社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－162　(4)
2004／7／15

テキストデータを用いた問題の早期発見手法

宅間 大介　　野美山 浩

日本アイ・ビー・エム株式会社 東京基礎研究所
{ta9ma,nomiyama}@jp.ibm.com

概要

継続的に蓄積されていくテキストデータを用いて，問題を早期に発見する手法を提案する．問題の表現に関しては，領域知識を反映したシソーラス上のカテゴリの組み合わせによって，表現可能な問題クラスを定義し，シソーラスと文書との対応付けを行った．表現可能なクラスに含まれる問題については，問題に関するテキストデータが非常に少ない段階でも検知できる相関分析手法を適用し，その他の未知の問題については，蓄積型テキストデータの特性を考慮し，時系列解析の精度向上を実現した．また，問題報告の冗長性，問題表現の詳細さにも留意し，冗長な報告を統計的に最も情報の価値のある粒度のもの一つに絞り込む手法を提案している．実験ではPCコールセンターのコールログを解析し，特定の機種において“ACアダプタから異音がする”など4件の問題を，週ごとの解析としては最短，またはその次週の時点で警告できた．

キーワード：テキストマイニング，変化点検出，時系列解析，コールセンター

Early Problem Detection Using Text Data

Daisuke Takuma　　Hiroshi Nomiyama

Tokyo Research Laboratory, IBM Japan
{ta9ma,nomiyama}@jp.ibm.com

abstract

We propose a method to detect problems using text data added constantly. In order to express problems, we defined thesauruses which represent the domain knowledge. Problems are represented by the thesauruses, and each category in the thesauruses is mapped to documents by natural language processing. The analysis is done in two ways: correlation analysis for conceivable problem class and time series analysis for unconceivable problems. We also propose the method to select statistically most important alarm from redundant problem reports of various granularity levels. In the experiments, we applied our method to call logs of a PC call center and detected four real problems, such as noise from an AC adapter for a certain machine type.

Keywords: text mining, change point detection, time series analysis, call center

1 はじめに

従来のテキストマイニング技術は，蓄積されたテキストデータを静的なものとみなして，その中から有用な情報を抽出することに主眼を置いてきた．しかし，企業の業務で利用されることを考えた場合，何らかの変化を適時に検知するような動的捉え方も重要になる．この問題の対象となる分野は広く,WEBドキュメントやメールや新聞記事においては，新しいトピックの発見に関する研究が進められ，企業のコールセンターにおいては，不具合報告の急増や顧客からの新たな要望を検知する試みがなされている．本論文では，PCコールセンターのコールログの解析を例に，問題の早期発見という課題に対する総合的なアプローチを提示する．

早期発見へのアプローチとしては，まず適用領域の知識をテキストデータと対応付け可能な形に構造化し，相関情報を利用した解析と時系列情報を利用した解析を試みる．領域知識の構造化では，製品の分類や製品に使われている部品の種類といった知識をシソーラスとして体系化し，各カテゴリと文書との対応付けを行う．相関情報を利用した解析では，警告の対象として

表現可能な問題パターンをシソーラスのカテゴリの組み合わせによって網羅的に定義し(表現可能クラス)，製品のシリーズやサービスなど，通常同じ問題の傾向を持つと考えられるカテゴリ群の中で，特定の問題が偏って起こっているものを，データ数の少ない段階で検出する．時系列解析では，上記の問題クラスに入らない問題(表現不可能クラス)を検出するために，コールセンターの電話件数の性質に適合した単語頻度時系列モデルを用いて，非定常な増加を検知する．これら一連の解析における新たな実践は，主に以下の二点と考える．

(1) 相関分析を定期的に行うことで，時系列解析では検知できないタイプの問題を早期に発見する
(2) コールセンターの電話件数の特性を考慮して，時系列モデルを作る

実験では,PCコールセンターのコールログを用いて,(1)に対しては相関分析によって検出できる製品の問題を実際に調べ,(2)に対してはコールログの単語頻度時系列に対して，様々な仮定を行ったモデルを比較している．その結果，相関分析では，ある閾値で切った上位24件の問題報告のうち，後のデータを調べることによって問題と判断できるものが4件見つかった.また,これらは単語共起の時系列として，初めて正の値をとった時点で報告されており，問題報告としては最短で実現したことが確認された．時系列解析精度に関しては，各手法について警告を発する閾値を変化させながら，同じ警告数における警告の的中率を比較し，提案手法の優位性を示した.

