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

2006－ICS－145 (7)
2006／10／25

知的Webサービスのための XML文書からの情報抽出

大川原雄也[†a)]　大園 忠親[†b)]　新谷 虎松[†c)]

Information Extraction from XML Docments for Intelligent Web Service

Yuya OKAWARA[†a)], Tadachika OZONO[†b)], and Toramatsu SHINTANI[†c)]

Abstract. 本稿は，記事を対象に記事間の関係を特徴付けるパターンを発見し，記事のトピックの抽出およびトピックの追跡を目的とする．本手法では，記事の言語モデルを学習して抽出を行う．言語モデルにはクラスモデルを用い，クラスモデルによって記事をクラスタリングする．トピックの抽出は，作成されたクラスタにラベル付けをすることに相当する．クラスタのラベルは，クラスタに属する記事の内容を表す語のうち，重要度が高い語で構成される．これらの語は，クラスモデルを用いたクラスタリングの際に抽出可能である．トピックの追跡は，作成されたクラスタ間の類似度を計算し，類似度が高いクラスタを時間軸上に配置することで行われる．

Keywords. トピック抽出，トピック追跡，クラスタリング，言語モデル

1. はじめに

本稿では，文書群から作成された言語モデルおよびXMLの意味情報を用いたXML文書からの情報抽出について述べる．近年大量の文書の配信や交換がネットワークを介して盛んに行われるようになった．今後も文書として配信される情報量はさらに増加し，文書データから必要な情報を発見することはより困難になると考えられる．種々の文書データの中でも，ニュース記事などのような時系列的に文書を配信する文書ストリームのコンテンツ分析技術の重要性が増加している．CRM，ナレッジマネージメントおよびWeb監視といった文書ストリームを扱う分野で，トピックの傾向や移り変わりを分析することが重要なポイントになっている．これまでに，文書データに含まれるトピックの検出，出現場所の発見等に関する研究などが行われている．

現在，Yahoo!, gooなどのポータルサイトで，様々なトピックに関するニュース記事を閲覧することができる．Yahoo!トピックス(注1)では，ニュース記事を1000個近くのトピックに分類し，一覧をリストで表示している．しかし，リストが示す内容を一見しただけでどのような事象が生起したかを理解するのは困難であり，同じトピックに属するニュース間の関連を理解するのは困難である．

この問題を解決する手法の一つとしてクラスタリングがある．ニュース記事をクラスタリングし，クラスタを分析することでトピックを抽出・追跡することが可能になる．また，クラスタの内容を表すラベルをクラスタに付加することでトピックの流れが一見して理解することが容易になると考えられる．このように，ニュース記事などのトピックを分析する研究として，TDT(Topic Detection and Tracking) プロジェクト [1] がある．TDT プロジェクトでは，ニュース記事などのデータからトピックの検出・追跡または新規トピックの検出などに関した議論がされている．

ニュース記事のような文書データは，社会で生起した事象が時間順に記述される文書であり，社会の縮図とも考えられる．これらの記事を時間順に並べてみると，単一事件に関わるものが多く，人間にとって整理することが容易になる．生起した事象を追跡して相互の関連性をとらえれば，全体を理解することが容易になる．本稿では，ニュース記事からのトピックの抽出・追跡，事象の予兆の抽出を目的とする．トピックの抽出は，クラスタリングによって作成されたクラスタにラベル付けをすることで実現される．本稿では，クラ

† 名古屋工業大学情報工学専攻，〒466-8555 名古屋市昭和区御器所町
a) E-mail: yuya@ics.nitech.ac.jp
b) E-mail: ozono@nitech.ac.jp
c) E-mail: tora@nitech.ac.jp
(注1)：http://dailynews.yahoo.co.jp/fc/

スモデルという言語モデルの一種を用いクラスタリングを行った．クラスモデルを用いることで，クラスタリングとラベル付けのためのキーワードの抽出が同時に行える．トピックの追跡は，クラスタの類似度を算出し，類似度の高いクラスタを関連づけて時間軸に沿って並べることで可能となる．

2. 関連研究

TDT をはじめ，ニュース記事から自動的にトピックを分析する研究は数多くある [4] [7]．トピック内のニュース記事の間の内容的な類似，相違の発見を行いながら，記事の時間変動を提示する研究が行われている [4] [5]．記事の内容だけでなく，記事の出現頻度の時間変動を利用し，内閣の支持率などの時系列データと関連のあるトピックを発見する研究もある [3]．本研究では，社会・経済などのジャンルを問わない様々なトピックが混在した記事群を関連のあるクラスタにまとめてトピックの検出を行い，クラスタ間の類似度を計算して時系列的に提示してトピックの追跡を行う．

