# 時系列ニュース記事集合に基づくニュース記事の順序付け

上嶋　宏† 三浦 孝夫† 塩谷 勇††

† 法政大学 工学研究科 電気工学専攻 〒184–8584 東京都小金井市梶野町3-7-2
†† 産能大学 経営情報学部 〒259–1197 神奈川県伊勢原市上粕屋1573
E-mail: †{i03r3207,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

**あらまし** 本稿では,タイムスタンプ（発行時間）を与えられていない文書集合にタイムスタンプを割り当てる（順序付け）手法を提案する. 発行時間が与えられている時系列文書集合を逐次クラスタリングにより学習し, その結果を利用して,発行時間が与えられていない記事の発行時間と話題を予測するものである. 本稿ではTDT2コーパスを用いた実験により，提案手法の有効性を検証する.

**キーワード** TDT,時系列データ,逐次クラスタリング,単一パスクラスタリング

# Temporal Ordering of News Corpus

Hiroshi UEJIMA†, Takao MIURA†, and Isamu SHIOYA††

† Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
†† Department of Management and Information Science, SANNO University 1573, Kamikasuya, Isehara city, Kanagawa 259–1197 Japan
E-mail: †{i03r3207,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

**Abstract** In this investigation, we propose a new mechanism to give *timestamps* to a collection of news corpus without any timestamps. Here we learn temporal data in advance to extract temporal feature by means of incremental clustering and then we estimate most likely timestamps to each news text. In this work, we examine TDT2 corpus and we show how well our approach works by some experiments.

**Key words** TDT, Stream data, Incremental Clustering,Single Pass Clustering

## 1. 前 書 き

インターネットの普及により,ニュースなどの報道記事がオンラインで幅広く利用可能になっている. そこで大量の日々報道されるニュースの動向を容易に素早く把握するための研究が, 近年行われている.

その代表的なものとしてTopic Detection and Tracking (TDT)プロジェクトがある[1]. TDTはオンラインニュースなどの文書データストリームから話題構造を自動で抽出する技術の確立を目指すプロジェクトで, 5つのタスクを設定している.

(1) 話題分割：記事内での話題の変更の検出,
(2) 話題追跡：与えられた話題の記事を検出,
(3) 話題検出：新規話題の発見と,同じ話題について述べている記事によるクラスタ生成,
(4) 事象検出：新規話題の最初の記事を検出,
(5) リンク検出：2つの記事が同一の話題についてのべているか決定

これらのタスクの主に(2),(3),(4)では,データが時間順に並んでいること, あるいはタイムスタンプを持っていることが非常に重要である. 例えば,「湾岸戦争に関する記事」と「イラクへの自衛隊派遣に関する記事」は非常に類似していると考えることができる. しかし，この2つの文章は時間的に大きく離れているということにより，異なった話題であることがわかる.

また，ある話題に関する続報記事を抽出するためには，時間とともに話題が変化していく過程を的確に捉えなくてはならない. 例えば「総理大臣の訪仏」という記事と,「総理大臣の米国での首脳会談」という記事が短い期間内で出現した場合,「総理大臣が仏国から直接渡米」という時間による話題の変化の過程と捉えることにより初めて，これらは「総理大臣の外遊」という同じ話題であると判断できる.

このようにニュースや話題の動向を把握するには記事の時間順序を取得できることが必要である. そのため一般的にこれらのタスクでは，あらかじめ時間順に並んだデータ，あるいはタイムスタンプを持っているデータを想定している. 逆に言うとタイムスタンプを持たないデータは，このようなタスクに貢献できないことを意味する. そのために，タイムスタンプを持たないデータにタイムスタンプを割り当てることは, データ空

間がより密になり，ニュースの話題や動向の変化をよりスムーズに把握することができ,話題追跡や話題検出に非常に有用である.

そこで，本稿ではタイムスタンプを持たない文書集合$N$に逐次，タイムスタンプを割り当てる手法を提案する．ここで，本稿ではタイムスタンプを持つ時系列文書集合$T$から$N$内の文書$n$のタイムスタンプを求める．基本的な考えは，$n$に最も近い，文書集合$T$内の文書$t_1$のタイムスタンプを$n$のタイムスタンプとするものである. これにより,タイムスタンプを持たない文書にタイムスタンプを割り当てる.

しかし，$n$のタイムスタンプを，単に$n$に一番近い記事のタイムスタンプを割り当てる事により決定した場合，正しく内容を考慮してタイムスタンプを与えたことにはならない．そのため，本稿では時系列文書集合$T$を，話題を細分化した事象により逐次クラスタリングし（事象検出），$N$を逐次，生成されたクラスタに割り当てる．これにより，$n$の示す事象を決定し，$n$の属するクラスタの文書集合から$n$に逐次，タイムスタンプを与える．ストリームデータでは，リアルタイムに最適な結果を取得可能であることが重要であり，逐次手法により処理を行うことが必要である.

通常，文書は，その内容に関する時間（有効時間）に従って理解されるが、必ずしも文書の内容時間が文書の作成された時間（トランザクション時間）と一致するものではない．文書の内容と生成時間に大きな差のある文書(過去の事件を振り返った記事や異なったニュースソースからの記事)など，タイムスタンプを持たない集合とタイムスタンプをもつ集合の時系列が異なる場合, 生成時間ではなく，記事の内容時間によりタイムスタンプを与えることができる.

2章では関連研究について述べ，3章で，文書ストリームデータの逐次クラスタリングと，タイムスタンプを持たない文書への事象の割り当て方法について示し，4章で，その結果に基づいた，タイムスタンプを持たない文書に対するタイムスタンプ割り当て手法を提案する，5章でTDT2コーパスを用いた提案手法の実験と結果を示し, 6章で結びとする.