2 関連研究

時刻情報を含むデータから，変化の検知を適時に行うことに関しては，対象分野，解析手法によらない問題クラスとして“Activity Monitoring”という見方がFawcettら[1]により提唱されており，その中で解決法の評価の基準として,

(1) Granularity：変化の内容を適切な粒度で報告すること
(2) Multiple alarms：同じ現象に対して複数の警告を出さないこと
(3) Benefit of timely alarms：できるだけ早期に報告すること

が挙げられている．本研究はこの三点の要件を満たすべく,(1),(2)に関しては複数の問題報告から統計的に最も意味がある粒度のものを選択する方法を論じており,(3)に関しては繰り返し行う相関分析を問題の早期でデータが少ない状況に対応させる手続きと，時系列解析において頻度スケールによらずに非定常な増加を検知する方法について述べている.

単語，係り受け，n-gram等の頻度時系列を利用したテキストデータにおける変化の検出としては，[2]，[3]が研究されている．Kleinbergは[2]において，トピック／サブトピックの階層構造を含めた頻度の急上昇(burst)の検出を行っている．時系列モデルとしては，定常状態で文書の到着間隔が指数分布に従うことを仮定しており，これは，藤木ら[3]の考察でも述べられているとおり，一定期間の文書数がポアソン分布に従うことに対応している．ただ，これらが解決する課題は状態の変化の検知を過去のデータについて行うものであり，問題の早期発見のように変化が起こった瞬間に警告を発するものとは，将来のデータの有無という点で異なる．ポアソン分布によるモデル化は，この論文の提案手法も用いているが，提案手法ではポアソン分布で平均と分散が等しいという性質を利用して精度を向上させており，その効果は実験で示されている．また，検出するトピックの階層構造については，我々が主に対象とするコールログで発見したい問題の粒度では，問題の早期の段階での頻度が低く(1〜数件程度)，データとしても1〜2区間分しか無いため，時系列解析には適さない.

一般のデータ解析での非定常事象の検知としては,[4],[5],[6]が研究されている．Zhuらは[4]において,計算区間をオーバーラップさせたHaar Waveletの係数を用いることで,異なるタイムスケールでのburst現象を非常に効率良く調べる手法を提案している．本研究が対象とする課題では，計算速度への要求はそれほど無いが，検出する問題のタイムスケールへの柔軟な対応は重要である．これについても，本論文で提案する相関分析によるアプローチは累積的な単語，係り受け頻度に基づいているため，短期間に急増する問題内容も長期間で緩やかに増加する問題内容も区別せずに扱う事ができる.

最後に,既存のテキストマイニング技術[7]での問題点に触れる．既存技術では，領域知識を取り入れる仕組みがないため，検出したい問題のクラスを定義することができない．それは即ち，統計処理を行うにしても,人手で与えた何らかの観点(相関を調べるカテゴリなど)に関する局所的な解析しかできないことを意味する．これに対し提案手法は，表現可能な問題クラスを予め定義してしまえば，解析は網羅的に行うことができる.

3 問題の早期発見

この章では，領域知識のシソーラスへの反映，カテゴリを用いた問題表現と文書との対応付け，相関を利用した問題発見手法及び時系列解析について述べる．相関分析はシソーラスを用いて定義できる問題に限定されるが，非常に少ないデータで問題を判定できる手法である．一方，時系列解析はあらゆる単語，係り受け，単語共起等を対象に，頻度の非定常な増加で未知の問題を検知するための手法である．これらは，表1に示すような特性の違いがあるため，相補的な効果が期待される.

表1 相関分析と時系列解析が対象とする問題の比較

手法	相関分析(3.2節)	時系列(3.3節)
対象とする問題クラス	表現可能クラス：シソーラスで表現可能な問題	表現不可能クラス：シソーラスでは表現不可能な問題
判定基準	単語等の共起の偏り	単語等の頻度の非定常な増加
手法の特性	緩やかな増加でも検知可能．少ないデータでも検知可能．	時系列として数期間分のデータが必要．急激な頻度変化のみに対応．

3.1 シソーラスによる領域知識の表現

この節では，シソーラスによって領域知識を表現する方法について説明する．例として，PCコールセンターにおける，

“ノートパソコンのシリーズAというタイプの製品がハードディスクを認識しない”

という問い合わせを想定してみる．この場合，キーワードとして“シリーズA”，“ハードディスク”が，係り受けとしては“ハードディスク…認識しない”が内容を特徴付けている．ここで，“シリーズA”は＜対象となる製品＞，“ハードディスク”は＜問題が生じている部品＞，“ハードディスク…認識しない”は＜不具合の内容＞という観点での情報となっている．そこで，このようなコールセンターにおける観点をツリー構造のカテゴリ(シソーラス)に反映する．図1，図2，図3はそれぞれ製品シソーラス，部品シソーラス，不具合シソーラスの例である．

