データベースシステム 123－8
（2001．1．23）

情報レベルに基づいたプッシュ型配信通知機構とその呈示方式

松本 好市 † 角谷 和俊 ‡ 上原 邦昭 ‡

† 神戸大学大学院自然科学研究科情報知能工学専攻
‡ 神戸大学都市安全研究センター 都市情報システム分野
{koichi,sumiya,uehara}@ai.cs.kobe-u.ac.jp

本稿では，インターネットにおけるオンラインニュースのためのマルチチャンネル型放送配信システムについて述べる．提案するシステムでは，トピックの時間的位置からなるオンラインニュースの情報レベルを提案し，それを基に続報群のもつトピックの展開を判断する方式について議論する．また，ニュースレベルの高い時系列文書群を構成し呈示する方法について述べる．さらに提案した方式をもとにしたプロトタイプシステムの実装について述べる．

Push based News Dissemination and Visualization System based on Information Value

Koichi Matsumoto † Sumiya Kazutoshi ‡ Kuniaki Uehara ‡

† Division of Computer and Systems Engineering,
Graduate School of Science and Technology, Kobe University
‡ Research Center for Urban Safety and Security, Kobe University
{koichi,sumiya,uehara}@ai.cs.kobe-u.ac.jp

In this paper, we describe a multi-channel dissemination system for on-line news articles on the Internet. We propose "the value" of the on-line news which is related to chronological relationship of topics, and the method which judges expansion of the topic of a follow-up article list.

Forthermore, we describe a method of reconstruction and presentation for the time-series document list with high news value.

Finally, we describe a prototype system based on proposed method.

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，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．従って，時事ニュースなどのリアルタイムな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．
- それぞれのチャンネルでは，ニュース記事の前後関係を保って配信されているが，チャンネル間のニュース記事の時間的前後関係は正しいとはいえない．すなわち，関連する記事の時間管理は十分とは言えない．

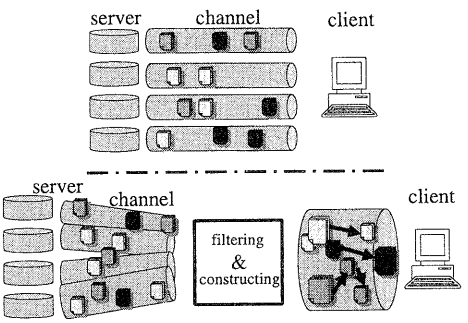

図 1: 従来システム（上）と提案システム（下）

本稿では，オンラインニュースのマルチチャンネル型放送配信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系列文書としての時間情報に基づいて続報の検出をおこなう．また，トピックと配信時間の時間情報から情報の重要度を計算する方式を提案する．また，トピックと配信時間の時間情報から得られた情報を通知する手法について述べる．

2 オンラインニュース記事

2.1 マルチチャネル配信システム

現在，WWW 上でのニュース配信サーバとして，goo のホットチャンネル [3]，Yahoo! ニュース [4]，Lycos の NEWS CENTER[5] や各新聞社のホームページ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の時事ニュースが一定の時間間隔で配信されている[1]．

例えば，あるニュースが配信された後に，そのニュースの訂正などが配信されるような状況は頻繁に起る．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．我々は，このように時系列に並んだ文書のことを**時系列文書**と定義する．

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．そして，ニュースサーバ全体としての更新間隔は一定だが，個々の記事についてはその更新間隔とは関係なく随時更新された内容の記事が新しい記事として追加されていく．

2.2 続報

オンラインニュースでは，あるイベントが発生し，時間が経過するとそのイベントに関する複数のニュースが複数のチャンネルで発生する．このようなニュース記事はそれ以前に配信されたニュース記事に関連があるので，これらを**続報**と定義する．また，これらのニュース記事は共通のトピックをもつと考えられる．従って，単独で扱うのではなくニュースの集合として扱う必要がある．このように，過去に起ったイベントについての一連のニュース記事のリストを **FA** (Follow-up Atricle list) と定義する．

[1] 実際にはユーザが定期的にプルしている．すなわち，周期的プルである．

2.3 続報予定

オンラインニュースでは，FA のようにあるイベントが発生し，それに関する複数のニュースが発生する．それらのニュース記事はイベントよりも時間的後に発生している．しかし，ニュースにはあらかじめ X 月 X 日にイベントが開催されると予定されているようなイベントが存在する．これを**続報予定**と見なす．また，一つの予定されているイベントに関連する続報予定の集合を **FC** (Follow-up Candidate list) とする．