トピックの統計的モデル化に関する研究として，有限混合モデルを用いたトピック分析がある [6]．この研究では，確率的トピックモデルと呼ばれる，文書の単語の確率分布を複数のトピック上の単語確率分布の線形結合としてモデリングしたものを用いてトピック分析を行っている．この研究では，モデルが時間的に変化しないと仮定し，最適なモデルを選択することでトピックの同定を行っている．最適なモデルが時間的に変化すると仮定して，各時刻における最適なモデルを選択する研究もある [2]．これらの研究では，トピックの検出とトピックのラベル付けの問題は別々に扱われている．これに対し，本研究では，記事中の単語の統計情報を用いてクラスタリングを行い，クラスタリングにある閾値以上の影響を与えた単語をクラスタのラベルの候補とする．すなわち，本研究では，クラスタリングと同時にクラスタのラベルの候補を抽出することができ，トピックの検出とラベル付けをより統一的に扱っている．

3. クラスタリングによるトピック分類

本稿で対象にしているニュース記事は，様々なトピックから構成されるため，各記事がどのトピックに属するか判断しなければトピックの追跡ができない．また，トピックは時間の経過によって事象間の関連が薄れたり，新しいトピックが発生する．したがって，様々なトピックが混在するニュース記事集合からトピックを検出・追跡するためには古い統計量を捨て新しい統計量を取り入れて処理を行う必要がある．

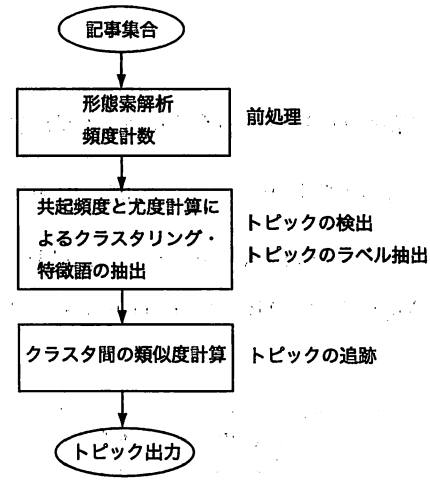

図 1 トピック分析システム

トピックの抽出と追跡を行う TDT の分野において，時間軸に沿ってクラスタリングすることが効果的であることが知られている．時間軸に沿って作成されたクラスタは，事象に対応することが多い．そのため，本稿では，記事を時間軸上に配置し，クラスタの結合を時間軸上の一定の期間で制限する．以上のように時間軸に沿ってクラスタリングを行うことで，様々なトピックが混在するニュースをトピックごとに分類し，時間に沿ってトピックの流れを把握することができる．

本稿では，クラスタをラベル付けすることでクラスタを表現する．$w_1, w_2, \cdots, w_n$ をクラスタの内容を表すラベルの集合とすると，クラスタ $C_i = w_1, w_2, \cdots, w_n$ と表現する．クラスタに属する記事の重要語のうち，クラスタの内容に関わりが強い語がラベル w_j として選択される．図 1 にシステムの流れを示す．

本稿では，クラスモデルに基づいた手法によってクラスタリングを行う．クラスモデル [9] は，学習用データから得られた情報をもとに単語を自動的に分類し，その分類結果を用いた言語モデルである [8]．クラスモデルでは，学習データにおけるクラスモデルの対数尤度を最大化させるクラス分類を最適な分類とする．この分類結果を，推定だけでなく他の用途に利用することも可能である．本稿では，クラスモデルによる記事の分類結果をクラスタリングに利用した．単語を分類

し，言語モデルとして利用する利点は，N グラムモデルでは十分に学習できない，学習パラメータが少ない場合でもクラスモデルは推定すべきパラメータ数が少ないため，有効な言語モデルとして用いることができる点が挙げられる．単語 w_n の属するクラスを c_n とするとき，クラスモデルは以下のように表される．

$$P(w_n|w_{n-1}) = P(w_n|c_n)P(c_n|c_{n-1})$$

確率 $P(w_n|c_n)$ は，単語 w_n がクラス c_n から生起する確率であり，次式により推定できる．

$$P(w_n|c_n) = \frac{N(w_n)}{N(c_n)}$$

ここで，$N(w_n)$ は，学習データ中で単語 w_n が出現した回数であり，$N(c_n)$ は，クラス c_n の単語が出現した回数である．

クラスモデルでは，学習データの対数尤度を最大化するクラス分類をより最適な分類と考える [9] [10]．クラスモデルにおける学習データの対数尤度 $L(\pi)$ は以下のように計算できる．

$$L(\pi) = \sum_{c_1,c_2} C(c_1,c_2)\log\frac{C(c_1,c_2)}{C(c_1)C(c_2)} + \sum_{w} C(w)\log C(w)$$