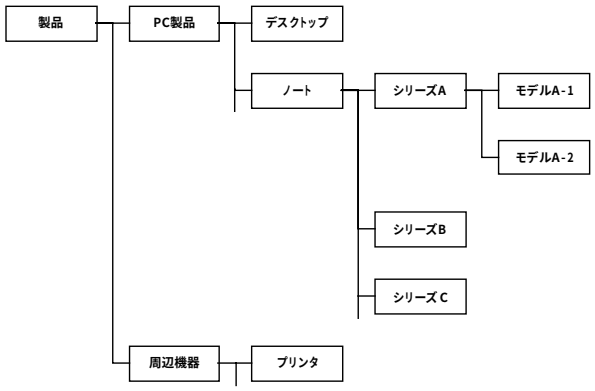

図1 製品シソーラスの例

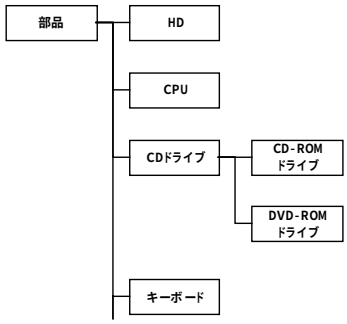

図2 部品シソーラスの例

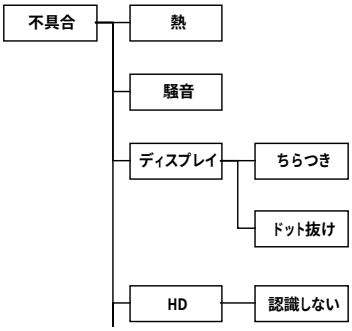

図3 不具合シソーラスの例

以上の例を一般化して，シソーラスの定義を与える．本論文で，“シソーラス”とは以下の(1),(2),(3)を満たすツリー構造を意味する．

(1) 各ノードはカテゴリに相当し，人間に認識可能なラベルが振られている
(2) 親ノードと子ノード(上位カテゴリと下位カテゴリ)は問題の表現として一般化／詳細化の関係にある
(3) １つのシソーラス内で分類の観点は一貫している

ただし(2)については，カテゴリそのものが“is-a”の関係にある必要は無く，例えば，“キーボード”カテゴリの下位カテゴリとして“part-of”の関係の“テンキー”があっても良い．これらは問題を表現する内容としては包含関係にある．

3.2 問題の表現と文書へのマッピング

前節で作ったシソーラス上のカテゴリを組み合わせると，“○○の製品で，○○の部品において，○○の不具合が生じる”という形で表されるパターンの問題を網羅的に定義できる．以下ではこれらの問題を，表現可能クラスと呼び，この節では表現可能クラスの問題を文書と対応付ける方法について述べる．ここで“文書”とは分割されたテキストデータの各分割単位を言

い，コールログならば1回のコール，メールならば一通のメール，掲示板ならば各書き込みを指すものとする．文書は予め単語や係り受けが抽出されていることを想定するが，ここではカテゴリから文書集合へのマッピングが満たすべき制約について先に述べた後，具体的な対応付け方法を定義する．

マッピングは各シソーラスについて定まるものであり，あるシソーラス：X上のカテゴリと文書集合を対応付けるマッピングT_Xは厳密には以下のように定義されなければならない．

T_X：{c|cはXのカテゴリ} → {文書の集合}

更に，カテゴリの親子関係が問題の表現として一般化／詳細化の関係に対応することを反映するための性質として，

cの直下の下位カテゴリが$c_1,c_2,\cdots,c_n$である時，
$T_X(c)=\cup T_X(c_k)$　(k=1,2,…,n)　　(性質1)

が満たされるべきである．そこで，シソーラスのリーフノードに相当するカテゴリに対し，関連する単語及び係り受けを登録する．これらは何らかのルールを定めて，合致するか否かで判定するのが現実的な方法である．不具合シソーラスの“騒音”カテゴリに分類するためのルールの例を以下に挙げる．

“騒音”，“異音”，“音”，“うるさ(い)”“音…する”，“音…鳴(る)”，“音…聞こえ(る)”，“音…大き(い)”

