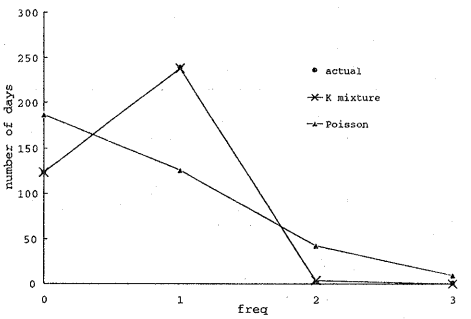

図 5: ”芦原すなお”の出現記事分布とその予想

表 1: 順位差が大きかった固有表現

ポアソン分布で高順位	K mixture 分布で高順位
日本	芦原すなお
自民党	佐藤愛子
国会	田波耕治
中央	森山良子
タイ	相生
ハウジング	佐倉
東京都	荒井義行
毎日新聞	相模原
東京	星野
米国	藤田健史
ドイツ	蓬田やすひろ
スポーツ	羽曳野
沖縄	吉井秀一
北海道	ブランチ
フランス	水野タケシ
大阪府	鴨川
長野	吹田
毎日新聞社	東京体育館
兵庫県	友利勝良
厚生省	野島孝一

出現に極端な集中日が存在するものとなるため大きな違いは確認できなかった．そこで両モデル間で順位差の大きかったものの観測を行った．表 1 は，順位差が大きかった各 20 の単語を示している．この表からもわかるように，ポアソン分布が K mixture 分布よりも高順位となる固有表現は，全期間を通じて頻繁に出現するようなありふれた固有表現が多い．一方，K mixture 分布で順位が高かった固有表現の分布を見てみると図 5に示した”芦原すなお”のように，1 記事だけ出現する日が他の日にに比べ極端に多い固有表現である場合が多い．このような単語は，総出現日数を入力としてあたえられないポアソン分布では，頻度が高いところでの確率を高く見積もりすぎてしまうため，相対順位が高くなってしまった．

この章で行った比較から我々は，今回のイベント検出に用いる確立分布としては，ポアソン分布よりも K mixture 分布が適当であると考える，

6 イベント検出実験

出現記事数の確率分布モデルとして K mixture 分布の方が優れていたとしても，それを使用して実際にイベントが検出できるかは明らかではない，そこで K mixuter 分布による統計検定によってイベントの検出実験を行った．使用するデータは 1997 年の毎日新聞全記事約 12 万件とし，閾値を決定する危険率を 5 %とし，それが 2 日以上連続して現れた期間を検出するとした．その結果，記事中より抽出された 82351 個の固有表現のうち 301 個の固有表現がイベントに関連した語として検出された．表 2 は検出された期間での最大出現記事数の危険率の低さ，つまりそのような日の発生確率の低さをスコアとしたときの上位 30 個の固有表現について示している．表の右部分はシステムが出力した記事の見出しを人間が確認しイベントとして読み取ったものを示している．表に示されたイベント群を見るとその殆んどが 1997 年に起きた大きなイベントであると人間が判断できるものである．また，ここに示した上位 30 個以外の固有表現についてもその大部分がその年に起きた何らかの大きなイベントに関するものであった．また，ファインのような固有表現抽出に失敗した語であっても検出できているものがいくつか確認できた．

7 今後の課題

- イベントのスコア
  今回の実験では，イベントの大きさを示すスコ

表 2: 上位 30 の固有表現とそのイベント内容

固有表現	イベント内容
大中	韓国大統領選、金大中氏が当選
小平	トウ小平氏、死去
ダイアナ	ダイアナ元英皇太子妃、事故死
サミット	デンバー・サミット開催
デンバー	デンバー・サミット開催
アルファ	ダイアナ元英皇太子妃、事故死 (アルファイド氏)
名護市	普天間代替ヘリポート建設での名護市民投票
橋本改造	第２次橋本改造内閣、きょう発足
エジプト	観光客狙い乱射テロ　日本人９人含む６８人殺害
香港	香港返還
広島市	原爆投下その日
ファイン	火星探査機マーズパスファインダー火星着陸
フジモリ	ペルー日本大使公邸人質事件
クーバー	日米外相会談－バンクーバー
バンカー	日債銀とバンカース・トラスト提携
山一証券	山一證券倒産
カストロ	日本大使公邸占拠
三菱自動車	三菱自工の幹部ら逮捕－総会屋へ利益供与
秋田県	秋田県知事選　新進党推薦寺田典城氏当選
海上保安庁	東京湾タンカー事故，原油量、国内最大に
内閣	第２次橋本改造内閣、きょう発足
天理	センバツ初Ｖの天理
エリツィン	デンバー・サミット，日露首脳会談
バンダイ	バンダイ、セガとの「合併」を解消
沖縄県名護市	普天間代替ヘリポート建設での名護市民投票
斎藤	オレンジ共済事件、参院出張尋問
佐藤孝行	佐藤孝行氏入閣－第２次橋本改造内閣
味の素	味の素商法違反事件
日本生命	都市対抗野球　大阪市（日生）３度目Ｖ
佐藤	佐藤孝行氏入閣－第２次橋本改造内閣

ア算出が単純であった．イベントが新聞記事で取り上げられている時間的長さ等を考慮したスコア付けを行いたい．

- イベント記事の抽出
  固有表現の検出後に行うイベント記事抽出の処理も [6] などと比べると単純であった．今後は [6] と同様に固有表現のクラスタリングを組み合わせ，スコアの算出方法についても考慮し，発展させたいと考えている．

- 特定固有表現の関連イベント
  験では１年分の全記事に対して統計検定を行ったが，その対象を特定の固有表現が出現した記事だけに限定して統計検定を行うことで，その固有表現に関係したイベントだけを検出できる可能性が有る．しかしその固有表現が出現する日の出現記事数が日によってまちまちであるので，それ対処する必要性が生じる．

- 他のメディアでの実験
  今回は対象が新聞記事であったが，同じようなメディアとして，ネット上のニュースサイト等がある．これらを対象とした実験も行いたいと考えている．

8　おわりに

新聞記事中のイベントを検出する方法として，固有表現の統計検定を用いる手法を提案した．その際統計検定に用いる分布モデルとしてポアソン分布と K mixture 分布を選択し両者の比較を行った．その結果我々がイベント検出に用いたパラメータで適用した K mixture 分布はポアソン分布と比較してより正確な分布予想を求められることが確認できた．また，その結果 K mixture 分布を用いた固有表現の統計検定によってイベントを効果的に検出する可能性が確認できた．

参考文献

[1] 木本治夫:”日本語新聞記事からのキーワード自動抽出と重要度評価”，電子情報通信学会論文誌 J-74-D-I-No.8

[2] 柳瀬隆史，仲尾由雄: “メールマガジンを利用した注目ニュースの自動抽出”，情処研報，NL136-20，2000

[3] 今井 亨，Richard Schwartz，小林彰夫，安藤彰男: “話題混合モデルによる放送ニュースからの話題抽出”，電子情報通信学会論文誌 J-81-II-No.9

[4] C.D.Manning and H.Schutze: “FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING”，PP.549，1999 THE MIT PRESS

[5] 仲村大也，梅村恭司:”IREX ワークショップ予稿集”，PP.207-210

[6] 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”，情処研報，NL113，1996