## 2. 関 連 研 究

文書に時間を割り当てる研究としては，Maniらの研究がある [5]．Maniらは,文書内の時制表現を抽出する手法を提案している. これらの時制表現を抽出する研究も盛んに行われている. しかし, これらの手法は文書の発行時間から，文書内の時制表現が示す相対的な時間を抽出し, 文書内の文の時間を求めるが, 他の文書との関係や, 文書自体の発行時間は考えない.

また，Papkaらは Inqueryによる単一パス法に基づいた話題検出法を提案している [3]．Papkaらは話題検出手法の提案を目的としており，本稿の目的とは大きく異なる．しきい値を時間距離により設定し逐次クラスタリングする点は，本稿で提案するクラスタリング手法と類似しているが，文書と質問の類似度の求め方やしきい値の設定方法は異なる．本稿ではIDF値は使用しないが，Papkaらは，補助コーパスから求めたIDF値を使用している.

石川らは，忘却の概念を利用した，逐次的な文書クラスタリング手法を提案している [6]．$C^2ICM$という文書クラスタリングアルゴリズムにより逐次クラスタリングしている．ここでは，統計的な確率に基づき，文書間の類似度を求めている．また，少ない計算量での差分によるクラスタの更新を行っている.

これらのように，時間を考慮した時系列文書データのクラスタリング手法は多く提案されている．しかし，これらの手法のほとんどは，すべての文書がタイムスタンプを持っていることを想定している．そのため，時系列文書とタイムスタンプを持たない文書を同時に扱う手法はあまり提案されていない.

本稿では時間距離を考慮するために，石川らの提案する忘却関数を利用して，クラスタリング，タイムスタンプ割り当てを行う [6].

## 3. 逐次クラスタリングとクラスタ割り当て

時系列文書集合を逐次クラスタリングする手法について述べる．TDTでは，話題発見や追跡において，時間を考慮したクラスタリングが非常に高い性能を示す [4]．すなわち，ニュースの話題や事象の出現は時間に大きく依存することを意味する.

最初に，本稿では逐次的な手法により，タイムスタンプを持った時系列文書集合を事象によりクラスタリングする．ここでのクラスタはタイムスタンプを考慮して作成され, 各クラスタはそれぞれニュース的意味を持つ.

また逐次クラスタリングすることにより, タイムスタンプを持たない文書が来るたびに, タイムスタンプを推定することができ,話題追跡等へ利用することが可能になる. プロセスは逐次的なので、アルゴリズムは現在のクラスタを維持し、追加されたデータだけを考慮する．これにより，その都度，最適な結果を得ることが可能で，効果的なクラスタリングを行うことができる.

### 3.1 文 書 表 現

文書とクラスタの表現は，テキストマイニングで一般的に使用される従来のベクトル空間モデルを利用する. 文書とクラスタは単語のベクトルで表現し, その次元は単語である．本稿では単語をステミングし,BrillTaggerにより, 名詞と固有名詞のみを抽出して利用する [2].

文書ベクトル$\vec{X}$の各属性の値は記事内での単語の重みで, 文書$\vec{X}$での$j$番目の単語$t_j$の値は, 単語$t_j$の文書$X$における出現頻度$TF(j)$（term frequency;TF)で表す．文書ベクトル$\vec{X}$は以下のように示される

$$\vec{X} = \frac{(TF(1), ....., TF(n))}{\sqrt{TF(1)^2 + .... + TF(n)^2}}$$

ここでベクトルは$\sum_{j=1}^{n} TF(j) = 1$のように正規化される.

クラスタはクラスタ内の文書ベクトルの重心で表現する.

ストリームデータのように逐次データが増える場合, IDF値(Inverce Document Frequency)を利用するためには，新しい文書が追加されるたびに再計算が必要である．Yangら [4]や，石川ら [6]は，話題検出タスク内での逐次クラスタリングで逐次IDFを更新する方法を提案しているが, 逐次クラスタリング

を行う際にIDFを更新することは, 一度決定した過去のクラスタリング基準が変わってしまう場合があり, クラスタの割り当てやタイムスタンプ割り当ての結果が時間により変わる可能性がある．このため，本稿では，IDF値を利用しない逐次クラスタリングを行う．

### 3.2 単一パスクラスタリング

本稿では, タイムスタンプを持つ時系列文書集合を単一パスクラスタリング法により逐次クラスタリングする [3], [4]．単一パスクラスタリング法は逐次クラスタリングに適している，また，非常に単純な手法である．文書$X_i$とクラスタ$C_j$の類似度がしきい値$th$を超えたら文書$X_i$をクラスタ$C_j$に追加し, どのクラスタに対してもしきい値を超えない場合は，文書$X_i$を新しいクラスタとする方法である．ストリームデータによる逐次クラスタリングなので, データが追加されるたびにその差分だけを計算する．すなわち, 一度決定されたクラスタ結果は永久に変更されることはない．

この単一パスクラスタリングは, 以下の手順で実行される．

（1） しきい値$th$を設定

（2） 最初は空の集合$S$から始め,1つ目のデータ$X_1$をクラスタ$C_1$とする

（3） 次の文書$X_i$ $(i > 1)$を取得し，既存の全クラスタ$C$との類似度$sim(\vec{X_i}, \vec{C})$を計算する

（4） 最も高い類似度$sc$を持つクラスタ$C_{X_i}$を最も類似したクラスタとし，もし$sc > th$なら，$X_i$をクラスタ$C_x$のメンバとし, そのクラスタの重心を更新する．

もし$sc < h$なら,$X_i$を新しいクラスタ$C_{x_i}$とする

$$sc = \texttt{MAX}\{sim(\vec{X_i}, \vec{C})\}$$

（5） 手順(3)，(4)をデータがなくなるまで繰り返す

文書$X$とクラスタ$C$の類似度はコサイン尺度と呼ばれる方法で求め，以下のように示される．クラスタ$C$の重心を$V_C$とあらわす．

$$sim(\vec{X}, \vec{C}) = \frac{\vec{X} \cdot \vec{V_C}}{|\vec{X}||\vec{V_C}|} \tag{1}$$