“…”は係り受けルール

単語と係り受けの登録により，任意のリーフカテゴリcに対して，そのカテゴリに属する単語または係り受けを含む文書集合が一意に決まるので，それをcに対応する文書集合：$T_X(c)$とする．cがリーフでない場合についても，性質1の要請により対応する文書集合は自動的に定まる．即ち，$T_X(c)$は，直下でないものも含めたcの全下位カテゴリに対応する文書集合の和集合となり，これは言い換えれば“いずれかの下位カテゴリに属する単語，係り受けを含む文書全体”ということになる．

以上により，問題を表現するためのシソーラスと，各シソーラス上のカテゴリから文書集合へのマッピングが定義された．以後では，複数のシソーラス$X_1,X_2,\cdots,X_n$に属するカテゴリの組み合わせ$\{c_1,c_2,\cdots,c_n\}$を問題の“カテゴリセット”と表記し，$\#(T_{X1}(c_1)\cap T_{X2}(c_2)\cap\cdots\cap T_{Xn}(c_n))$を問題の“頻度”として，$F(\{c_1,c_2,\cdots,c_n\})$で表すことにする．

3.3 表現可能クラスの問題の相関分析

ここでは，シソーラス上のカテゴリの組み合わせによって表現できる表現可能クラスの問題を早期に検知し，冗長性を取り除いて報告する方法について述べる．従来使われている時系列解析による問題警告では，単語の頻度データが一定期間以上蓄積されない限り，データ数の少ない問題の警告は困難である．また，製品やサービスに関する問題の記述の多くは，新製品のリリースや新しいサービスの開始の直後からテキストデータに現れる．しかし，問題頻度の単純な増減分析だけでは，新しい製品やサービスに関連するカテゴリセットの頻度増加が大量に検知されてしまい，問題として警告する必要のあるカテゴリセットの頻度増加もそれに隠れてしまう．そこで，この節では増減分析ではなく，特定の製品カテゴリにおいて，偏って多く起こっている問題の早期発見方法を提案する．

以後では，再びPCコールセンターにおける製品の問題の例を用いることとし，製品シソーラスは$X_{Product}$という記号を用いて特別に扱うものとする．まず，$X_{Product}$上のカテゴリで，それ以下の下位カテゴリで問題の傾向が似ていると考えられるカテゴリc_{base}を定める(図4参照)．これは自動的に決めることも不可能ではないが，ここでは人間が決定するものとする．c_{base}を含む各カテゴリセット$\{c_{base},c_1,\cdots,c_n\}$について，製品カテゴリを除いた問題の内容を表す部分$\{c_1,c_2,\cdots,c_n\}$をC_λ($\lambda\in\Lambda$：問題部分のカテゴリセットの全パターン)と表すことにすると，平均的な問題の傾向を表す指標として，

$F(\{c_{base}\}\cup C_\lambda)/F(\{c_{base}\})$　　$\lambda\in\Lambda$

を計算できる．同様にc_{base}より下位の各製品$c_{product}$(図4参照)についても，問題の傾向

$F(\{c_{product}\}\cup C_\lambda)/F(\{c_{product}\})$　　$\lambda\in\Lambda$

を計算できる．これらを用いて，製品$c_{product}$における問題C_λの相対頻度を

$$R(C_\lambda,c_{product},c_{base})=\frac{F(\{c_{product}\}\cup C_\lambda)/F(\{c_{product}\})}{F(\{c_{base}\}\cup C_\lambda)/F(\{c_{base}\})}$$