$C(c_i)$ は，クラス c_i の単語が出現する回数である．右辺の第2項は，クラスに依存しないため，第1項を最大化することで最大の $L(\pi)$ が求まる．

クラスモデルでは，各単語に1つのクラスを割当て，$L(\pi)$ を最大化するようにクラスを次々に併合していく方法でクラス分類を求める．以下にクラス分類を求めるアルゴリズムを示す．

（1） すべての単語に対して1つのクラスを割り当てる．

（2） 各クラスの単語を，他のクラスに移動したときの $L(\pi)$ を計算する．

（3） $L(\pi)$ を最大化させるクラス C_m を求める．

（4） 単語をクラス C_m に移動する．

（5） ステップ2〜4を $L(\pi)$ が収束するまで，または決められた回数繰り返す．

3.1 記事のクラスタリング

3.1.1 提案手法

本稿では，クラスモデルのクラス分類に基づいたクラスタリングを行う．提案する手法では，記事のクラスタリングの結果からクラスタの内容を表すラベルの候補を同時に抽出可能である．

通常，クラスモデルは，データ中のすべての単語を対象に処理を行う．本稿では，記事を対象としている．記事のタイトルは，人手によって，記事内容を最大限に要約したメタデータである．本稿では，その特性を利用し，タイトル中のすべての名詞および本文中の重要度が高い名詞を対象に処理を行う．名詞を抽出するための前処理として，形態素解析を行う．形態素解析には茶筌(注2)を用いた．重要度は，TF-IDF によって計算する．重要度が高い語は，記事の本文での重要度が上位の語のことである．

また，通常のクラスモデルでは，クラスのメンバは単語であるが，記事をクラスタリングするために，クラスモデルにおけるクラスに属するメンバの単位を記事とする．クラスの移動は，前述の語群を用いて，対数尤度 $L(\pi)$ の評価をすることで行われる．$L(\pi)$ を最大化させる移動先が求まったとき，記事をその移動先に移動する．

記事 A_i のタイトル中のすべての名詞および本文中の重要度が高い名詞の集合を W_i とする．対数尤度 $L(\pi)$ の計算は，他のクラス C_k に対して，W_i に含まれる語と，$W_j(A_j \in C_k)$ に含まれる語のすべての組み合わせに対して行われる．W_i の語数を m，$W_j(A_j \in C_k)$ の語数を n_k とするとこの組み合わせは mn 個ある．他のすべてのクラスでこの組み合わせに関して対数尤度の評価を行う．通常のクラスモデルと同様に，対数尤度を最大化させるクラスを求め，クラスのメンバである記事をそのクラスへ移動する．

以下にクラスタリングのアルゴリズムを示す．図2にクラスタリング処理の例を示す．

（1） 各記事にそれぞれ1つのクラスを割り当てる．

（2） 各記事のタイトル中の名詞および本文中の重要度が上位の語を抽出する．記事 Ai より抽出された，これらの語の集合を W_i とする．

（3） W_i 中の語を，他のクラスに移動したときの $L(\pi)$ を計算する．

（4） $L(\pi)$ を最大化させるクラス C_m を求める．

（5） 記事 A_i および W_i をクラス C_m に移動する．

（6） i を変更し，適当なクラス数になるまでステップ3〜5を繰り返す．

(注2)：http://chasen.naist.jp/hiki/ChaSen/

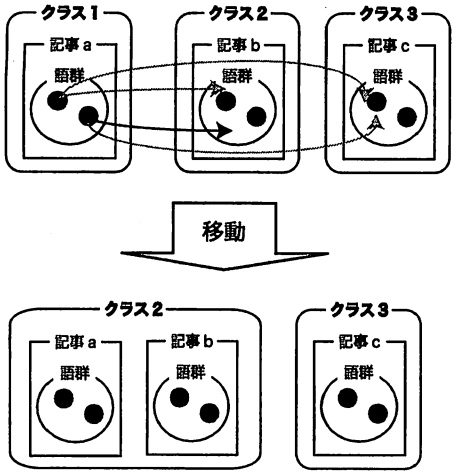

図2 クラスタリング処理の例

3.2 クラスタのラベル付け

記事 A より抽出された W 中に含まれる，語 t のクラス移動に伴う対数尤度 $L(\pi)$ の変量を ΔL とする．閾値 α に関して，$\alpha < \Delta L$ を満たす語 t を，語 t が属するクラスタおよび記事 A の内容を表す重要語とする．語 t は，常識的・基本的な高頻度語と共起する語である．こうしてクラスタリングと同時に抽出された語をクラスタのラベルの候補 W' とする．