### 3.3 忘 却 関 数

本稿では, 文書とクラスタ間の時間距離を考慮した類似度計算を行うために，忘却関数$w_\lambda(t)$を適用する [6], [9]．ここで$t$は時間距離を示し，その単位は日数である．また$0 < \lambda < 1.0$で$w_\lambda(t) = \lambda^t$とする．

ある期間に集中して発生するニュース記事では, 2つ文書が文書ベクトル的に類似度が高くても, タイムスタンプが離れていると，同じ事象について述べている可能性は低くなる．逆にタイムスタンプの非常に近い2つの文書がある場合、文書ベクトルの類似度は多少低くても同じ事象について述べている可能性は高い．

忘却関数を用いることにより，これらの問題を解決することができる．忘却関数とは, 文書とクラスタのタイムスタンプが離れているほど類似度を小さくするものである．図1は，$\lambda = 0.97$の時の，現在$(w_\lambda(0))$と30日後$(w_\lambda(30))$のクラスタの状態を示している．忘却関数により類似度が小さくなることはクラスタが小さくなることに対応する．

また, 本稿ではクラスタ$C$のタイムスタンプ$time_C$はクラスタ内の一番新しい文書のタイムスタンプとする．現在を$time_{now}$とすると, 現在到着した文書$X$とクラスタ$C$の忘却関数を考慮した類似度$sim'$は，以下のように与えられる．

$$sim'(\vec{X}, \vec{C}) = w_\lambda(|time_{now} - time_C|) \times sim(\vec{X}, \vec{C})$$

この類似度を用いて, 上記の単一パスクラスタリングを行う．文書$X$とクラスタ$C$の類似度$sc'$は以下のようになる．

$$sc' = \texttt{MAX}\{sim'(\vec{X}, \vec{C})\} \tag{2}$$

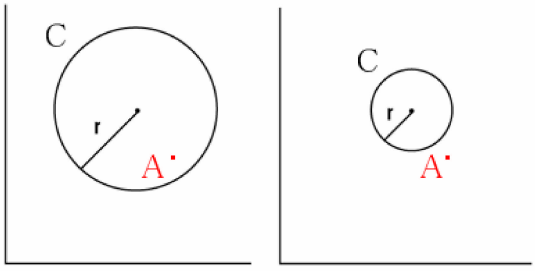


図1 $w_\lambda(0) = 1.0$ $w_\lambda(30) = 0.5454$

### 3.4 上位10点法と最近点法によるクラスタ割り当て

上記クラスタリング結果に基づき，タイムスタンプを持たないデータ集合$N$内の文書$n$が示す事象（クラスタ）を決定する．

文書$n$のクラスタの決定は，文書集合$T$内で，$n$との類似度の高い上位10個の文書による投票により決定する上位$10$点法により行う．つまり, この10個の文書内で一番多くを占めているクラスタを$n$の属するクラスタと決定する．図2の場合、$n$が属するクラスタは$C_1$クラスタとなる．

また，本稿では，上位10点法の有効性を評価するために，最近点法と比較する．最近点法とは，$n$に一番近い点が属するクラスタを$n$が属するクラスタとする非常に単純な方法である．図2の場合，最近点法による$n$の属するクラスタは$C_3$クラスタとなる．

この際の類似度計算はコサイン尺度により求め，式(1)より以下で示される．

$$sim(\vec{n}, \vec{T_i}) = \frac{\vec{n} \cdot \vec{T_i}}{|\vec{n}| \times |\vec{T_i}|}$$

### 3.5 バッチ手法

本稿では，一般的なバッチ手法によるクラスタリングと，その結果に基づくクラスタ割り当ても行う．このバッチ手法と逐次手法の結果を比較することにより，逐次手法を評価する．

バッチ手法として，本稿ではk-means法を用い，逐次手法による単一パスクラスタリング結果と比較する．ここで使用する

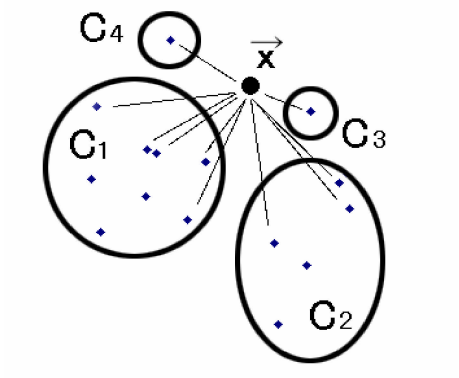


図2　文書ベクトル$\vec{n}$の投票によるクラスタ割り当て

k-means法では，単一パス法と同様，コサイン類似度と忘却関数を用い文書間距離を計算する．また，忘却関数で使用するパラメタ$\lambda$は単一パス法と等しい値とする．k-means法の結果に基づいた文書集合$N$のクラスタ割り当ても逐次手法と同様，最近点法と上位10点法により行う．

## 4. タイムスタンプ割り当て

本稿では，上記のクラスタリングとクラスタ割り当ての結果に基づき，文書$n$の示す事象を考慮した，逐次タイムスタンプを割り当てる手法を提案する．また，本稿ではニュース記事の内容時間と生成時間は一致すると考える．

### 4.1 上位10点法と最近点法によるタイムスタンプ割り当て

文書$n$の示す事象を考慮したタイムスタンプ割り当てを行うために，前章でのクラスタ割り当て結果に基づいたタイムスタンプ割り当てを行う．そのため，上位10点法と最近点法のそれぞれの基づき，タイムスタンプを割り当てる．本稿では，上位10点法と最近点法の割り当て方法はそれぞれ異なり，2種類の割り当て結果を比較する．