と定義する．

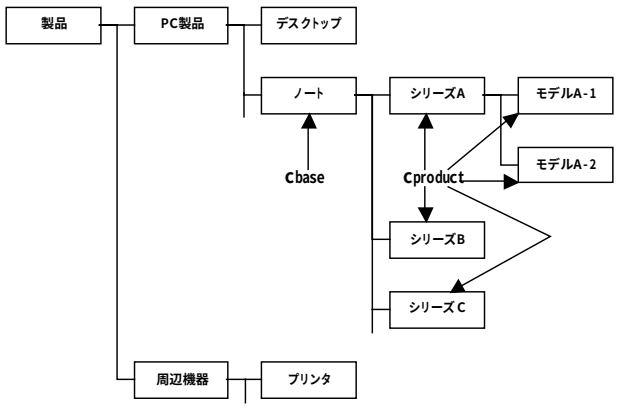

図4 製品シソーラス$X_{product}$における$c_{base},c_{product}$の例

以後の議論では$c_{product},c_{base}$を固定し，相対頻度を問

題のカテゴリセットC_λのみの関数として$R(C_\lambda)$と略記する．この相対頻度を，製品がリリースされた時点からデータが追加されるごとに繰り返し計算していくことで早期の段階から問題を調べることができる．ただし，ここで必ず考慮しなければならないのは，早期発見のために解析が重要となる製品リリース直後では，まだ相対頻度の値が信頼性を持つのに十分なデータが揃っていない可能性が高いことである．そのため，少ないデータによる偶然の偏りを検知することの無いように，相対頻度の区間推定を行って信頼区間を求め，その左端の値を問題の検定，あるいはランキングに用いる．上記の相対頻度は，$F(\{c_{product}\})$，$F(\{c_{base}\})$が十分大きいと考えることにより，“二項分布／二項分布”とみなすことができるため，現実的には二項分布の信頼区間計算をすれば良い．以下，$R(C_\lambda)$の信頼区間の左端の値を$R^*(C_\lambda)$と記す．

このようにして偏って起こる問題と製品のカテゴリセットを取り出した場合，次に課題となるのが，問題報告の冗長性の排除である．ある製品において，問題表現用シソーラス(製品以外のシソーラス)のカテゴリcの相対頻度が高い時,そのカテゴリよりも下位のカテゴリも平均的にcと同じ相対頻度を持つ．また，問題表現用の2つのシソーラスのカテゴリc_1とc_2の相対頻度が高い時，組み合わせ $\{c_1,c_2\}$ の相対頻度も必然的に高くなる．これはc_1とc_2が関連性を持っていない場合でも起こるため，問題内容の誤解に繋がる可能性がある．例えば，マウスの問題とファンの騒音の問題が起こっている場合，{“マウス”}と{“騒音”}という2つのカテゴリセットの相対頻度はそれぞれ高い値になり,2つの問題が独立であったとしても共起のカテゴリセット{“マウス”,“騒音”}の相対頻度は高くなってしまう．以上を踏まえると，シソーラス上のカテゴリセット2組が詳細化／一般化の関係にある場合,次のような3つのパターンに分類できる．

(a) シソーラス上でカテゴリが上位下位の関係にある場合
カテゴリセット1：{“不具合/ハードウェア/ポインティング・デバイス”}
カテゴリセット2：{“不具合/ハードウェア/ポインティング・デバイス/マウス”}
→カテゴリセット1よりカテゴリセット2が詳細.

(b) 一方に別のシソーラスのカテゴリが追加されている場合
カテゴリセット3：{“部品/ハードウェア/ファン”}
カテゴリセット4：{“問題/ハードウェア/騒音”}
カテゴリセット5:{“部品/ハードウェア/ファン”,“問題/ハードウェア/騒音”}
→カテゴリセット3, カテゴリセット4よりカテゴリセット5が詳細.

(c) (a),(b)の組み合わせ

これらを一般化して，2つのカテゴリセット $C=\{c_1,c_2,\ldots c_n\}$と$D=\{d_1,d_2,\ldots d_m\}$に対し，問題の粒度に関する半順序を以下で定義する．

DがCより“詳細”(⇔ CがDより“一般的”)
⇔ $\forall c_i \in C$に対し，$\exists d_j \in D$ such that d_jはc_iの下位カテゴリであるか等しい

これは意味的には上記の(a),(b),(c)のいずれかの関係が成り立つことに相当する．この半順序によって比較可能なカテゴリセット同士は問題警告としては冗長になると言える．そのため

カテゴリセットC,Dが“粒度において比較可能”
⇔ CがDより“詳細”またはCがDより“一般的”

である全てのC,Dのペアについて問題報告の優先規則“≧”が定義される必要がある．以下，DがCより“詳細”と仮定する．

(1) Dが(a)の意味で，Cより“詳細”である場合
この場合，m=nで，Dの各カテゴリは全てCのいずれかのカテゴリの下位なので，d_iはc_iの下位(i=1,2,…,n)と書ける．よって，各c_iを下位に下ることで該当する製品との相関が新たに生じない限り，Dの相対頻度とCの相対頻度は等しくなる．よって，Cより“詳細”なDを優先するのは，DがCに比べて高い相対頻度となる時とするのが妥当である．そこで，以下のように“≧”を定義する．

$R^*(D) \geqq R^*(C) \Rightarrow D \geqq C$ (DをCより優先)
$R^*(D) < R^*(C) \Rightarrow C \geqq D$ (CをDより優先)

R^*は信頼区間の左端の値を取っているため，Dの方が対応する文書が少ないことによって偶然の偏りが生じ易いことも考慮されており，上記の式には統計的な有意性もこめられている．