次に，$w \in W'$ の重み付けを行い，重要度が低い語をフィルタリングする．まず，語 w の出現確率 $tfp(w)$ は以下のようになる．

$$tfp(w) = \frac{tf(w)}{n}$$

n は，単一文書中の単語数を示す．すなわち，tfp は正規化された出現頻度を表し，文書の語数の違いによる頻度差を考慮している．次に，w の全文書における出現確率 $gfp(w)$ を計算する．$gfp(w)$ は以下のようになる．

$$gfp(w) = \frac{1}{N}\sum_{i=0}^{N} tfp_i(w)$$

N は，全文書数を表す．ここで，gfp が高い語には，ニュース記事に広く使用され頻出する語も多数存在するため，高い順にラベルと判定すると，ニュース記事に特有の語がラベルとして多く登録されてしまう．そこで，ニュース記事に特有の語を考慮するため，以下の式によって再度重み付けを行う．

$$keyword(w) = \begin{cases} gfp(w), & (\frac{df(w)}{N} \leqq \alpha) \\ (1 - \frac{df}{N}) \cdot gfp(w), & (\frac{df(w)}{N} > \alpha) \end{cases}$$

閾値 α は，0 から 1 の範囲の定数である．また，$\frac{df(w)}{N}$ は，語 w が文書全体に対して出現する確率を表す．この重み付けにより，ニュース記事特有の語の重要度が上位になる問題を解消した．以上の重要度計算を行い，重要度が上位の語をクラスタのラベルとする．

3.3 トピックの追跡

トピックの追跡は，クラスタの類似度を計算して行う．クラスタに付加されたラベルを用いてクラスタ間の類似度を計算し，類似度が高いクラスタを関連付ける．古い情報を捨て，新しい情報を取り入れて追跡を行うために，クラスタの時間間隔を重みとして類似度計算に取り入れる．クラスタの時間間隔は，クラスタのタイムスタンプの差によって求められる．クラスタのタイムスタンプは，クラスタに属する記事が作成された時間の平均とする．クラスタのタイムスタンプを t_1, t_2 とした場合，重み $weight$ を以下のように設定する．

$$weight = \lambda^{|t_1 - t_2|} \qquad (0 < \lambda < 1)$$

クラスタの時間間隔を考慮した重み $weight$ を利用し，コサイン類似度によってクラスタ間の類似度を計算する．類似度が高いクラスタを同一のトピックに関するクラスタと見なす．同一のトピックに関するクラスタを時間軸にそって出力することでトピックの追跡を行うことが可能である．

4. 実　　験

提案手法の有用性を示すため，様々なトピックが混在した社会に関する 7981 記事を対象に実験を行った．実験環境は，OS が Mac OSX，CPU が 1.6GHz PowerPC G5 およびメモリが 1GB DDR SDRAM である．まず，同じトピックの記事をまとめるためにクラスタリングを行う．次に，クラスタ間の類似度を計算し，関連度が高いクラスタを関連づけてトピックの追跡を行う．そして，クラスタリングの際に抽出されたラベルの候補から重要度が高い語をクラスタのラベルとする．対象とした記事は，2006 年 1 月 10 日から

6月6日までの記事であり，記事はタイトルと本文で構成されている．

4.1 クラスタリング結果

対象データは，社会のジャンルに属するニュースであり，社会での事件・事故に関するニュースが多く見られた．そのうち，一度報道されたニュースに関して，続報があるニュースと続報がないニュースが存在する．続報がないニュースには，話題性や重要性が低いニュース，地方のニュースや一つの記事で完結するような内容のニュースなどが多く見られた．また，続報がないニュースの数が多く，そのほとんどが単一のニュースで単一のクラスタを構成したため，結果的にクラスタの数も多くなった．単一のニュースが単一のクラスタを構成したニュースの例を以下に示す．このようなニュースのほとんどは，他のクラスタとの類似度が低く，関連するトピックがないと判断された．

