マルチチャンネル型ニュース配信システムのための時系列クラスタリング

角谷和俊[†] 松本好市[††]
高橋美乃梨[†††] 上原邦昭[†††]

近年，インターネットを用いたプッシュ型ニュース配信システムが注目を浴びている．プッシュ型ニュース配信システムでは，配信される記事は追加・更新される時系列データである．ニュース記事の情報価値は時間に依存しているので，ユーザにとって価値の高い情報を選択するには，従来のプル型配信では毎回情報を取得する必要がある．本論文では，オンラインニュース配信における時系列文書のための時系列クラスタリングとマルチチャンネル型放送配信システムについて述べる．まず，オンラインニュースの時間情報と内容情報から，同一のトピックを持ったニュース記事を配信順に並べたリストである続報リストにクラスタリングする方式を提案する．また，記事中から過去や未来の時間を抽出し，それをもとにより精度の高いクラスタリングを行う方式について述べる．さらに，実際のオンラインデータを用いた実験結果と考察について述べる．

A Time-series Clustering Mechanism for Multi-channel Type Information Dissemination Systems

Kazutoshi Sumiya,[†] Koichi Matsumoto,[††] Minori Takahashi[†††] and Kuniaki Uehara[†††]

In this paper, we propose a clustering technique for time-series on-line news articles on the Internet. The articles are time dependent data deseminated from news servers via multi-channels based on their topics. We describe a detecting technique for articles whose topics have any relations, called follow-up articles. We also describe a technique of clustering articles based on time information extracted from news text. Furthermore we describe a prototype system based on the proposed technique. The system can reconstruct all of the articles to follow-up article clusters and distributes the articles to users. Finally, we describe the result of an experimental evaluation of the system.

1. はじめに

近年，インターネットを用いたプッシュ型配信サービス[1)]が注目を浴びている．Infogate[2)]や Yahoo!News[3)]をはじめとする WWW 上でのニュースサイトによるインターネットニュース配信などがその代表例である．従来のプル型配信ではインタラクションによって欲しい情報を検索し，絞り込む必要があった．しかしながら，プッシュ型配信では，情報をリアルタイムに送ることが可能である．したがって，時事ニュースなどのリアルタイムに配信される情報はプッシュ型配信に適していると考えられる．しかし，従来のプッシュ型配信には以下の問題がある：

- チャンネルが多数あり，大量のニュース記事が配信されているので，目的のニュース記事を見つけるのが困難である．
- ある 1 つのトピックに関して複数のニュース記事が存在し，それらの記事が異なるチャンネルに分散している場合がある．

大量の情報からユーザの要求に見合ったものだけを選択するための手法として，ユーザプロファイルによる情報フィルタリングがある．しかし，ユーザプロファイルを使用する手法では，あらかじめユーザがプロファイルを入力する必要があり，インタラクションを必要とせずに情報を取得できるプッシュ型配信システムの

† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University
†† 株式会社大林組エンジニアリング本部情報エンジニアリング部
Obayashi Corporation
††† 神戸大学大学院自然科学研究科情報知能工学専攻
Division of Computer and Systems Engineering, Graduate School of Science and Technology, Kobe University

メリットが失われてしまう．そこで我々は，ユーザプロファイルではなく配信される情報自身の内容から重要度を計算すべきであると考えた．

本論文では，オンラインニュースのマルチチャンネル型ニュース配信システムと時間情報に基づくクラスタリング方式について述べる．本方式では，オンラインニュースの内容情報と時間情報に基づいて続報の検出を行う．また，続報予定の概念を提案し，それをもとに，ニュース記事群を再構成する方法について述べる．さらに，実際のオンラインデータを用いた実験結果と考察について述べる．

以下，本論文では，2章では関連研究について述べ，3章では，時間的性質に関するオンラインニュースの特性について述べる．4章では，提案する時系列クラスタリングと続報予定について述べる．5章では実験結果を示し，評価を行う．最後に，6章ではまとめと今後に課題について述べる．

2. 関 連 研 究

2.1 情報配信システム

Infogate[2]は，チャンネルメタファを用いたニュース配信システムである PointCast[4]の後継システムである．Infogate はデスクトップ上に専用のツールバーを配置し，そこにニュースのヘッドラインを表示するシステムである．PointCast のパーソナライズ機能であったチャンネルの追加・削除に加えて，“alerts lists”によって，ニュースにあらかじめ登録しておいたキーワードが含まれていた場合に通知することができる．このシステムはパーソナライズ型マルチチャンネル型放送配信システムであると考えられる．

Yahoo!News[3]，goo のホットチャンネル[5]や Lycos の NEWS CENTER[6]などの WWW 上でのニュースサイトによるインターネットニュース配信もマルチチャンネル型放送配信システムである．どのニュースサイトも国際，社会，政治，経済などのチャンネルにカテゴライズしてニュースを配信している．また，ニュースサーバと Web クライアントの間は，Web クライアントによる周期的プルを利用しており，完全なプッシュ型サービスではない．

Omniviewer[7]はニュースビューアークライアントである．与えられた URL のホームページからティッカー☆やヘッドラインと呼ばれるものを生成する．ニュースを対象に作られているが，ニュースの時間情報や複数チャンネルを考慮していない．

2.2 配信コンテンツの再構成

山本[8]の LiveText は情報をコンテンツとコンテクストに分けてプッシュ型配信する方式である．コンテンツは情報の中身，コンテクストは情報を見せる枠組みを表すものである．すなわち，サーバから文字情報（コンテンツ）のみを配信し，レンダラーというクライアントでコンテクストに基づいて画像を付加して視覚化するシステムである．本研究と同様に，Web コンテンツを素材として扱っているが，情報自体の性質や重要度は考慮せずに呈示している点が本研究とは異なる．