最近点法によるタイムスタンプの割り当ては，前章と同じく，$n$に最も近い文書を$n$のタイムスタンプとする非常に単純な手法である．図2の場合、最も近い，クラスタ$C_3$内の文書のタイムスタンプを$n$のタイムスタンプとする．

次に上位10点法によるタイムスタンプ割り当て方法について述べる．上位10点法では，文書$n$に,最も近い10個のデータのうち, データ$n$が属するとしたクラスタ$C$に属するデータ集合$T_C$のみを使用し，タイムスタンプを割り当てる．図2の場合，文書$n$に近い上位10個の内，クラスタ$C_1$に属する5つの文書から文書$n$のタイムスタンプを予想する．

データ集合$T_C$内の1つのデータ$t_C$による$n$のタイムスタンプの予測を以下の式により与える．

$$TimeStamp_{n,t_C,\lambda}(day) = sim(t_C,n) \times distr_C(time_{t_C}) \times w_\lambda(|time_{t_C} - day|)$$

この式は以下の図3のようなタイムスタンプ予想曲線を与える．ここで$distr_C(time_{t_C})$は，クラスタ$C$内のタイムスタンプ分布（図4）の時間$time_{t_C}$での値を示す．ニュース記事では，大抵のニュースはある期間に集中して起こる特性があり，タイムスタンプ分布を考慮することは重要である [1]．クラスタ内のタイムスタンプの分布は$n$が持つタイムスタンプの発生確率と考えることができる．そのために,本手法ではタイムスタンプ分布$distr_C(time_{t_{ci}})$と，$n$と$t_{ci}$の類似度の積を頂点としたタイムスタンプ予想曲線を$T_C$内の各データ毎に求め総和を取る。すなわち,以下の式になる．

$$\begin{aligned} TimeStamp_{n,T_C,\lambda}(day) = & \\ & TimeStamp_{n,t_{c_1},\lambda_1}(day) \\ & + TimeStamp_{n,t_{c_2},\lambda_2}(day) \\ & \vdots \\ & + TimeStamp_{n,t_{c_k},\lambda_k}(day) \end{aligned} \quad (3)$$

図2の場合，$C_1$内の5つの文書から，図5のようなタイムスタンプ予想曲線を得る．

ここで，タイムスタンプの誤差許容範囲を設定する．これは，タイムスタンプ割り当ての精度を評価するときの，実際のタイムスタンプと予測したタイムスタンプの差の許容範囲であり，この許容範囲に従って予想曲線からタイムスタンプを求める．タイムスタンプの割り当ては，予想曲線の誤差許容範囲内の総和が最大になるタイムスタンプを文書$n$に割り当てるものである．例えば許容誤差範囲が$m$日の場合，前後$m$日間のタイムスタンプ予想曲線の総和が最大になる日付を文書$n$のタイムスタンプとする [図5]．$n$のタイムスタンプ$day_n$は以下の式で表すことができる．

$$day_n = \mathtt{MaxArg}_{day} \int_{day-m}^{day+m} TimeStamp_{n,T_C,\lambda}(day)$$

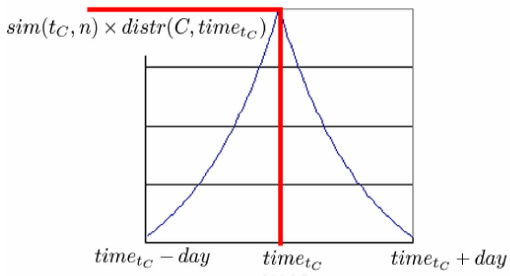


図3　1つの文書によるタイムスタンプ予想曲線 ($\lambda = 0.97$)

### 4.2 逐次手法とバッチ手法

本稿では，逐次的なタイムスタンプ割り当て手法を提案しているが，2章のクラスタリング同様，逐次手法を評価するために，逐次手法とバッチ手法のそれぞれのクラスタリング結果に基づいた上位10点法によるタイムスタンプ割り当てを比較する．同様に最近点法によるタイムスタンプ割り当ても比較する．

## 5. 実　　験

### 5.1 TDT2コーパス

本稿では,実験にTDT2コーパスを用いる [7]．TDT2コーパ

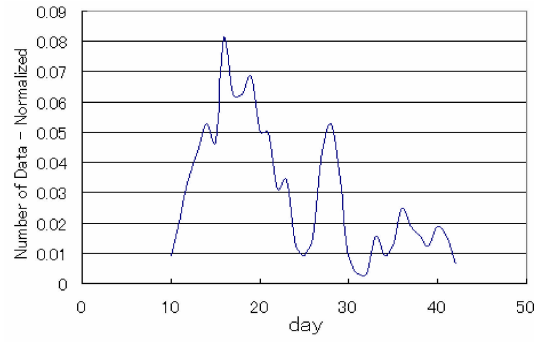


図4　クラスタのタイムスタンプ分布

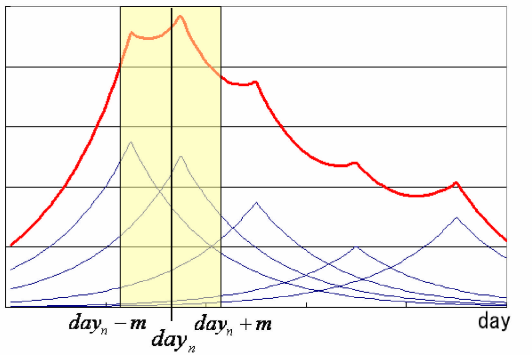


図5　5つの文書によるタイムスタンプ予想曲線

スは，放送されたニュースを書き写したものと，ニュース通信の2種類からなり，それぞれ1998年1月から6月までの6ヶ月間分のデータからなる．また，TDTコーパスには英語と中国語のデータが含まれるが，今回は英語の記事のみを利用する．放送データとして（ABC，CNN，VOA，PRI），ニュース通信として（APW，NYT）の計6つのニュースソースを利用する．本稿では，放送データ，ニュース通信データ共に速報性が高く，内容時間と発行時間は一致すると考えるが，その中でも速報性が高い放送データをタイムスタンプを持ったデータ（訓練データ）とし，ニュース通信のデータを，タイムスタンプを予測するデータ(テストデータ)とする．そして,実際のタイムスタンプと予測結果を比較することにより，提案する手法を評価する．