・客室乗務員はパンツスーツ　スター社、機内も披露（新規航空会社の制服や航空機の内装に関するニュース）
・ラーメン店で偽ブランド　ショーケースに陳列、販売（偽ブランド販売で逮捕された内容と供述内容に関するニュース）
・静岡で文化財防災シンポ　21日午後1時から（静岡で行われた，災害から文化財を守るためのシンポジウムの内容に関するニュース）
・放置自動車はリサイクル　千葉・市原市が初の条例（放置自転車の処分に関する条例の施行のニュース）

また，話題性のあるニュースや，重大なニュース，全国規模のニュースなどは続報が存在することが多い．

センター試験のリスニングテストでのトラブルに関するトピッククラスタが3つ作成された．作成された3つのクラスタ間の類似度が高く，クラスタが関連付けられた．それぞれのクラスタの記事数は，クラスタ1が4件，クラスタ2が3件およびクラスタ3が3件であった．それぞれのクラスタのラベル，記事のタイトルを以下に示す．

＜クラスタ1＞
ラベル：トラブル，大学入試センター試験，リスニング，リスニングテスト，受験生，ICプレーヤー
・初のリスニングを実施　センター試験50万人受験
・リスニングでトラブル　300人超に再テスト
・43都道府県でトラブル　再テスト対象448人
・再テスト対象は461人　リスニングで故障相次ぐ

＜クラスタ2＞
ラベル：大学入試センター試験，イヤホン，リスニング，追試験，実施，小坂大臣
・「受験生におわびする」　リスニングで文科相謝罪
・4%が「よく聞き取れず」　リスニングで予備校調査

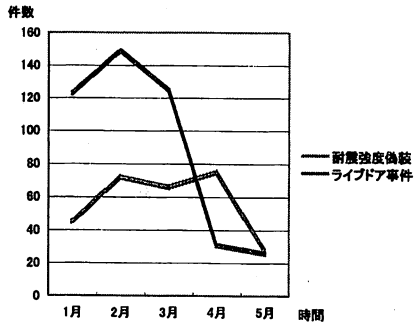

図3　耐震強度偽装およびライブドア事件に関するニュースの件数

・再テスト対象は465人　センター入試リスニング

＜クラスタ3＞
ラベル：得点，大学入試センター試験，受験生，追試験，公民，病気，リスニング
・得点調整はない見込み　センターが平均点中間発表
・得点調整は実施せず　センター試験
・センター試験で再・追試験　リスニング不調の9人も

それぞれのクラスタの内容を観察すると，クラスタ1がトラブルの発生に関する記事の集合，クラスタ2がトラブルの詳細および2次的な事象の発生に関する記事の集合，クラスタ3がトラブルの収束に関する記事の集合であった．このトピックは1ヶ月程度の期間に渡った話題であり，3つのクラスタに分割されたが，1週間程度で収束する話題は，1つのクラスタにまとまる場合が多かった．これは，記事の時間間隔を考慮して，時間間隔が短いほどクラスタにまとまりやすいようにクラスタリングを行っているためである．

対象データの報道された期間を通じで，耐震強度偽装に関する一連のニュースとライブドア事件に関する一連のニュースが目立った．クラスタリングでは，それぞれのニュースに関するクラスタが多く作成され，ライブドア事件に関するニュースが最も主要なトピックとして抽出された．それぞれのニュース件数の変動を図3に示す．以下では，最も主要なトピックと判断されたライブドア事件に対するクラスタリング結果について述べる．

クラスタリング結果からライブドア事件に関するクラスタを抽出したところ，5つのクラスタが抽出された．それぞれのクラスタは類似度が高く，関連づけら

れていた．以下に，クラスタのラベルとクラスタに属しているニュースのタイトルの例を示す．

＜クラスタ１＞
ラベル：ライブドア，幹部，自殺，ホテル，堀江貴文社長，那覇
・証券会社副社長が自殺か　ライブドアの企業買収関与
・ライブドア系列役員も　自殺？証券会社副社長・知人ら参列「残念」　自殺元役員の告別式

＜クラスタ２＞
ラベル：逮捕，還流，新株売却，ライブドアマーケティング，堀江貴文容疑者，ライブドアファイナンス，株式交換，買収
・宮内氏を３日連続聴取　ライブドア事件で特捜部
・堀江社長を逮捕　宮内取締役ら３人も
・「社長に報告、了承得た」　堀江容疑者の関与認める