清田ら[9]は，オンラインニュース記事を自己組織化マップを用いてクラスタリングし，自動整理するシステムを作成している．しかし，オンラインニュースの時間属性についてはまったく考慮していない．

2.3 時系列データ処理

マサチューセッツ大学の Allan ら[10]とカーネギーメロン大学の Yang ら[11]は，TDT（Topic Detection and Tracking）を提案している．TDT は，オンラインニュースの記事列からトピックを検出し，トラッキングする機能を有している．トピックを検出し，その続報となるニュース記事を監視することは，本研究の目的と同様であるが，記事の内容のみに基づいたクラスタリングであり，時間属性は考慮されていない．

一方，Yang らは IncrementalIDF 法も提案している．これはオンラインニュースの性質として，文書数が時間とともに増えてくるのに対応した IDF 法である．本方式では IDF 値の計算のために過去のすべての文書を対象としている．したがって，随時配信されてくるニュース記事に対して毎回，全記事を対象に再計算が必要である．

宗像ら[12]は，周期的に発生する数値データ列が複数あり，それぞれ異なる周期で発生している状況から，データの鮮度と同期度に基づいてデータの組合せを選択する手法を提案している．時間情報からデータの重要度を算出しているが，この方式の対象は，周期的データであり，ニュース記事のような周期性を持たないデータに対しては適用できない．

馬ら[13]は，ユーザプロファイルに加えて時系列的な特徴量を用いて，時系列文書のフィルタリングを行っている．時系列的特徴量を用いている点が本研究と類似しているが，マルチチャンネル環境での時系列文書を対象としていない点で異なる．

☆ 電光掲示板のように文字が固定領域内で左右に流れて表示されるもの．

3. オンラインニュースの特性

本章では，オンラインニュースの特性について述べる．まず，時系列文書について考察し，その特性について述べる．また，時系列文書間の関係についても論じる．

3.1 マルチチャンネル型配信システム

現在，WWW 上でのニュース配信サーバとして，Yahoo!ニュース[3]，goo のホットチャンネル[5]や Lycos の NEWS CENTER[6]などのニュースサイトがある．これらのニュースサイトではリアルタイムに時事ニュースが一定の時間間隔で配信されている☆．

たとえば，あるニュースが配信され，その数時間後に追加や訂正などが配信されることがある．また，一連のニュース記事のまとめのニュース記事が配信される場合もある．すなわち，この状況は，あるトピックに関するニュースの時間的な列となったニュース群が存在すると考えられる．このように時間情報を含む関連文書を**時系列文書**と呼ぶ．

各ニュースは社会・経済・国際などの分野ごとにカテゴライズされたチャンネルで配信されている．また，ニュース記事はその複数のチャンネルで連続的に（ストリーム）配信されている．従来システムでは，サーバはニュース記事を独立に配信し，クライアントはニュース記事を直接に受け取っていた．本研究では，サーバからすべての記事を受け取り，フィルタリングして再構成してからクライアントに再配信する方式を提案する（**図 1**）．すなわち，サーバとクライアント間に存在する中間サーバとして機能するシステムを想定している．

なお，本論文では，マルチチャンネルとして，分野のカテゴライズとマルチソース（サイト）が混在している環境を対象としている．しかし，実際には，1 つのソースに対して適用するのが一般的であると考えられる．

3.2 時系列文書

一般に，時系列文書は以前に配信した内容に関する情報に追加や誤りがあった場合，追加・訂正が配信される．このような過去に配信されたニュース記事と同一のトピックについて配信された記事を**続報**と定義する．続報とその元となるニュース記事との間には以下のような関係がある．

- 文書間の類似度が高い．

☆ 実際にはクライアントが定期的にプルしている．すなわち，周期的プルである．

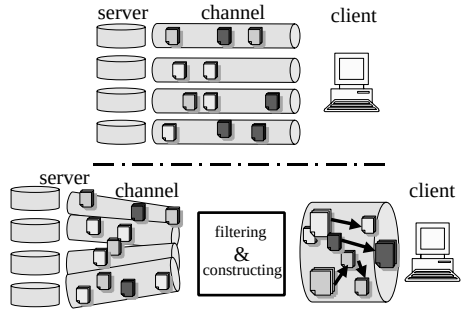

図 1 従来システム（上）と提案システム（下）
Fig. 1 Conventional system and proposed system.

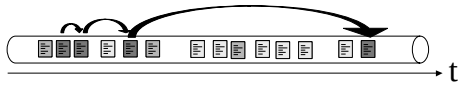

図 2 ノーマルパターン
Fig. 2 Normal pattern.

- 配信時間の差がそれほど大きくはない．

あるトピックに関して，後で発生した追加・訂正などの続報と，その元となったニュース記事との間には，深い関連がある．そこで，続報を単独で扱うのではなく，まとめて扱う必要がある．同一のトピックに関するニュース記事の発生順に並べたリストを**続報リスト**（以下，FA: Follow-up Article list と表記する）と呼ぶ．続報リストは次節に示すような性質がある．

3.3 続報リストの分類

我々は，続報リストをニュース記事のつながり方から以下の 3 種類のパターンに分類した．（（1）ノーマル，（2）トランジット，（3）リターン）このうち，（2），（3）のパターンはマルチチャンネル環境においてのみ発生することに注意を要する．

3.3.1 ノーマルパターン

図 2 のように，ある続報リストに属するニュース記事のすべてが同一のチャンネルから配信される場合である．たとえば，銀行強盗事件が起こった場合，まもなく社会チャンネルで，その事件の発生を伝えるニュースが流される．その数日後，その強盗事件の犯人が捕まったニュースもまた，社会チャンネルで配信される場合である．

3.3.2 トランジットパターン

図 3 のように，チャンネル B において，同一のトピックに関する 3 つのニュース記事が配信される状態がある．その後，チャンネル A において，前述の 3 つのニュース記事と同一のトピックに関する新しいニュース記事が配信されてくることがある．この新着のニュース記事は，3 つの記事と配信されてきたチャンネルは異なるが続報と見なされるべきである．

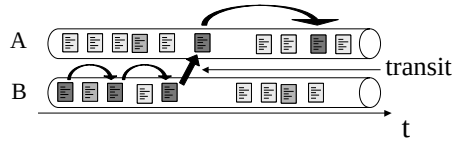

図 3 トランジットパターン
Fig. 3 Transit pattern.