また,本稿では,放送記事,ニュース通信記事の双方に200件以上の"YES"タグが付与された記事を持つ4つの話題（$T_1 - T_4$）を利用して実験を行った。実験に使用した話題とそのデータ数を以下に示す．

| | Topic | Broadcast | Newswire |
|---|---|---|---|
| $T_1$: | Asian Economic Crisis | 476 | 657 |
| $T_2$: | Monica Lewinsky Case | 747 | 222 |
| $T_3$: | 1998 Winter Olympics | 222 | 318 |
| $T_4$: | Current Conflict with Iraq | 1022 | 464 |
| | Total | 2467 | 1661 |

話題ごとのタイムスタンプの分布を以下の図6に示す．

### 5.2　実 験 手 順

本稿では発行時間によるタイムスタンプ割り当てを行う．そ

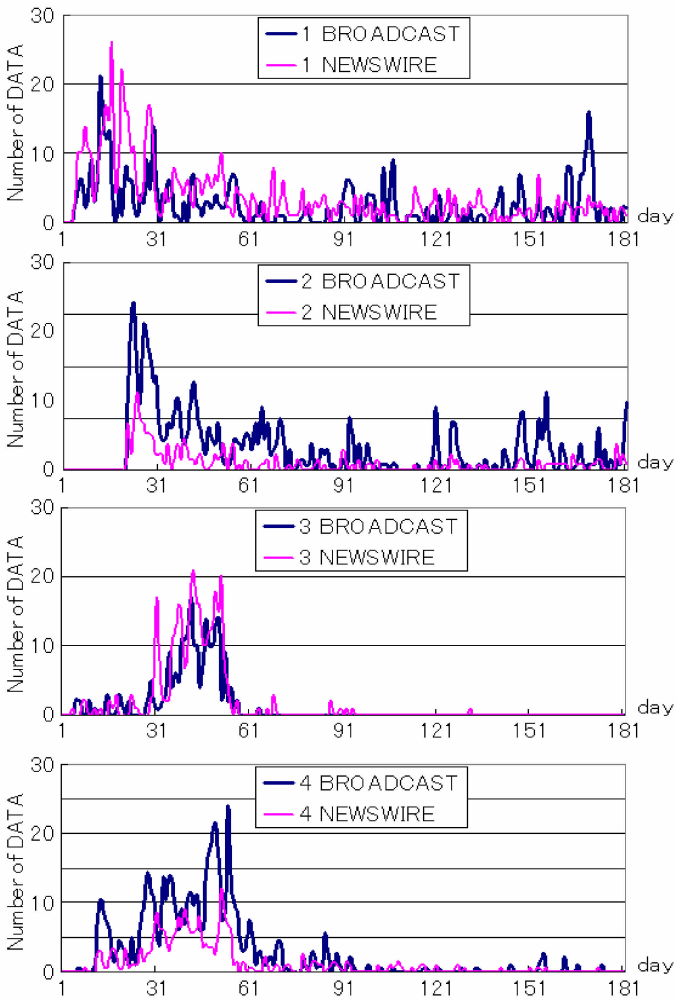


図6　話題のタイムスタンプ分布

して，逐次方式とバッチ方式の２つのタイプにより，それぞれ最近点法と上位10点法により，クラスタへの割り当てとタイムスタンプ割り当ての比較を行う．

逐次処理により1ヶ月分クラスタリングするごとに，それ以前のすべてのニュース通信記事のタイムスタンプ割り当てを試みた．バッチ処理は，その月迄のデータを使用したクラスタリング結果により，それ以前のニュース通信記事のタイムスタンプ割り当てを行った．すなわち,最初1月の終わりのデータまでクラスタした時点で，1月のニュース通信記事のタイムスタンプ割り当てを行う．次に，2月の終わりまでクラスタリングした時点で,1月,2月両方のニュース通信記事のタイムスタンプ割り当てを行う．これを6月まで繰り返す．

まず最初に，クラスタリング精度により，逐次単一パスクラスタリングによるクラスタリング（事象検出）の評価を行う．ここで,「逐次単一パス法による結果」を,「バッチk-means法による結果」と比較することにより，逐次方式による性能の有効性を評価する．さらに,「TDTコーパスに与えられえている話題をクラスタとしたもの」と比較することにより，逐次クラスタリングにより生成された事象（クラスタ）が，正しく，話題を細分化したものになっているかを評価する．

単一パス法によるクラスタを$C_1 - C_x$とおき，k-means法に

よるクラスタを$K_1 - K_y$とすると，単一パス法とk-means法との比較方法は，まず，クラスタ$K_i$が$C_1 - C_x$のどのクラスタを意味するかを決定する．決定方法としては，$K_i$内の文書集合の，属する割合がもっとも多いクラスタ$C_j$を$K_i$が示すクラスタとする．図7の場合，$K_1, K_2$は$C_1$を意味する．次に，逐次，バッチのそれぞれのクラスタリング結果に基づき，文書$n$が属するクラスタ$C_j$と$K_i$を比較して，$K_i$の意味するクラスタが$C_j$ならば正解とする．