(2) Dが(b),(c)の意味で，Cより“詳細”である場合
E=(DとCで,カテゴリの上位下位の違いを無視した差分のカテゴリセット)とする時,以下で優先度を定義する．

$R^*(D) \geqq R^*(C) \times R^*(E) \Rightarrow D \geqq C$
$R^*(D) < R^*(C) \times R^*(E) \Rightarrow C \geqq D$

$R^*(C) \times R^*(E)$はCとEが独立だった場合のDの相対頻度に相当する．

この定義において，“粒度に関して比較可能”なカテゴリセットC_1,C_2,C_3に対して

$C_1 \leqq C_2,\ C_2 \leqq C_3 \Rightarrow C_1 \leqq C_3$

を満たしていることが数学的に確かめられる．よって“粒度に関して比較可能”なカテゴリセット同士については比較の仕方に依らずに優先度が定義されている．

そこで，冗長なカテゴリセットに対し，この優先度によって自身より優先度の高いものが無いもののみを警告する．これにより，粒度に関して比較可能な複数の問題が報告されることはなくなり，真に相関があるとみなせる問題のみが適切な粒度まで詳細化される．

3.4 表現不可能クラスの問題のための時系列解析

問題の早期発見というタスクにおいては,3.1節で定義した表現可能クラスの問題に合致しない未知の問題(表現不可能クラス)も検出できることが望ましい．表現不可能クラスについては，従来の多くの手法と同様に，何らかの単語や係り受けの頻度の増加を検知する方法を用いる．以下では，テキストデータをコールログに限って，時系列解析の方法を論じる．

時系列解析は，観測された時系列を一般的な関数ではなく，モデルとなる何らかの関数集合の元（あるいは確率過程として何らかの性質を持つもの）であると仮定することで非自明な予測を可能にしている．そのため，モデルを構築するにあたっては，対象とする時系列の性質を十分に検討しなければならない.例えば,掲示板やメールでは発言者に以前のテキストが見えるため，頻度時系列の過去の値が作成者に対し直接フィードバックを与える可能性があるのに対し，コールセンターでは電話をかける人が過去のコールログにおける単語等の頻度の影響を受けることはない．また，コールセンターでは，いかなる状態においても電話の頻度はポアソン分布に従う揺れを持つ．提案手法では,これらの性質を考慮し,モデルを構成するに当たって,時系列に以下を仮定することが有用であり，実データと比較しても妥当であると判断した.

(1) 時系列は定常状態と非定常状態とを遷移する．定常状態 → 非定常状態で頻度が増加することが多い.
(2) 定常状態では，各タイムスパンでの頻度はパラメータの等しい独立なポアソン分布に従う.

(1)は経験則であり，(2)はコールセンターでは電話をかける人に過去の頻度のフィードバックが無いこと(独立性)と，定常状態のノイズがポアソン分布に従うことを反映している．問題の検出には定常状態から外れる瞬間を捉えれば良いので，過去のデータから(2)の分布を推定できれば良い．ポアソン分布は，平均であり分散でもあるパラメータ(λとおく)により定まるので，より収束の速い平均の推定値によってλを推定する．即ち，現在までの頻度時系列がF_1，F_2，…F_{n-1}の時，λをF_1，F_2，…F_{n-1}の平均として$Po(\lambda)$を最新データF_nの予測分布とする．実際には，データが古いほど定常性が崩れている可能性が高いことを考慮して，近い過去のデータを重視する加重平均を使う．F_nの“非定常性”（予測平均より大きい場合に限る)の指標は,予測分布に対するF_nの情報量を採用するのが自然と言える.

以上で，コールセンターデータの特性も考慮して，単語，係り受けの頻度の非定常性を評価する方法が定義された.

4 実験

4.1 表現可能クラスの問題の相関分析

3.3節で述べた定義済み問題の検知手法を検証するために,PCコールセンターのコールログ20ヶ月分(約35万文書)を実際に解析した.相対頻度の区間推定における信頼係数は90%を用いた．その結果，R^*を閾値4で切った上位24件のカテゴリセットのうち，実際に警告として価値があると判断できるものが4件あった．表2に問題の内容と報告されたカテゴリセットを示す．問題と判断しなかった報告には，報告後に相対頻度が下がったものの他，特定の機種にのみ取り付けられているコンポーネント名に関するものなどが挙げられる.