例えば，アメリカ大統領選挙を例に挙げると，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 日に疑問表手集計判決という続報予定が誕生する．

続報予定は以下の要素からなる．

- チャンネル情報
- 発生予定時刻
- 元記事の属する FA 情報

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生み出した元ニュース記事が存在する．また，すべてのニュース記事が続報予定を生み出す訳ではない．

図 2 において，現在時刻より過去にあるのはすでに配信されているニュース記事をあらわす．また，それらは一つのトピックに関する FA 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 FC であり，イベント発生予定のみでニュース記事はまだ存在しないので，文書を思わせるような図は用いずに" 時" をイメージした△であらわしている．

情報を抽出するニュース記事に関連がある場合，すなわちそれらの記事が FA を構成している時に，それらの記事から作ることができるそれぞれの続報予定も関連があると考えられ，それらは FC であると考えられる．そこで，次に FC の検出方式について述べる．

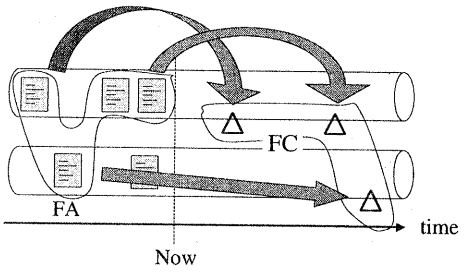

図 2: 続報予定

2.3.1 続報予定検出方式

あるチャンネルにおいて，新しいニュース a_n が配信されてきた時に，以下の手順で続報予定の検出をおこなう．

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，FC 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 $now < ext\_time(a_n)$
   then $Inf_{time} = ext\_time(a_n)$
   else END

2. a_n がどの FA に属しているかという情報を取得し，その FA に対応する FC を続報予定の追加候補とする．もし，対応する FC がなければ，新規作成．

   if $isthere(fa2fc(whichFA(a_n)))$
   then $Inf_{FC} = fa2ua(whichFA(a_n))$
   else $Inf_{FC} = \text{create} < UA_{new} >$

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

   $Inf_{ch} = channel(a_n)$

4. 配信されてきたニュース記事から，続報予定のトピックを抽出する．

$Inf_{topic} = ext_topic(a_n)$

5. これまでのステップで得られた情報である

   - Inf_{time}
   - Inf_{FC}
   - Inf_{ch}
   - Inf_{topic}

   の各情報から新しい続報予定を生成し，既存あるいは新規のUAに追加する．

上記のUA検出方式により，同じトピックに関するまだ起っていない複数イベントを検出することができる．

複数の続報予定をどのようにして関連付けるかについては，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に注目した．元ニュース記事はFC検出より先にFAにクラスタリングされているので，元ニュース記事と同じFAに属すニュース記事を元とする続報予定もまた関連があると考えれるので，同じUAに属するとする．

3 情報レベル

マルチチャネル環境において，時間の経過と共に配信されるニュースの価値は，そのニュースのトピックの大きさだけでなく，トピックのもつ時間と配信時間との関係や配信されているチャンネル情報などにも依存する．本論文では，これらの特徴量を**情報レベル**と呼ぶことにする．具体的には以下からなる：

- トピックの時間的位置
- トピックの分離・依存
- イベントの延期に伴う情報価値の変化

3.1 トピックの時間的位置

ニュースは話題の中心となるトピックの時間（以下，トピック時間とする）配信時間の時間的に前後関係によって，以下のように大きく2つに分けることができる．

- 事前ニュース
- 事後ニュース

これらのニュースの性質は一般的に異なるので，どちらのニュースであるかを考慮して，ニュースを取り扱うべきである．

さらに，あるニュースがこれらの一方から他方へ変化する場合についても論じる．

3.1.1 事前ニュース

事前ニュースとは，トピックがもつ時間情報が現在配信されているニュースに対して未来にあるニュースである．図3に示されるように，トピック時間をt_{topic}，対象ニュースの配信時間をtとすると，

$$t < t_{topic} \tag{1}$$

このニュースは，現在配信されているニュースは時間的未来にあるトピックに直接関係あるものではなく，トピックの中心を決定する材料に関するニュースである場合がほとんどである．