$\alpha$を全記事数とし，$\beta$を，文書$X$の属するクラスタ$K_i$の意味するクラスタが，逐次手法による文書$X$の属する$C_j$と一致する記事数とすると，クラスタリング精度は$\beta/\alpha$と定義される．ここで$i > j$で，本実験ではk-means法のkの値は単一パス法により生成されたクラスタ数の2倍とした．

図7の場合，$K_1, K_2$は$C_1$を意味し，$K_2$のみでのクラスタリング精度は14/20となる．

単一パス法と「TDTコーパスに与えられている話題」との比較方法としては，k-meansとの比較同様，クラスタ$C_j$内の文書が示す話題のうち，割合がもっとも多い話題をそのクラスタの話題とし，文書自身の話題とクラスタの話題が等しいか否か比較する．この場合$\beta$を，記事の話題と，その記事が属するクラスタの話題が一致する記事数とする．

クラスタリング評価は一ヶ月毎にその時点でのクラスタリング結果を評価する.[表1]

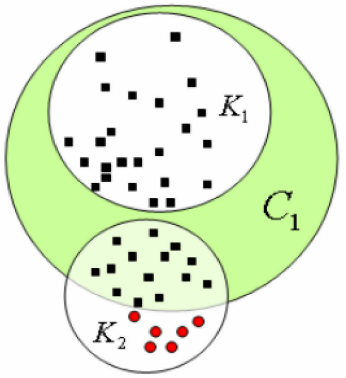


図7　クラスタリング精度の例

次に,テストデータのクラスタ割り当ての評価を行う．上記の,「逐次的クラスタリングの結果」に基づいたテストコーパスの割り当て結果と,「バッチk-meansによるクラスタリング結果」,「TDTコーパスに与えられている話題」の2つに基づいた割り当て結果との比較を最近点法，上位10点法のそれぞれに対して行う．評価方法としては，k-means法との比較の場合，k-means法に基づき，文書$n$が割り当てられたクラスタ$K_i$の意味する事象が，逐次法に基づき，文書$n$が割り当てられたクラスタ$C_j$と等しければ$(K_i = C_j)$正解とする．

この割り当ては割り当て精度により評価し，$\alpha$を全記事数，$\gamma$を正しく割り当てられた$(K_i = C_j)$記事数とすると，割り当て精度は$\gamma/\alpha$と定義される．

話題との比較する場合，文書$n$に割り当てた事象[図2]の意味する話題が，文書$n$に与えれている話題（答え）と一致するかどうかを評価する（話題追跡）．[表2] この場合$\gamma$を正しい話題が割り当てられた記事数とする．

最後に，タイムスタンプ精度 によりタイムスタンプ割り当ての評価を行う．逐次手法とバッチ手法それぞれのクラスタリング結果に基づいた，上位10点法による結果と最近点法による結果の計3つを比較する．さらに,3章で示した許容誤差範囲を指定する．許容誤差範囲は1週間（7日）と1ヶ月(30日）とする．つまり，予想されたタイムスタンプと実際のタイムスタンプの差が1週間以内なら正解[表3図9], 1ヶ月以内なら正解[表4図10]とする計2パターンの許容誤差範囲を比較する．

$\alpha$を全記事数とし，$\delta$を正しくタイムスタンプが与えられた記事の数とすると，タイムスタンプ精度 は$\delta/\alpha$と定義される．[表3,4]

1月毎にタイムスタンプ割り当ての性能を評価するが，日数が増えるごとに，タイムスタンプを割り当てる範囲が広くなるので，タイムスタンプ精度は低下すると考えられる．

### 5.3　実 験 結 果

クラスタリング精度を表1に示す。

| | Jan. | Feb. | Mar. | Apr. | May | June |
|---|---|---|---|---|---|---|
| K-means(%) | 86.16 | 79.81 | 77.72 | 81.98 | 79.26 | 88.68 |
| Topic(%) | 96.89 | 96.11 | 96.38 | 96.55 | 96.7 | 97.04 |

表1　クラスタリング精度

ここでは，単一パスクラスタリングにより生成されたデータ集合のうち，データを5つ以上含む集合をクラスタとみなした．また、6か月分クラスタリングした後の最終的なクラスタ数は21クラスタであった．すなわちK-means法では最終的には$k = 42$としてクラスタリングを行った．

次に逐次，バッチの両手法での割り当て精度の結果を示す.[表2,図8]

| | | Jan. | Feb. | Mar. | Apr. | May | June |
|---|---|---|---|---|---|---|---|
| K-means | TOP10 | 83.72 | 85.78 | 75.95 | 81.29 | 81.93 | 86.03 |
| K-means | NN | 80.23 | 83.74 | 71.57 | 75.83 | 74.62 | 74.59 |
| Topic | TOP10 | 93.80 | 96.62 | 96.53 | 96.67 | 97.07 | 96.81 |
| Topic | NN | 95.16 | 96.31 | 95.90 | 96.40 | 96.25 | 96.69 |

表2　割り当て精度

続いて，これらの結果に基づいたタイムスタンプ精度を，誤差を1週間とした場合[表3,図9]と1ヶ月とした場合[表4,図10]のタイムスタンプ精度をそれぞれ以下に示す.

| | | Jan. | Feb. | Mar. | Apr. | May | June |
|---|---|---|---|---|---|---|---|
| | N N | 70.93 | 57.11 | 51.91 | 48.27 | 44.91 | 43.59 |
| K-means | TOP10 | 69.19 | 59.15 | 53.96 | 50.93 | 49.55 | 46.6 |
| On-line | TOP10 | 73.06 | 60.88 | 55.30 | 53.00 | 49.81 | 47.38 |