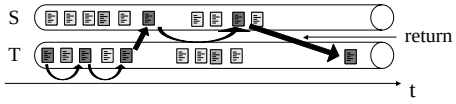

図 4 リターンパターン
Fig. 4 Return pattern.

たとえば，台湾地震のニュースが国際チャンネルで頻繁に配信されている状況において，しばらくして経済チャンネルで PC 用のメモリの値段が高騰したというニュースが配信されるような場合が考えられる．これら 2 つの出来事は，配信されるチャンネルは違うが，PC 用のメモリの大部分が台湾で生産されているという事実を考えると，同一トピックを扱った記事であると考えられるからである．

3.3.3 リターンパターン

図 4 に示すように，チャンネル S，T 間において，事前にトランジットパターンがあったとする．その後，S チャンネルから T チャンネルへのトランジットが起こることが考えられる．これをリターンパターンと呼ぶ．なぜなら，この S チャンネルから T チャンネルへのトランジットは元々配信があったチャンネルへのトランジットであるためである．

たとえば，トランジットパターンの例であげた，台湾地震とメモリ高騰のニュースの後，台湾経済の回復のニュースが台湾地震のニュースが以前に配信された国際チャンネルにおいて，配信されるような場合である．

4. 時系列クラスタリング

本章では，続報リストの生成方式と続報リストのパターン分類について述べる．

4.1 続報リスト 生成方式

続報を検出するために，ニュース記事を特徴ベクトルを用いて類似度を計算する．新しいニュース記事 a_{new} があるチャンネルで配信されてくると，特徴ベクトルが計算される．本研究では，ニュース記事群から抽出したキーワードを要素とする 1500 次元の特徴ベクトル☆を用いる．新着ニュースと続報リストの類似度は特徴ベクトルのコサイン相関値により計算される．続報リストの特徴ベクトルは続報リストに属する最後のニュース記事の特徴ベクトルとする．この理由は，最後のニュース記事は，そのクラスタの最新の特性を持っていると考えるからである．しかし，もちろん a_{new} がクラスタの最後のニュース記事以外の記事との高い類似性を持つ場合が考えられる．一般的にはクラスタのトピックは時間的に変化するために，現時点でのクラスタを代表する記事は，最後の記事を採用する．2 つのベクトル $\vec{a}_1$，$\vec{a}_2$ のコサイン相関値 $sim(\vec{a}_1, \vec{a}_2)$ は，以下の式によって計算される．

$$sim(\vec{a}_1, \vec{a}_2) = \frac{\vec{a}_1 \cdot \vec{a}_2}{|\vec{a}_1||\vec{a}_2|} \quad (1)$$

ある続報リストと新着ニュース記事との類似度がしきい値以上であり，かつ，他のどの続報リストとの類似度よりも大きいならば，新着ニュース記事はその続報リストの要素として追加される．

次に，マルチチャンネル型配信システムで配信されるニュース記事のためのクラスタリング機構を提案する．提案アルゴリズムでは，主に以下の 2 つの情報に注目して，オンラインニュースをトピックごとにクラスタリングする．

- 新着ニュース記事と既存の続報リストの類似度
- 新着ニュース記事と既存の続報リストに最後に追加された記事との配信時間差

マルチチャンネル型配信においても，類似度判定にはコサイン相関値を用いる．配信時間差の判定は，単一チャンネル環境とマルチチャンネル環境とで異なる．もし，新着ニュース記事が同じチャンネルに存在する続報リストに追加されるならば，その配信時間間隔は制限されない．つまり，どんなに新着記事の配信まで時間間隔が大きくても既存の続報リストに追加する☆☆．もちろん，新着ニュース記事と続報リストの類似度がしきい値より高い場合のみである．

一方，新着ニュース記事が続報リストと違うチャンネルで配信された場合は，続報リストへの追加が制限される．もし，配信間隔が大きい場合は，新着ニュース記事は続報リストに追加されない．なぜなら，時系列文書の性質として，トピックが他のチャンネルに移り変わるというのは，あるニュースの影響を受けて新しいニュースが発生したということであり，2 つの配信間隔が大きくならないからである☆☆☆．しかしながら，チャンネルが異なり，配信間隔が大きくても続報

☆ これについては，詳しくは 5.1 節で述べる．

☆☆ 時系列クラスタリングサーバに蓄積される記事の有効時間は，しきい値よりもはるかに大きいと仮定している．

☆☆☆ 逆にいうと，チャンネルが異なり，かつ配信間隔が大きい記事どうしは，トピックが違うと考える方が妥当である．

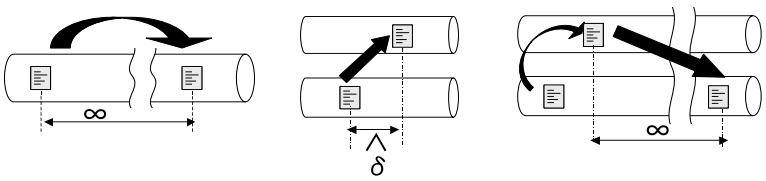

図 5 配信時間間隔制限
Fig. 5 Time interval of adding a follow-up article.