従ってこのタイプのニュースは，この内容の核となるべきトピックのトピック時間は現在の配信よりも未来にあり，時間的余裕があるために即時性に重点をおく必要はない．また，トピックの結果はまだ出ていないため，ニュースとしての価値も低いと考えられる．

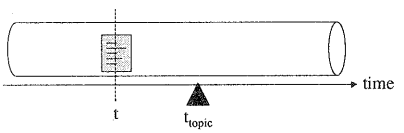

図 3: 事前ニュース

例を挙げると，シドニーオリンピックのように開催期日が予定されているようなトピックにおいては，オリンピックの数ヶ月前からたくさんのニュース配信される．このようなニュースが事前ニュースである．これらはすべてオリンピック本番というメイントピックに先駆けてのニュースである．大抵の場合，オリンピックより時間的に先に配信されているので，急いで配信しないと情報が古くなるということはない．なぜなら，より注目すべきニュースはオリンピックが開催以降のニュースであるはずである．

3.1.2 事後ニュース

事後ニュースとは事前ニュースよりも一般に多くみられるニュースであり，図4のようにトピック時間が配信時間よりも過去の場合のニュースである．

トピック時間を t_{topic}，対象ニュースの配信時間を t とすると，下式のように定義できる．

$$t_{topic} < t \tag{2}$$

事後ニュースでは，現在配信されているニュースのトピックの中心は既に過ぎ去っているので時間が経つにつれ，情報としての価値が下がってしまう．従って，配信には，内容の正確さはもとより，即時性も要求される．

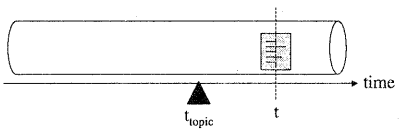

図 4: 事後ニュース

あらかじめ予定されていない時刻，すなわち”突然”に起こったイベントについて生成された内容が配信されたものである．従って，内容の中心は第一報が配信された時間より過去にある．

例えば，いつ起るか予想のつかない地震や事故などのイベントが発生し，その被害，余震や逮捕などのニュースがその経過に伴って配信される．

3.2 トピックの分離・依存

次に，マルチチャネル環境でのトピック時間の変更について考える．

図5,6は共にあらかじめ予定されていたトピック時間 (t_{topic}) が t'_{topic} 延期された場合の図を示している．

3.2.1 トピック分離

図5では，あるチャンネルでトピック時間が t'_{topic} に延期されたにも関わらず，他のチャンネルで関連するニュースが配信されている．従って，他のチャンネルで配信されたニュースはこのFAの現時点でのメイントピックである延期されたニュースの内容の影響を受けたニュースではないと考えられる．

他チャンネルで配信されたニュースは，メイントピックの時間に依存していないため，このFAのトピックは事前であるが，このようにトピック時間の変更の影響を受けないニュースは，中心となるトピックは延期されたものではないと考えられる．つまり，事前ニュースではなく，事後ニュースである．

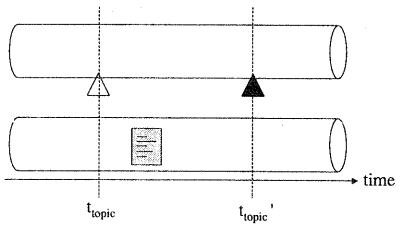

図 5: トピック分離

例えばアメリカ大統領選挙に関するニュースにおいて，フロリダ州での投票の最終結果が延期されるというニュースが国際チャンネルで配信された状況において，その後に経済チャンネルで，パソコン関係のイベントでの講演でビルゲイツ氏がパンチカードを使った投票方式を皮肉ったといったニュースが配信されたとする．

これら2つのニュースは大きく分類すると共に大統領選に関するニュースではあるが，国際チャネルで流れたニュースが大統領選の結果という時間的に未来にあるメイントピックに関する記事であるのに対し，経済チャネルで流れたニュースは，ビルゲイツ氏の講演という出来事がメイントピックになっている．従って，国際チャンネルで流れたニュースは事前ニュースであり，経済チャネルで流れたニュースは事後ニュースである．

3.2.2 トピック依存

一方，図6では，あるチャンネルでトピック時間が t'_{topic} に延期されると，t'_{topic} 以降に他のチャンネルにおいて，関連するニュースが配信されている．従って，他のチャンネルで配信されたニュースはこのFAの現時点でのメイントピッ