表3　タイムスタンプ精度（誤差許容範囲7日）

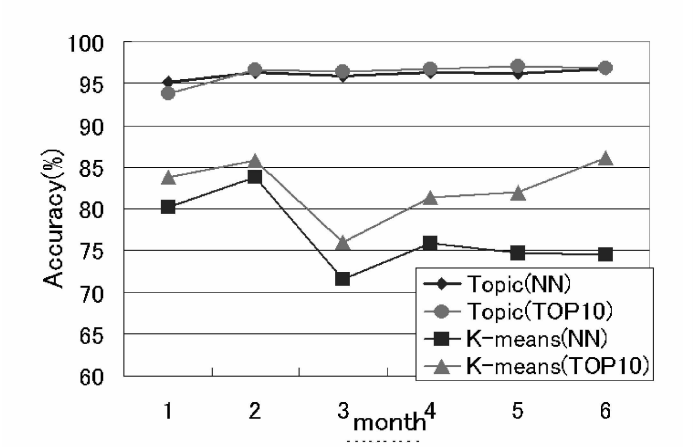


図8　割り当て精度

| | | Jan. | Feb. | Mar. | Apr. | May | June |
|---|---|---|---|---|---|---|---|
| | NN | 100 | 93.72 | 87.27 | 80.83 | 75.13 | 72.55 |
| K-means | TOP10 | 100 | 96.78 | 90.95 | 83.69 | 81.36 | 77.24 |
| Online | TOP10 | 100 | 98.59 | 91.09 | 85.62 | 82.12 | 78.33 |

表4　タイムスタンプ精度（誤差許容範囲30日）

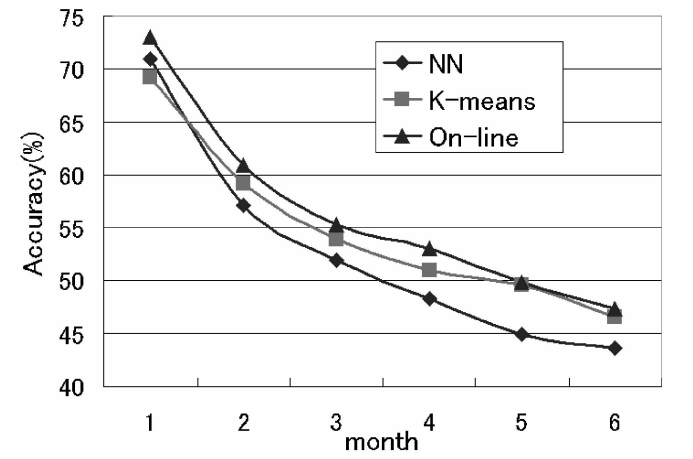


図9　タイムスタンプ精度（誤差許容範囲7日）

### 5.4 考　　察

表1により，バッチ手法による結果と比較しても，80%以上一致し，非常に高い精度でのクラスタリングが可能なことがわかる．また，話題と比較した場合も，常に96%以上の正解率を得ていることから，話題を細分化した，事象によるクラスタリングができていることがわかる．

表2より，逐次手法により，テストデータ$n$に割り当てた事象の正解率も，バッチ手法の場合と80%程度一致しており，高い精度で，事象の割り当てが可能になっていることがわかる．表1により，事象（クラスタ）が正しく話題を細分化しているため，話題割り当ての正解率も，95%以上を示している．

これらより,逐次,バッチどちらの場合でもほぼ同性能のタイムスタンプ割り当て結果を得た.[図9,図10]

タイムスタンプ精度は誤差範囲が7日で, 6か月分のデータに

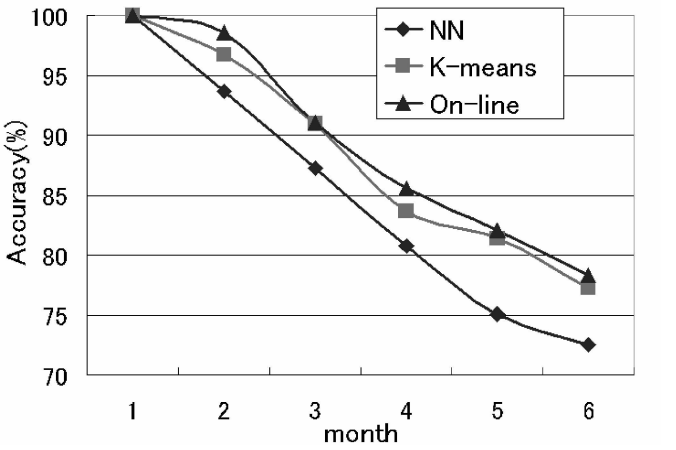


図10　タイムスタンプ精度（誤差許容範囲30日）

タイムスタンプを割り当てた場合, 精度が約50%と高い性能で，タイムスタンプを割り当てることが可能となった．また，誤差範囲が1ヶ月の場合，最低の場合でも70%以上の精度でタイムスタンプを割り当てることができた．これによりタイムスタンプが与えられていないデータにもタイムスタンプを与えることが可能である．上位10点法と最近点法によるタイムスタンプ割り当てを比較した結果, 上位10点法による結果の方が一般に優れた性能をもち, この手法を取ることは妥当であるといえる．上位10点法の場合，事象に基づいてタイムスタンプを割り当てているために，事象追跡の性能にタイムスタンプ精度が大きく依存すると考えられる．今回は4つの話題に限定して実験を行ったが，話題数が非常に多い場合も，事象追跡の性能が高ければ，高い精度でのタイムスタンプ割り当てが可能であると考えられる．事象は時間に深く依存しているために，最近点法に比べ上位10点法は割り当て期間の増加による影響が少ない．そのため，図9，図10を見ると，割り当て期間が大きくなるにつれ，最近点法と上位10点法の精度の差が大きくなっていることがわかる．これにより事象による割り当てが，期間の増大にも対応できることがわかる．