リストに追加される場合もある．続報リストが新着ニュース記事と同じチャンネルから配信されたニュース記事を含んでいる場合である．つまり，トピックは複数のチャンネルに点在していると考えられる．これはリターンパターンの典型である．

図 5 に，新着ニュース記事と続報リストの最終ニュース記事との配信時間間隔の制限を示す．左の図は，ノーマルパターン，中央の図はトランジットパターンを示し，右図はリターンパターンを示している．

本研究で提案するクラスタリングのアルゴリズムを**図 6** に示す．関数 $interval(x, y)$ はニュース記事 x と y の配信時間間隔を返す関数である．

ただし，$channel(a)$ は記事 a の配信チャンネルを返す関数であり，$articles(FA)$ はその続報リストに属するすべてのニュース記事のリストを返す．また，$last(FA)$ は続報リストに最後に追加されたニュース記事を返す関数である．アルゴリズムの概略を以下に示す．

1. 新着ニュース記事 a_{new} とすべての続報リスト FA_i の類似度を計算し，最も類似度が高い続報リストを追加先続報リスト候補とする．
2. 選ばれた追加先続報リスト候補と新着ニュース記事との類似度があらかじめ設定したしきい値を超えるかどうかを判定する．超えなければ，新しい続報リストを作成し，新着ニュース記事を最初の要素とする．
3. 追加先続報リスト候補の最後のニュース記事と新着ニュース記事の配信チャンネルが等しいならば，新着ニュース記事は追加先続報リスト候補に追加されて終了（ノーマルパターン）．
4. 追加先続報リスト候補の中に，新着ニュース記事と同じチャンネルで配信されてきたニュース記事がないかを調べる．もしあれば，新着ニュース記事は追加先続報リスト候補に追加されて終了（リターンパターン）．
5. 追加先続報リスト候補の最後のニュース記事と新着ニュース記事の配信時間間隔がしきい値より小さければ，新着ニュース記事は追加先続報リスト候補に追加されて終了（トランジットパターン）．
6. Step1 から 5 までに，あてはまらなかったときは，新規続報リストを作成し，新着ニュース記事を最初の要素とする．

---

1. **for** $1 \leq i \leq n$
　if
　$(sim(FA_i, a_{new}) > sim(FA_{cand}, a_{new}))$
　　$cand = i$;
　end
2. **if** $sim(FA_{cand}, a_{new}) < \theta$
　then create $\langle FA_{new} \rangle$
3. **else if**
　$channel(a_{new}) = channel(last(FA_{cand}))$
　then $a_{new} \rightarrow \langle FA_{cand} \rangle$
4. **else if** $channel(a_{new}) \in$
　$channel(articles(FA_{cand}))$
　then $a_{new} \rightarrow \langle FA_{cand} \rangle$
5. **else if** $interval(last(FA_{cand}), a_{new})) < \delta$
　then $a_{new} \rightarrow \langle FA_{cand} \rangle$
6. **else** create $\langle FA_{new} \rangle$

図 6 時系列クラスタリングアルゴリズム
Fig. 6 Time-series clustering algorithm.

上記のステップを新たに配信されてきたオンラインニュース記事を対象に行うことで続報リストを生成する．

4.2 続報予定

前章で述べた配信時間と類似度のクラスタリングの方式では，同じトピックについて書かれてはいるが，類似度がしきい値より低く，配信時間が離れている場合などはうまくクラスタリングされない，といった問題点が存在する．

そこで，記事中から過去や未来の時間を抽出し，クラスタリングのための要素とする**続報予定とイベントリスト**を提案する．

4.2.1 続報予定

オンラインニュースは，あるイベントが発生した後，それに関する複数のニュースが時系列で配信される．このニュース記事はイベントよりも後に発生している．一方，イベントにはあらかじめ X 月 Y 日にイベントが開催されるといったような，未来の内容を含んだニュース記事が存在する．このようなイベントの発生予定時刻近辺には，そのイベントに関するニュース記事が配信される確率が高いと考えられる．このような配信予定のことを**続報予定**（FC: Follow-up Candidate）と定義する．つまり，配信されてきたニュースの文中に時間表現があって，その時間が未来ならば，その時間に続報の配信が予想される．このような配信の予定のことを続報予定と呼ぶ．

たとえば，アメリカ大統領選挙を例にあげると，12 月 11 日の国際チャンネルで,「大統領選の当落を決する米フロリダ州における大統領選で，同日中に予想されていた 疑問票手集計判決は 12 月 12 日に持ち越されそうだ．連邦最高裁が手集計を退ければ，ブッシュ氏の当選が事実上，確定する」というニュース記事が配信されると，このニュース内容より，元のチャンネルと同じ国際チャンネルに 12 月 12 日に疑問票手集計判決という続報予定が仮想的にマッピングされる☆．

続報予定は以下の情報からなる．

- 配信チャンネル（FC_{ch}）
- イベント発生予定時刻（FC_{time}）

これらは，配信されてきたニュース記事内容から抽出されるので，続報予定には必ずそれを掲載した元ニュース記事が存在する．もちろん，すべてのニュース記事が続報予定となるわけではない．

この 2 つの情報，および，記事間の類似度を用い，新しい記事が到着したときに，それが続報予定群に含まれる続報予定で予定された記事であることを判定する．つまり，続報予定の持つ情報と同配信チャンネルで同イベント発生時刻が記された記事が発見されれば，それと元記事との類似度を計算する．この類似度がしきい値よりも高ければ，この記事は予定された記事であると判断され，続報リストに追加される．

図 7 において，現在時刻（Now）より過去に記事があるのはすでに配信されているニュース記事を表す．また，それらはトピックに関する続報リストを形成しているものとする．現在時刻よりも未来の部分が続報予定（図 7 では “△” で示している）である．

☆ 図 7 の “△” 印が元記事からマッピングされた続報予定である．

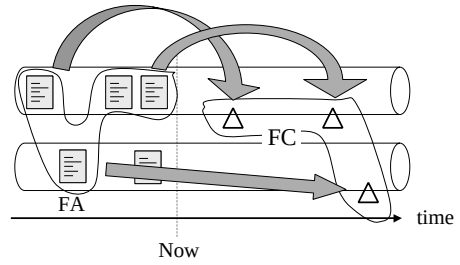

図 7 続報予定
Fig. 7 Follow-up candidates.