クである延期されたニュースの内容に依存したニュースであると考えられる．

従って，Bチャンネルで配信されたニュースのメイントピックは，t'_{topic} にAチャンネルで配信されたニュースであるので，事後ニュースであると考えられる．

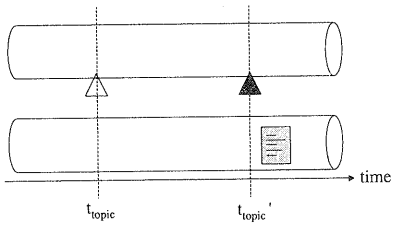

図 6: トピック依存

例えば，Aチャンネルを国際チャンネルとし，延期されたニュースは「アメリカ合衆国大統領がブッシュ氏に決定」というニュースだとする．そして，Bチャンネルが政治チャンネルであるとするならば，Bチャンネルで流れたニュースは「森首相，訪米意思をブッシュに生電話」という内容である．森首相が生電話したい相手は，大統領選というメイントピックの結果が出ないと解らない．よって，このニュースは t_{topic} の後ではなく，延期後の t'_{topic} よりも時間的後に配信される．

3.3 待機通知

図7の様に，同一のFAから生まれた3つの続報予定からなるFCがある．これらは，あらかじめ時系列な関連あるトピックであり，それらのイベント時間をそれぞれ t_k,t_{k+1},t_{k+2} とする．この状態において，一連のイベントのうちの一つのイベント時間が

$$t_k \longrightarrow t'_k \qquad (3)$$

のように延期される場合を考える．

図7は，e_1 の発生時刻が，t_1 から t'_1 に延期した場合をあらわしている．この場合，e_2 の続報予定時刻である t_2 は t'_1 よりも時間的順序は先であるが，e_1 の影響を受けているかもしれないニュース記事であるから，C_2 で流れるニュースは内容の正確さに問題があると考えられる．そ

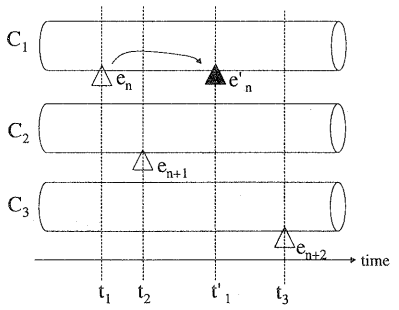

図 7: イベント予定の延期

こで，図8の様に，e_2 関係のニュースの配信を e_1 に関する正確なニュース記事が流れるであろう t'_1 まで見合わせることにする．また，t_1 で待機通知 a_{notify} をそのFAに C_2 からの新着記事として配信する．

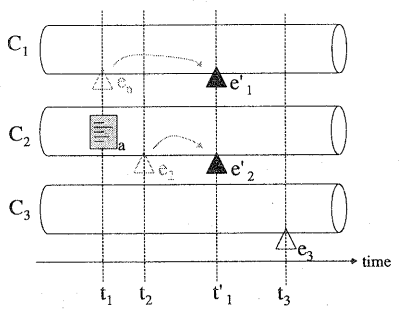

図 8: 待機通知

4 ニュースの価値

- 確信度

  記事の内容がどれだけ信用できるかを示す尺度

- スクープ度

  新鮮さとインパクトを示す尺度

- 展開度

以前，ニュースの価値を判断する尺度として，確信度とスクープ度の2つを定義した．本論文で

は，それに加えて**展開度**なる尺度を提案する．

4.1 展開度

ニュース記事の内容だけではそのトピックに関するFAが発展するのか，あるいは，衰退するのかの判断は困難である．しかし，現在よりも未来の情報である続報予定を利用することによって，トピックの今後の盛り上がり具合を判断することができると考える．ここで我々は展開度をいう尺度を次のように定義する．

$$Expand(a_n) = w_1 \frac{1}{D} \sum_{i=1}^{n} \frac{1}{|x_i|} + w_2 \sum_{j=1}^{n} \frac{D}{y_j} \quad (4)$$

ここで，各変数は図9に示すように，w_1, w_2 は重み，D は対象記事と同一チャンネルにある一番近い未来の続報予定との時間差，$|x_i|$ は対象チャンネルでの続報予定間の時間距離，$|y_j|$ は他チャンネルでの各続報予定までの時間距離とする．