表2 実験で検出された問題

問題内容	問題カテゴリセットのラベル
ACアダプタから異音がする	機種/○○○ 部品/ハードウェア/電源/バッテリ・ACアダプタ 問題/ハードウェア/騒音
ディスプレイにドット抜けがある	機種/××× 部品/ハードウェア/ディスプレイ 問題/ハードウェア/ディスプレイ/ドット
ハードディスクを認識しない	機種/■■■ 問題/ハードウェア/ハードディスク
ネジ穴に関する問い合わせが多い	機種/▼▼▼ 部品/ハードウェア/ネジ

警告時期については，ACアダプタの騒音，ドット抜け，ネジに関する問い合わせ及びハードディスクの認識の問題は週次の解析としては最短で検知できており,電源の発熱の問題は他と比べ判定にやや遅れて検知されている.

冗長な問題報告の排除処理については,{電源},{バッテリー／ＡＣアダプター},{電源 ,騒音},{騒音},{バッテリー／ＡＣアダプター,騒音}から{バッテリー／ＡＣアダプター,騒音}が選ばれるといった効果が見られた.

4.2 表現不可能クラスの問題のための時系列解析

3.4節で提案した時系列解析の精度については，4.1節と同じコールセンターデータの“一般名詞(その他)”,“固有名詞”に分類される400語の週ごとの頻度時系列89週間分を用いて比較検証した.“一般名詞(その他)”は分類辞書の分類に当てはまらないものであり，専門用語の可能性が高い.表3は,提案手法と比較実験に用いた3つの手法について,増加を判定するのに用いられる指標と，各手法の時系列に対する仮定をまとめたも

のである．

表3 比較実験に用いた時系列解析手法

手法	増加判定の指標	時系列に対する仮定
提案手法	過去の平均値をパラメータとするポアソン分布における最新の頻度の情報量	各週の頻度は独立同分布なポアソン分布に従う
差分	今週の頻度 - 先週の頻度	時系列は連続的に変化する
正規化	(今週の頻度 – 平均)／分散	各週の頻度は独立同分布に従う
畳み込み	加重平均の増加分	大きなスケールで見て連続的に変化する

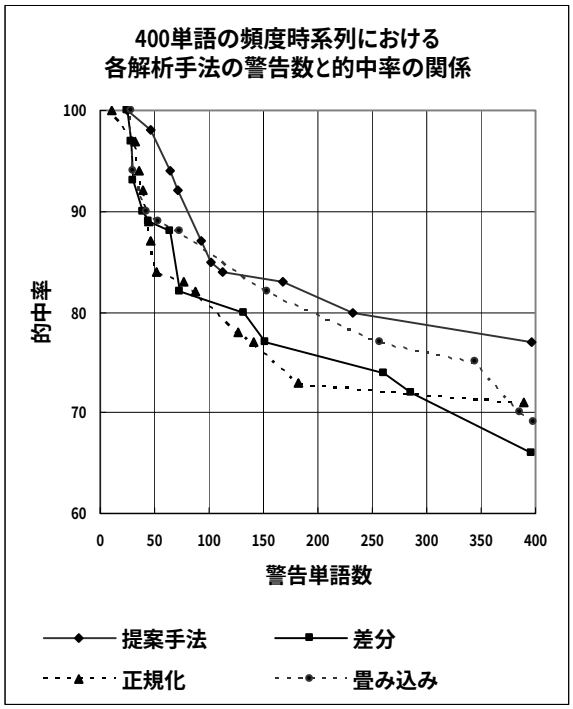

図5 400単語の頻度時系列における各解析手法の警告数と的中率の関係

警告の有無は，増加判定の指標が閾値を越えるか否かで決まるため，比較実験では各手法について閾値を変化させながら警告された単語数と的中率の関係を調べた．的中率については，警告週の前後5週間の平均頻度を比べて，後5週間の方が大きければ“的中”とした．また，多重警告回避のため，どの手法も1つの単語に対する警告は最初の1回(初めて閾値を越えた時)のみとした．グラフ4.1に比較実験の結果を示す．このグラフから提案手法では，差分手法，正規化手法と同じ警告の的中率を約倍の警告数で実現でき，畳み込み手法と比べても優れていることが確認できる．

しかし，実際に警告された単語は，非定常な増加は見られても，情報としては価値の低いものが多く，新しいコールテーカーの名前，特定のコールテーカーの口癖，支払い・発注方法の変更に伴う新しい専門用語などが大量に検知された．有用と思われた情報としては，特定の不具合でしか表示されないエラーメッセージ，新種のウィルス名が挙げられる．