4.2.2 続報予定検出方式

次に続報予定の検出方式について述べる．あるチャンネルにおいて，新しいニュース記事 a_n が配信されてきたときに，以下の手順で続報予定の検出を行う．

(1) a_n の文章中から抽出した時間情報が現在時刻よりも未来であれば，それは予定されたイベント時間であるので，続報予定の時間情報として抽出する．もし，未来の時間情報が抽出できなければ終了．

if $now < extract\_time(a_n)$
then $FC_{time} = extract\_time(a_n)$
else END

$extract\_time(a_n)$ は，記事 a_n 中から時間情報を抽出する関数とする．

(2) a_n がどの続報リストに属しているかという情報を取得し，その続報リストに対応する続報予定群に続報予定を追加する．もし，対応する続報予定群がなければ，新規作成．

if $isthere(fa2fc(whichFA(a_n)))$
then Add $(fa2fc(whichFA(a_n)), a_n)$
else $FC_{fc} =$ create $< FC_{new} >$

$whichFA(a_n)$ は記事 a_n が属している続報リストを返す関数であり，$fa2fc(FA)$ は，続報リスト FA に対応する続報予定群 FC を返す関数である．また，$isthere(FC)$ は対応する続報予定群が存在しているかどうかを返す関数である．

(3) 配信されてきたニュースのチャンネル情報を取得する．このチャンネルに続報予定を配置する．

$FC_{ch} = channel(a_n)$

得られた情報から新しい続報予定を生成し，既存あるいは新規の続報予定群に追加する．

複数の続報予定を関連付けるために，続報予定を生成するための情報を抽出する元ニュース記事が属する続報リストを用いる．すなわち，同じ続報リストに属すニュース記事を元とする続報予定は同じ続報予定群に属するとする．これにより，同じトピックに関するイベントを検出することが可能である．

5. 時系列クラスタリングの精度評価

5.1 予備実験 (1)：オンラインニュース記事の次元と類似度

5.1.1 対象とするニュース記事

時系列クラスタリングの精度を評価するために，Yahoo!JAPAN NEWS の社会チャンネル，経済，国際の 3 チャンネル，2000 年 1 月 24 日から 2000 年 1 月 30 日までの 1 週間分のニュース記事を対象に実験を行った．総記事数は 1,514 記事であった．

Yahoo!JAPAN NEWS のニュース記事のみを対象とした理由は，あらかじめトピックを表すキーワードがタイトルの最初に付けられて配信されてくるためである．精度評価の際に，このキーワードが同一のものを同一トピックとすることで，クラスタリングの評価が正しく行えると考えられる．

5.1.2 抽出するキーワード

予備実験 (1) では，下記のようなルールで特徴ベクトル生成のためのキーワードの抽出を行った．

- 茶筌[14]によって，「名詞」と品詞分類された単語のうち次の条件をすべて満たす場合に限ってキーワードとする．
  - 2 文字以上である．
  - 少なくとも 2 つの文書に含まれている．
  - 最初の 1 文字が数字でない．

（IT）などの半角記号に囲まれているものは，IT として扱った．

5.1.3 実験結果・考察

1,514 記事からプログラムを用いてランダムに 308 記事を選んだ．そして，それらについて上述のルールに従ってキーワードを選び，それらの中から tf·idf 値上位 n 語を用いて，n 次元特徴ベクトルを作り，類似度をコサイン相関値によって計算した．

次元が上がるにつれて，類似度が下がっていく傾向がみられた．これは，類似度を比較する文書間において共通する単語から得られる特徴ベクトルの値が，次元が上がることによって，正規化の際に小さくなってしまうことが原因と考えられる．1500 次元以下では，文書セット間の類似度の開きがかなり大きい傾向がみられた．これは対象とするニュース記事の文書の長さが比較的短いのに加えて，共通する単語が次元数の低い段階で両文書に現れたためと考えられる．

1500 次元以上では類似度は，ほぼ一定の値となった．これはすなわち，類似度判定に必要な特徴ベクトルの要素となる単語が 1500 次元程度でカバーできていることを示していると考えられる．

したがって，本研究では対象がニュース記事であるので，リアルタイム性が必要とされることを考慮して，精度をあまり下げずになるべく低い次元で高速にクラスタリングを行うことを目的とし，以下，1500 次元で評価実験を行うものとする．

5.2 予備実験 (2)：続報リストと類似度しきい値

5.2.1 続報リスト生成

以下のようにニュース記事から続報リストの生成実験を，類似度しきい値を 0.1 から 0.9 まで変化させて行った．クラスタリングの際の類似度判定には，予備実験 (1) と同じく，$tf \cdot idf$ 値により重み付けしたキーワードからなる特徴ベクトルのコサイン相関値による類似度を用いた．

(1) 既存のすべての続報リストとの類似度を計算する．
(2) 類似度が最大の続報リストを追加先候補とする．
(3) しきい値以上なら，追加．未満なら，新規続報リストを生成し，その第 1 要素とする．

5.2.2 評価方法

生成した続報リストがトピックごとのニュース記事のリストになっているかどうかを再現率（式 (2)）・適合率（式 (3)）・F 値を用いて評価した．

評価の際のパラメータは以下のとおりである．

総正解ニュース数 今回は，あらかじめ各記事に正しいトピックが付けられている Yahoo!JAPAN NEWS の毎日国内，毎日経済，毎日国際の 3 チャンネルのニュース記事のみを実験データとしたため，すべての記事にトピックが付けられてあるが，そのうち 2 つ以上の記事に付いてるトピックの記事数とする．本実験では 1,175 であった．