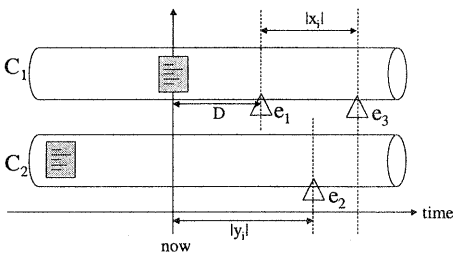

図 9: 展開度

従って，以下のような性質をもつ．

- 続報予定の数が多いほど展開度は大きい
- 続報予定時刻に近づけば展開度は大きくなる
- 続報予定が集中していればしているほど大きい

これにより，続報予定を用いることによって，あるトピックについてのFAの今後の盛り上がり具合を知ることができる．

例えば，続報予定が全くない場合よりも，「2000年12月31日，アントニオ猪木が21世紀に向けてカウントダウン．1,2,3ダー！ in 大阪城」というような元ニュース記事から続報予定が生まれているトピックの方が大晦日に近づくにつれて，どんな選手が登場するかというニュースが配信されることが期待できる．

5 プロトタイプシステム

時系列クラスタリングによってできたFAのうち，情報レベルの統合値が閾値を越えた記事をFAの情報と共に配信する．呈示イメージとしては，トピック一覧 (FA) 表示ウィンドウとFA情報とニュース内容を表示するウィンドウの2部構成となる．確信度の高いニュース記事は強調して，スクープ度が高いニュース記事は配信順序を早くする．ニュース記事はMacromediaのGENERATOR[6]を使用して，ティッカーやポップアップテキストなどのFlashコンテンツに変換されてJavascriptを含んだHTML文書としてWebサーバから配信される．

現在，ニュース配信サーバとして，gooのホットチャンネルとyahoo!ニュースを用いる[2]．これらのニュース配信サーバがもつチャンネルの内，gooのホットチャンネルからは朝日新聞提供の社会・経済・国際・政治を，Yahoo!ニュースからは時事通信・毎日新聞・NNA・ロイター通信提供の国内・経済・政治・国際などを対象とする．また，ニュース収集サーバはSunのUltra10 (Solaris7) 上で，perlを用いて実装している．ニュースの重要度計算部分とニュース呈示部分は現在開発中である．

6 関連研究

馬ら[7]は，オンラインニュースをユーザプロファイルと併用して，新鮮度・流行度・緊急度の3つの時系列的特徴量でフィルタリングしている．本研究との相違点は，トピックが現在時間より過去にあるニュースしか想定していないことである．

[2] これらは厳密には周期的プル型と分類されるが，一般にはプッシュ型サービスと考えることができる．

宗像ら [8] は，周期的に発生するデータ系列から，データの鮮度と同期度に基づいてデータの最適な組み合わせを選択する手法を提案している．本研究の着眼点である，発生が予定されているイベントという点で類似点が見られるが，本研究が対象としているニュースは完全に周期化されたタイミングで発生しないので，対象が異なる．

7 おわりに

本稿では，インターネットにおけるオンラインニュースのマルチチャネル型放送配信システムをについて述べた．本システムでは，オンラインニュースの内容情報と時系列文書としての時間情報からニュースとして重要度を計算する方式を提案した．

今後の課題としては，提案している手法を評価するためにプロトタイプシステムを完成し，評価実験をおこなう．

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)，および文部省科学研究費「マルチメディアコンテンツの放送型配信に関する研究」(課題番号 12780308) 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, Vol. 40, No. SIG8(TOD4), pp. 141–157, 1999.

[2] infogate. http://www.infogate.com/.

[3] ホットチャンネル． *http://channel.goo.ne.jp/*, NTT-X.

[4] Yahoo! JAPAN NEWS. *http://news.yahoo.co.jp/headlines/top/*, Yahoo Japan Corporation.

[5] NEWS CENTER. http://www.lycos.co.jp/news/. Lycos Japan Inc.

[6] GENERATOR. http://www.macromedia.com/jp/software/generator/.

[7] 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．Vol. 41, No. SIG6(TOD7), pp. 46–57, 2000.

[8] 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, Vol. 41, No. SIG1(TOD5), pp. 140–153, 2000.

[9] Yiming Yang, Tom Pierce, and Jaime Carbonell. A study in retrospective and on-line event detection. In *Proc. of ACM SIGIR98.*

[10] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *ACM SIGIR98.*