ここで表5に，6ヶ月分のデータを逐次クラスタリングした後の，話題毎のクラスタ割り当て精度とタイムスタンプ割り当て精度を示す．これより，タイムスタンプ割り当て精度が，クラスタ割り当て精度に依存していることがわかる．$T_1$のようにピークのあまりない話題[図6]に関しては，タイムスタンプ割り当て精度が低い結果となった．時間を考慮し，クラスタリングしているために，ピークを持たない話題は，事象により細分化することが困難であると考えられる．

しかしながら，ニュース記事の特性として，短期間に集中して現れるという特性があり，$T_1$のような，ピークの分散している話題は少ない．

次に，話題の割り当てとタイムスタンプ割り当ての関係について述べる．表6，7は逐次的，上位10位法による6ヶ月すべての文書にタイムスタンプ割り当てをした場合の，話題割り当ての正否と，タイムスタンプ割り当ての正否の文書数の関係を示

| 話題 | 割り当て | 誤差7日 | 誤差30日 |
|---|---|---|---|
| $T_1$ | 94.67 | 24.66 | 63.77 |
| $T_2$ | 97.3 | 48.65 | 74.32 |
| $T_3$ | 98.74 | 72.96 | 96.23 |
| $T_4$ | 98.28 | 61.42 | 88.58 |

表5　割り当て精度とタイムスタンプ精度

している.

| | 時間正解 | 時間間違い |
|---|---|---|
| 話題正解 | 778 | 830 |
| 話題間違い | 9 | 44 |

表6　時間の間違い比率（誤差範囲7日）

| | 時間正解 | 時間間違い |
|---|---|---|
| 話題正解 | 1262 | 346 |
| 話題間違い | 39 | 14 |

表7　時間の間違い比率（誤差範囲30日）

これらから，誤差範囲7日の場合，正しく話題が割り当てられた1608件の文書中，間違ってタイムスタンプが割り当てられたものは，48.38%の778件に対し，間違った話題が割り当てられた53件中，83.02%の44件の文書は間違ったタイムスタンプが割り当てられた．誤差範囲30日の場合でも，正しく話題が割り当てられた文書で，間違った時間が割り当てられたものより，話題，時間の両方が間違って割り当てられたものの割合が高くなっている．この結果から，事象を考慮したタイムスタンプ割り当てが有効に働いていることがわかる．表5より，タイムスタンプ割り当てが，クラスタ割り当て精度に依存していることがわかったが，話題割り当てが間違った文書はタイムスタンプ割り当ても間違っている傾向が非常に強いため，これらの文書は他の文書とは異なった例外的な文書である可能性が高い.

本実験では,しきい値$th = 0.24$, 忘却関数$\lambda = 0.97$の時にもっとも良い結果を示した．しかし今回の実験では,クラスタリングに関して，忘却関数は結果にはそれほど大きく影響しなかった．TDT2コーパスは6ヶ月間という期間のデータで，話題の時間的大きさに対して，データの期間が短いために，あまりうまく働かなかったようである．今後より長い期間のデータによる評価も必要と考えられる.

## 6. 結　　び

本稿では,ニュースストリームを実時間でクラスタリングしつつ、タイムスタンプを持たないデータにタイムスタンプを割り当てる手法を提案した．TDTコーパスによる実験の結果1週間という狭い誤差にもかかわらず,50%程度の精度でタイムスタンプを割り当てることが可能であった。

また,今後の展開としては，忘却関数でのパラメタ$\lambda$や，しきい値$th$などのパラメタ設定手法の提案や，や，タイムスタンプ割り当て手法を実際の話題追跡や話題発見に応用していく予定である．また，ニュースストリームの場合,記事の生成時間と内容時間はほぼ一致するが,必ずしも内容時間順に並んでいるわけではない.記事の内容時間よる並べ変えの処理を行うことにより今までに抽出できなかった情報を抽出することを考えている.

謝辞

本研究に対して貴重な支援をいただいた Wai Lam 教授 (香港中文大学) に感謝します。本研究の一部は文部科学省科学研究費補助金 (課題番号14580392) の支援による

## 文　　献

[1] Allan, J., Carbonell,J., Doddington, G., Yamron,J. and Yang,Y.: Topic Detection and Tracking Pilot Study: Final Report, Proc. DARPA Broadcast News Transcription and Understanding Workshop (1998).

[2] Grossman,D. and Frieder,O.: Information Retrieval – Algorithms and Heuristics, Kluwer Academic Press, 1998

[3] Papka, R. and Allan, J.: On-line new event detection using single-pass clustering, Technical Report UMASS Computer Science Technical Report 98 - 21, Department of Computer Science, University of Massachusetts, 1998.

[4] Yang, Y., Pierce, T. and Carbonell,J.: A Study on Retrospective and On-Line Event Detection Proc. of SIGIR-98, ACM Intn'l Conf. on Research and Development in Information Retrieval,1998

[5] Mani,I. and Wilson,G.: Robust temporal processing of news. Proc.of Annual Meeting of the Association for Computational Linguistics (ACL 2000), New Brunswick, New Jersey, 2000. Association for Computational Linguistics.

[6] Ishikawa,Y. ,Chen,Y. and Kitagawa, H.: An On-line Document Clustering Method Based on Forgetting Factors, in Proc. of 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'01),2001

[7] Wayne,C., Doddington,G. et al.: TDT2 Multilanguage Text Version 4.0 LDC2001T57, Philadelphia: Linguistic Data Consortium (LDC), 2001

[8] 福本 文代, 鈴木 良弥, 山田 寛康: 話題の推移に基づく続報記事の自動抽出, 情報処理学会ジャーナル Vol.44 No.07,2003

[9] 石川 佳治, 北川 博之: 忘却の概念に基づくクラスタリングの改良手法, 日本データベース学会 Letters Vol.2, No.3, 2003

[10] 上嶋 宏，三浦 孝夫，塩谷 勇: 同義語，多義語の考慮による文書分類の精度向上, 電子情報通信学会論文誌 Vol.J87-D-I No.2, 2004