5 考察

5.1 表現可能クラスの問題の相関分析

警告の指標が高い値を示しているカテゴリセット中に，実際に問題が見つかった割合は，実用的な観点から見て十分に高かったと言える．また，警告はカテゴリセットの頻度の値が正となる第一週目から出されているので，報告としては最短と言える．当然第一週目の時点ではカテゴリセットの頻度時系列に特徴を見出すことはできないため，従来の時系列解析を用いた手法で同じ問題を検知するのは困難である．

警告のタイミングについては，カテゴリセットが問題に特異的にカウントされる場合は早い時期に警告でき，カテゴリセットに対応する単語や係り受けが一般的に頻出するものでは警告が遅れやすい傾向が見られた．そのため，適用領域において問題がうまく分類されていること，即ち問題うまく表現できるシソーラスを準備することが，本手法の精度向上のために重要と言える．

5.2 表現不可能クラスの問題のための時系列解析

時系列解析の非定常性の検知精度については，図5の結果から，提案手法の優位性は明らかである．比較手法を個別に見ていくと，差分手法は時系列に関する仮定が短絡的であり,本質的な変化と3.4節で述べた電話件数が内在的に持つノイズが区別されていないために精度が悪いと考えられる．正規化手法は，仮定は妥当だが，指標の分母の分散を極めて収束の遅い不偏標本分散で推定しているためと考えられる．提案手法が独立同分布を特にポアソン分布と仮定することによる優位性は，この分散推定値の収束速度によるところが大きいと考えられる．畳み込み手法の精度が比較的良いのは，仮定が妥当であることと，畳み込む関数をデータに合わせて十分にカスタマイズしたことによると考える．また，畳み込みは平均化の効果を持つため，電話件数のノイズ除去にも寄与している．

一方，実用的側面を見ると，定義済み問題の検出手法の導入により製品リリースに伴う単語群を除いても，尚，有用な警告の密度は低く，不要な単語が大量に検出される問題は解決されていないと言わざるを得ない．

6 結論

本研究ではテキストデータを用いた問題の早期発見というタスクに対し，シソーラスを用いた領域知識の利用，カテゴリとテキストデータの対応付けを土台として，相関と時系列の二種の情報に基づいたアプローチを試みた．相関情報は従来，問題の早期発見の用途では利用されていなかったが，(1)シソーラスと組み合わせて用いること，(2)カテゴリの組み合わせによって人間に分かりやすい報告をすること，(3)定期的に計算を行うこと，(4)データ量を考慮した統計処理を導入すること，により実用レベルで有用な結果を得ることができた．時系列解析では，コールログの特性を反映したモデルを再構築し，解析精度を改善できた．

上記の二つのアプローチを併用することによる相補的な効果については，それぞれが表現可能なクラスの問題と表現不可能なクラスの問題を発見できているという点では意味があったと言えるが，シソーラスを分けることにより，頻度増加を警告する必要の無い単語を時系列解析の対象から除くという点では，効果を実感できるレベルには達していない．今回の実験で警告された単語群を見る限り，警告された場合に有用な単語と有用でない単語を，シソーラスレベルで予め定義するのは極めて困難であり，時系列の特徴から判断することもほぼ不可能と考えられる．そのため，時系列解析ついても何らかの言語的な情報と組み合わせて精度を上げることが必要となると推測する．

本研究で領域知識が問題の早期発見に非常に有効であることが分かったが，今回の実験で用いたシソーラス及びルールの構築コストは比較的小さかった．これは，コールログの情報が意味的に狭い範囲に限られており，単語，係り受けの抽出ルールが比較的単純なものでも機能したことと，製品分類など担当部門によって既に作成された情報を有効利用できたことによるところが大きい．

今後は，シソーラスを半自動的に構築することと，シソーラスを様々な用途に活用することで構築コストあたりのリターンをより高めていくことが必要であると考える．

7 参考文献

[1] Tom Fawcett and Foster Provost. Activity Monitoring: Noticing interesting changes in behavior, In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 53-62, 1999.

[2] Jon Kleinberg. Bursty and hierarchical structure in streams. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[3] 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，pages85-92.

[4] Yunyue Zhu and Dennis Shasha. Efficient Elastic Burst Detection in Data Streams. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge
Discovery and Data Mining, pages 336-345, 2003.

[5] Yunyue Zhu, Dennis Shasha. Statstream: Statistical Monitoring of Thousands of Data Streams in Real Time. In Proceedings of 28th International Conference on Very Large Data Bases, pages 358-369, 2002.

[6] K. Yamanishi and J. Takeuchi. A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data, In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[7] Tetsuya Nasukawa, Tohru Nagano. Text Analysis and Knowledge Mining System. IBM SYSTEMS JOURNAL, VOL 40, NO 4, 2001.