クラスタリング記事数 生成された続報リストのうち，2 つ以上の要素数を持つ続報リストに属するニュース記事の総数

クラスタリング正解記事数 システムによって生成された続報リストに含まれるニュース記事のうち，以下の条件を満たすニュース記事の総数

- クラスタリング文書であること

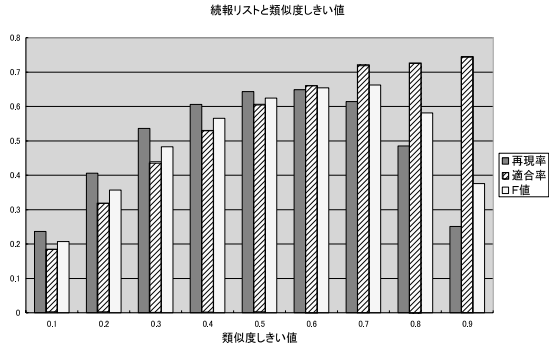

図 8 類似度しきい値
Fig. 8 Threshold of similarity.

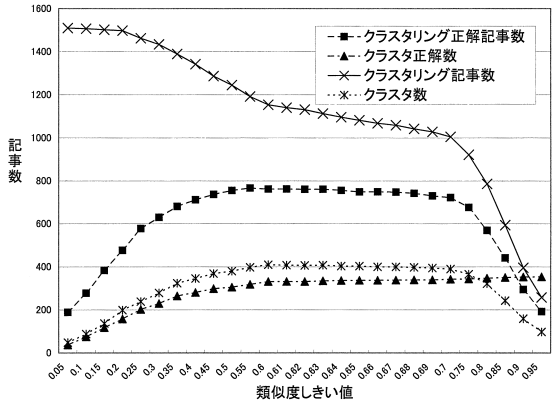

図 9 記事と続報リスト
Fig. 9 Articles and clusters of FA.

- 全続報リストのうち，あらかじめ付けられているトピックが同じニュース記事が，その記事が属している続報リストに最も多く存在していること

あらかじめ付けられている正しいトピックごとに上記の条件を満たす記事数を計算し，合計したものがクラスタリング正解文書数となる．

$$recall = \frac{\text{クラスタリング正解記事数}}{\text{総正解ニュース記事数}} \quad (2)$$

$$precision = \frac{\text{クラスタリング正解記事数}}{\text{クラスタリング記事数}} \quad (3)$$

$$F-measure = \frac{(b^2+1)\times precision \times recall}{b^2 \times precision + recall} \quad (4)$$

実験結果を**図 8** に示す．

5.2.3 実験結果・考察

図 8 から，適合率は類似度しきい値が小さいところで低い値を示している．これは予備実験 (2) のときの評価パラメータの推移を示した**図 9** を参考にして考えると，しきい値が低い場合のクラスタリング記事数が非常に多いことが原因と考えられる．

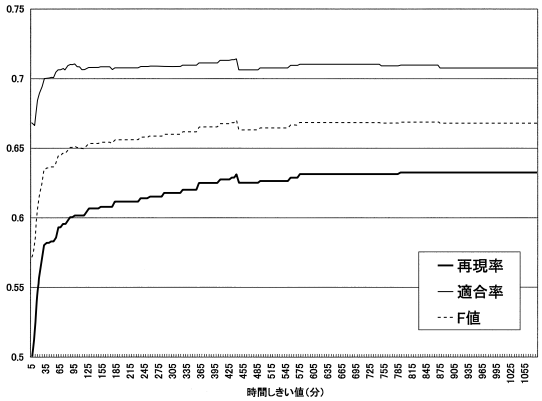

図 10 続報リストと配信時間しきい値
Fig. 10 Threshold of dissemination interval.

一方，一般的に適合率とトレードオフの関係にあるはずの再現率がしきい値の低いところで大きくなっていない．これは正解記事の定め方に原因がある．一般的には，しきい値を低くすると，多数の解を得ることができるため，多くの不正解と同時に多くの正解を得ることができる．しかしながら，今回の実験では，しきい値を低くすると，1 つの続報リストに誤った記事が追加される可能性が高くなり，続報リストが大きくなってしまう．これは，本来別々のトピックのはずの記事が 1 つの続報リストにまとまってしまうために，正解クラスタリング記事数が小さくなってしまったのが原因であると考えられる．また，しきい値が高くなると非常によく似た記事のみ続報リストに追加されないために，本来追加されるべき記事が追加されずに新規続報リストを生成してしまう．結果として，クラスタリング正解記事数が増えずに再現率が低くなってしまう．したがって，しきい値の高いときと低いときの両方で再現率が小さい値を示したと考えられる．

したがって，F 値（式 (4)）を用いて評価することにした．適合率に対する再現率の相対的な重みを示すパラメータ b は今回の評価で b = 1 として扱った．F 値によって評価すると，0.67 で最も良い値を示したため，この値をしきい値とする．

5.3 実験 1：時系列クラスタリングの配信時間差とその精度

予備実験 (2) で得られた結果をもとにして，類似度しきい値を 0.67 として時系列クラスタリングを用いて続報リスト生成精度を評価する実験を行った．その実験結果の評価を**図 10** に示す．