＜クラスタ３＞
ラベル：粉飾決算，ライブドア，ライブドアマーケティング，関連会社，コンサルタント会社，堀江容疑者，配当，還流，出資，買収
・コンサル会社、深く関与　ライブドア事件で捜索
・別の２投資組合も介在　ライブドア出資隠ぺいか
・株売却益還流は６件　ライブドア

＜クラスタ４＞
ラベル：粉飾決算，再逮捕，起訴，バリュークリックジャパン，ライブドア，発行，新株，証券取引，投資事業組合
・投資ファンドに１６億円　出版社買収でライブドア側
・堀江前社長ら４人起訴　本体粉飾決算で再逮捕へ
・堀江前社長ら４人再逮捕　ライブドア事件で特捜部
・５百億円超社債引き受け　ライブドア、数十億円稼ぐ

＜クラスタ５＞
ラベル：保釈，宮内前取締役，堀江前社長，逮捕，ライブドア，上場廃止，被害拡大，風説
・堀江前社長ら１３日起訴　粉飾決算容疑で再逮捕へ
・堀江被告の保釈認めず　東京地裁が弁護人請求却下
・東京拘置所前に２５０人　堀江被告の保釈待つ報道陣
・堀江前社長保釈へ　東京地裁が準抗告棄却

各クラスタを観察すると，クラスタ１は，事件の関係者が亡くなったことに関するクラスタであり，ラベルにもその事件の内容を表す語「自殺」，「那覇」，「ホテル」が付加された．クラスタ２は，「逮捕」，「還流」，「買収」などのラベルで表されるように，関係者の逮捕や事情聴取に関する記事が集まった．クラスタ３とクラスタ４の内容は類似しており，それぞれ粉飾決算の詳細に関するニュースが集まった．クラスタ５には，関係者の保釈および事件の影響に関する記事が集まった．ラベルの「保釈」や「上場廃止」「被害拡大」などからその内容が予想できる．

以上の結果から，本提案手法を用いて，クラスタを時間軸上で追跡し，ラベルを観察することで，トピックの流れが把握できると考えられる．時間軸にそってクラスタを配置すれば，容易にトピックの流れの全体像を把握することができる．

5. おわりに

本稿では，クラスモデルにおけるクラス分類の考え方をもとに，トピック検出およびトピック追跡の枠組みを提案した．クラスタリングによって，トピックの検出を行い，作成されたクラスタの類似度を計算することでトピックの追跡を行った．クラスモデルに基づいた手法により記事をクラスタリングし，クラスタリングに影響を与える語をクラスタのラベルの候補として抽出した．クラスタは，あるトピックに関する記事の集合に相当し，クラスタリングによって抽出されたラベルはトピックの内容を表現している．トピックの検出とトピックのラベル付けは，クラスタリングによって同時に行える．従来になかった統一的な枠組みを提案した．

文　献

[1] J. Allan, R. Papka, V. Lavrenko, "On-line new event eetection and tracking", Proc. ACM SIGIR, pp.37-45, 1998.

[2] 森永聡，山西健司，“有限混合モデルを用いたトピック傾向の動的マイニング”，情報論的学習理論ワークショップ IBIS2004, 2004.

[3] 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トピックの発見”，DBSJ Letters, Vol. 5, No. 1, pp. 129-132, 2006.

[4] Nadamoto, A, Tanaka, K., “Time-basede contextualized-news browser (t-cnb)”, Proc. 13th international World Wide Web Conference, pp. 458-459, 2004.

[5] J. Allan, V. Khandelwal, R.Gupta, “Temporal summeries of news topics”, Proc. 24th annual international ACM SIGIR conference on Research and development of information retrieval, pp. 10-18, 2001.

[6] Hang Li, Kenji Yamanisi, “Topic analisys using a finite mixture model”, Proc. ACL Workshop on Very Large Corpus, pp. 35-44, 2000.

[7] 森正輝，三浦孝夫，塩谷勇，“時制クラスタのトピック追跡”，データ工学ワークショップ DEWS2006, 2006.

[8] 北研二，“確率的言語モデル”，東京大学出版会，1999.

[9] S. Martin, J. Liermann and H. Ney,“Algorithms for Bigram and Trigram Word Clustering”, Proc. EUROSPEECH-95, Madrid, pp.1253-1256,1995.

[10] MacMahon, J. G. and Smith, F. J., “Improving statical language model performance with automatically generated word hierarchies”, Computational Linguistics, 22(2), pp.217-247, 1996.