図 10 は，1 週間分 1514 記事に対して，時間しきい値を 5 から 1,080(分) まで変化させて時系列クラスタ

リングを行った結果である．実験自体は時間しきい値10,080(分) まで行ったが，実際に結果として有効であるのは，1,080(分) までである．

グラフより，時間しきい値が小さい範囲で再現率，適合率が低くなっていることが分かる．これは本来同一のトピックに関する記事であり，類似度から判断しても同じ続報リストに追加されるはずであろう記事が配信時間差がしきい値以上であったために，別の続報リストにクラタリングされてしまった結果と考えられる．また，時間しきい値 5 の場合でも，再現率と適合率が 0 に近い値をとらないのは，クラスタリングが進むにつれ，リターンパターンでの記事の追加が増え，トランジットパターンでの記事の追加が発生しなくなるからと考えられる．

しかし，時間しきい値が大きくなると，再現率，適合率ともに，配信時間差を用いない類似度のみのクラスタリングよりも良い結果を示している．これは，類似度は大きいがトピックは異なる記事を時系列文書の特性を生かして，同一の続報リストに誤ってクラスタリングされることを防げた結果であると考えられる．

時間しきい値を大きくしていくと再現率，適合率が大きくなっていくが，時間しきい値を大きくしすぎると配信時間差を用いないクラスタリングの結果に近づいていく．したがって，適当な時間しきい値をとることによって，クラスタリング精度を向上させることが可能であると考えられる．さらに，できるだけ小さい時間しきい値をとることによって，トランジットパターンでのニュース記事の続報リストへの追加の場合，追加先続報リスト候補を類似度計算の前に絞れるために，計算量を格段に減らすことができる．これらより，図 10 で，再現率，適合率の上昇が安定して，かつ，なるべく小さい時間しきい値が時系列クラスタリングに最適であると考えられる．

評価実験より，時間しきい値は 575(分) が適当であると考えられる．また，そのときの再現率は 0.63，適合率は 0.71，F 値は 0.67 を示した．時間しきい値を用いない従来の類似度のみのクラスタリングよりも良い精度で計算量を減らせることが確認された．

5.4 実験 2：続報予定

2000 年 11 月 8 日から 2001 年 2 月 1 日までの間の，毎日国内，毎日国際，毎日経済，時事国内，時事国際，時事経済，ロイター国際，NNA 国際の計 8 チャンネルの記事を対象に時系列クラスタリングを行った．

続報予定（FC）が生成されたトピックとして，センター試験・エンデバー・紅白歌合戦・被災地サミットがあった．これらはすべて，開催日が明記されている

表 1 FC の実験結果
Table 1 Experimental results of FC.

トピック	FC 考慮前のクラスタ数	考慮後のクラスタ数	正解記事数	不正解記事数
センター試験	7	5	5	8
エンデバー	4	2	4	2
紅白歌合戦	7	3	8	10
被災地サミット	3	2	2	2

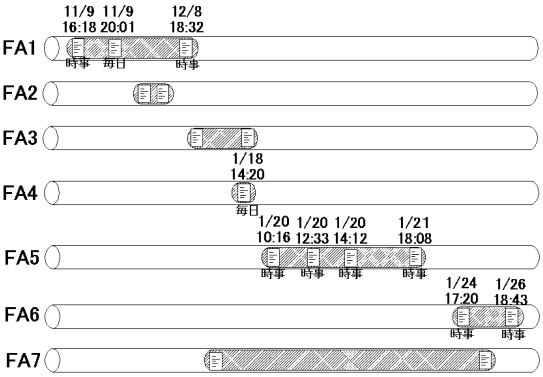

図 11 続報リスト
Fig. 11 FA.

イベントに関する記事である．各トピックについて，続報予定考慮前のクラスタ数・正確に結合された記事数（正解数）・結合されなかったり誤って結合されたりした記事数（不正解数）をまとめ，**表 1** に示す．

例として，2001 年度センター試験に関する続報リストと続報予定について詳しく述べる．その続報リストは時事国内 11 月 9 日 16 時 18 分の記事から始まり，毎日国内，時事国内の 2 チャンネルによって配信された記事から構成されていた．

本来ならば，センター試験の続報リストは 1 つにクラスタリングされるべきだが，3 章に提案したアルゴリズムによる続報リスト生成方式では 7 つに分散された．その結果，各々の続報リストの時間幅が短くなり，本来のトピックの時間幅とは異なる結果を得た．作成された続報リストを**図 11** に示す．

一方，上記の続報リスト生成方式に加えて，続報予定を考慮し，クラスタリングを行った場合，同一のトピックを持つ記事が同一のクラスタに分類された．

実験結果では，生成された続報リスト内の記事のうち，続報予定が含まれていた記事は 1 月 20 日以前の 6 記事であった．その 6 記事にはいずれも，「1 月 20，21 日に行われるセンター試験」といった記述があった．これにより，1 月 20，21 日付近には，実施されたセンター試験に関する記事が配信されるだろうことが予想できる．実際の配信においても，1 月 20 日に

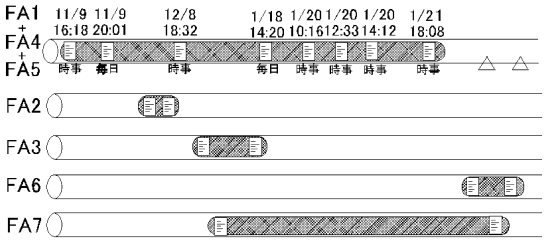

図 12 続報予定を考慮した続報リスト
Fig. 12 FA in consideration of the FA candidate.

はセンター試験に関する記事が配信された．しかし，内容を調べてみると，FA1 は受験志願者数に関する記事，FA4 はセンター試験を利用する大学に関する記事，FA5 は実際にセンター試験の開催に関する記事であった．これらの記事は，時間と類似度のみの計算結果では関連が薄いと計算され，それぞれ別々の続報リストが生成された（図 11）．これに対して，続報予定を考慮した方式を用いると，分散されていた F1，F4，F5 の 3 つの続報リスト内の記事に続報予定の記述があることにより，これらの記事が同一の続報リストに分類された（**図 12**）．

続報予定を考慮した場合でも同一クラスタに結合されなかった原因として，図 11 の F2 の記事には，「来年 1 月に実施されるセンター試験」という記述があり，日時の特定ができなかったことがあげられる．その他の記事についても開催日時の記載がなかったものについては結合されていなかった．

また，その他のトピックで結合されなかった原因として，紅白歌合戦については「大みそか恒例の … 」と書かれていた場合，また，エンデバーについては「20 日にも及ぶ今回の … 」という日時ではないが文字列では誤って日時と判別できる記述があったことがあげられる．記事中の時間記述に対する処理精度の向上は今後の課題である．

6. おわりに

本論文では，マルチチャンネル型放送配信システムにおけるオンラインニュースをトピックごとにクラスタリングする手法を提案した．提案するクラスタリング手法は，従来のクラスタリングで一般的に用いられていた内容の類似度に加えて，時系列文書の特徴である時間情報を用いてクラスタリングを行うものである．これにより，従来手法に比べて再現率と適合率が改善されることを確認した．また，オンラインニュースのトピック・クラスタリングにおいて，精度が向上することを確認した．

また，続報予定の概念を定義し，これを用いることによって，内容類似度と時系列のみの場合よりもさらに効果的なクラスタリングを行えることを確認した．

マルチソースのカテゴライズが存在した場合，同一トピックに対する記事でも，各サイトの記述の違いなどにより類似度に差が生じてくる可能性があると考えられる．これらの場合の対応については，今後の課題である．

謝辞 本研究の一部は，日本学術振興会未来開拓学術研究推進事業における研究プロジェクト「マルチメディア・コンテンツの高次処理の研究」（プロジェクト番号 JSPS-RFTF97P00501）による．ここに記して謝意を表します．

参考文献

1) 角谷和俊，宮部義幸：放送型情報配信のためのモデルとシステム，情報処理学会論文誌：データベース，Vol.40, No.SIG8(TOD4), pp.141–157 (1999).
2) Infogate: http://www.infogate.com/
3) Yahoo! JAPAN NEWS: http://news.yahoo.co.jp/headlines/top/, Yahoo Japan Corporation.
4) PointCastNetwork: http://www.pointcast.com.
5) ホットチャンネル：http://channel.coo.ne.jp/, NTT-X.
6) NEWS CENTER: http://www.lycos.co.jp/news/, Lycos Japan Inc..
7) Omni Viewer: http://www.digiportal.com/, DigiPortal Software LLC.
8) 山本 強：多ソース融合型表現メディアシステム LiveText の開発，情報処理学会デジタルドキュメント研究会報告，98-DD-13 (1998).
9) 清田陽司，黒橋禎夫，中村順一，長尾 眞：構文情報を利用した電子ニュース記事のクラスタリングシステムの作成と評価，信学技報，NCL98-17, 7, pp.15–22 (1998).
10) Allan, J., Papka, R. and Lavrenko, V.: On-Line New Event Detection and Tracking, *Proc. 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.37–45 (1998).
11) Yang, Y., Pierce, T. and Carbonell, J.: A Study in Retrospective and On-Line Evbent Detection, *Proc. 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.28–36 (1998).
12) 宗像浩一，吉川正俊，植村俊亮：鮮度と同期度に基づく周期データの選択方式，情報処理学会論文誌：データベース，Vol.41, No.SIG1(TOD5), pp.140–153 (2000).
13) 馬 強，角谷和俊，田中克己：放送型情報配信システムのための時系列性を考慮した情報フィル

タリング，情報処理学会論文誌：データベース，Vol.41, No.SIG6(TOD7), pp.46–57 (2000).
14) 松本祐治，北内 啓，山下達雄，平野義隆，松田 寛，高岡一馬，浅原正幸：日本語形態素解析システム『茶筌』version 2.2.1 使用説明書第 2 版，NAIST Technical Report NAIST-IS-TR00912, 12 (2000).

(平成 13 年 12 月 20 日受付)
(平成 14 年 4 月 3 日採録)

（担当編集委員 有川 正俊）

角谷 和俊（正会員）

1988 年神戸大学大学院工学研究科修士課程修了．同年松下電器産業株式会社入社．ソフトウェア開発環境，マルチメディアデータベース，データ放送の研究開発に従事．1998 年神戸大学大学院自然科学研究科博士後期課程（情報メディア科学専攻）修了．1999 年神戸大学都市安全研究センター都市情報システム研究分野講師，2000 年同助教授．2001 年京都大学大学院情報学研究科社会情報学専攻助教授，現在に至る．博士（工学）．ACM, IEEE Computer Society，映像情報メディア学会，地理情報システム学会各会員．

松本 好市（正会員）

1999 年神戸大学工学部情報知能工学科卒業．2001 年同大学大学院自然科学研究科情報知能工学専攻修了．同年株式会社大林組入社，現在に至る．マルチメディアデータベースの研究に従事．

高橋美乃梨（学生会員）

2001 年神戸大学工学部情報知能工学科卒業．同年同大学大学院自然科学研究科情報知能工学専攻入学，現在に至る．マルチメディアデータベースの研究に興味を持つ．

上原 邦昭（正会員）

1978 年大阪大学基礎工学部情報工学科卒業．1983 年同大学大学院博士後期課程単位取得退学．大阪大学産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授，同大学都市安全研究センター教授．2002 年同大学大学院自然科学研究科教授．情報知能工学科を兼任，現在に至る．1989 年より 1990 年まで Oregon State University，Visiting Assistant Professor．1994 年より 1996 年まで神戸大学総合情報処理センター副センター長．工学博士．人工知能，特に機械学習，マルチメディアデータベース，自然言語によるヒューマンインタフェースの研究に従事．1990 年度人工知能学会研究奨励賞受賞．人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員．