T-Scroll：時間的トピックの推移をとらえる可視化システム

長谷川幹根† 石川 佳治††

† 名古屋大学 工学部 電気・電子情報工学科 情報工学コース
†† 名古屋大学 情報連携基盤センター 〒 464–8601 名古屋市千種区不老町
E-mail: †hasegawa@dl.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

あらまし インターネット上では，ニュースなどの大量のテキストデータの配信が日々なされている．本論文では，このようなテキストデータにおける，時間的なトピックの推移をとらえるための可視化システム T-Scroll について述べる．本システムは，下位の時系列的な文書クラスタリングシステムのクラスタリング結果をもとに，クラスタの関連を巻き物（scroll）状に提示する．本論文では，システムのアイデア，機能，実現手法等について述べる．
キーワード 時系列文書，クラスタリング，可視化，インタフェース

T-Scroll: A Visualization System for Temporally Changing Topics

Mikine HASEGAWA† and Yoshiharu ISHIKAWA††

† Department of Information Engineering, School of Engineering, Nagoya University
†† Information Technology Center, Nagoya University Furo-cho, Chikusa-ku, Nagoya, 464–8601 Japan
E-mail: †hasegawa@dl.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

Abstract On the Internet, delivery of a large amount of documents such as news articles is continually performed everyday. In this paper, we describe an information visualization system *T-Scroll* to show the transition of topics contaned in such documents to the user and to provide an overview of their trends. The system is built on a clustering system for time-sereis of documents and presents relationships between clusters like a scroll. This paper describes the idea, the functions, and the implementation of the system.
Key words time-series documents, clustering, visualization, interface

1. ま え が き

インターネット上の情報提供・配信サービスの進展により，今日では，ネットワークを介したニュース配信が盛んに行われている．それに伴い，大量の情報を要約しフィルタリングするための，オンラインテキスト情報処理の重要性がさらに増してきており [1]，時々刻々と配信される時系列的な文書データに適した情報の要約と提示に関する新たな技術の開発が求められている．

このような背景を受け，本研究では，一般のユーザが大量のニュースのトピックの大まかな推移を容易に把握できるようにするためのユーザインタフェースである **T-Scroll**（Topic/Trend-Scroll）システムの開発を行っている．T-Scroll は文書クラスタリングシステムの上位に位置し，その出力を利用して，クラスタリングされた結果を可視化してユーザに提示する．その特徴は，各時点で得られたクラスタをラベルを付与して時間軸上に配置し，クラスタ間の関連性を表すリンクを示すことで，トピックの流れを表す点にある．画面上にクラスタリングの結果を巻き物上に表示することから，システムを T-Scroll と呼んでいる．あるトピックに興味をもったユーザは，対話的な操作により，必要に応じてより詳細な情報を得ることが可能となる．T-Scroll のアイデアは，[5] においてその概念が提案された ClusterFlow システムの考え方に基づいているが，具体的な実装を行い，さまざまな機能を追加し，評価を行った点において異なっている．

以下ではまず，2. において，T-Scroll システムの基盤となる，新規性に基づく時系列文書のクラスタリング手法について説明を行う．次いで 3. では，T-Scroll システムの機能およびその設計の概要について述べる．4. では実装したシステムをもとにその機能について説明する．5. では実装方式について述べ，6. では評価結果を示す．最後に 8. でまとめと今後の課題について述べる．

2. 新規性に基づく時系列文書のクラスタリング

本研究が基礎とするのは，[4], [6], [7] において提案されている，新規性に基づく文書クラスタリング手法である．その特徴は以下の 3 点である．

（1） 類似度計算において，文書の内容の類似度だけでなく

文書の新規性も考慮することで，新規性の高い文書により着目したクラスタリング結果を導出する．ポイントとなるのは，文献書誌学で用いられる**老化**（aging）の概念を取り込んだことにあり，若い文書（入手されて間もない文書）ほど，クラスタリング結果に与える影響が高くなるようにしている．

（2） 新たに文書の追加の際にはインクリメンタルな更新処理を行い，更新コストを削減している．クラスタリングのアルゴリズム自体は ***k*-means 法**に基づき，それを拡張することでインクリメンタルな処理を実現している．

（3） 上述のように本手法では老化の概念を導入しており，文書が古くなると，他のどの文書とも類似しなくなり，外れ値（outlier）となる．外れ値がある場合にクラスタリング結果を悪化させないための処理が工夫されている．また，十分古くなった文書は，寿命に達したとされ，自動的にクラスタリングの対象から削除される．

このようなアプローチにより，新規なトピックに重点をおいて，時系列的に配信されてくる文書データを定期的にクラスタリングを行い，ユーザが最新の状況を把握することを容易にする．

[4], [6], [7] で用いられた影響力の逓減モデルでは，文書の価値（重み）が時間の経過にしたがって指数的に逓減していくと想定し，文書 d_i に対する文書の重みを以下のように与える．

$$dw_i = \lambda^{\tau - T_i} \qquad (0 < \lambda < 1) \tag{1}$$

ただし，τ は現在の時刻を表し，T_i は文書 d_i が入手された時刻を表す．λ は文書の影響力の逓減の度合いを表すパラメタである．一方，n 個の文書からなる文書集合 $d_1, \ldots, d_n$ の文書の重みの総和を

$$tdw = \sum_{l=1}^{n} dw_l \tag{2}$$

で与え，文書 d_i の文書集合中での生起確率を

$$\Pr(d_i) = \frac{dw_i}{tdw} \tag{3}$$

という主観確率で定める．この確率は，古い文書ほど値が小さくなり，古い文書を考慮の対象から外す（忘却する）というアイデアを表現している．

文書の類似度は，上記の式や他の仮定をもとに確率的なモデリングに基づいて導出される [4], [6], [7]．その一般形は

$$sim(d_i, d_j) = \Pr(d_i)\Pr(d_j)\frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{len_i \times len_j} \tag{4}$$

であり，文書ベクトルの内積を文書長の積で割ったものに各文書の生起確率を掛けたものとなる．よって，この文書類似度は，単に文書どうしが類似しているかどうかだけでなく，各文書がどの程度古いかも考慮し，十分古くなった文書は他のどの文書にも類似しなくなるという性質を有している．このような類似度をクラスタリングに用いることにより，文書の新規性を重視したクラスタリングの実現を図っている．

以上のアプローチに基づき，このクラスタリング手法では，追加の文書集合が与えられると線形時間で統計情報の更新と再クラスタリング処理を行い，最新のクラスタリング結果を出力する．各時点のクラスタリング結果はその時点のトピックの情報を表しており，それらを保持しておくことで後の分析に役立てることができる．このアイデアに基づき，視覚的な表現による分析用インタフェースとして現在開発を行ったのが，以下で述べる T-Scroll システムである．

3. T-Scroll システムの概要

3.1 システムの特徴

本研究で開発を進めている **T-Scroll** (Topic/Trend-Scroll) システムの特徴は，主として以下のようになる．

（1） 継続的なクラスタリングにより得られた各時点のクラスタリング結果を時間軸上にトピックを表すラベルとともに表示することで，各時点における主要なトピックを把握可能とする．ニュース記事などのトピックやトレンドの流れが巻き物のように表示されることから，本システムを T-Scroll と呼んでいる．

（2） 興味のあるクラスタを選択することで，より詳細な情報（関連キーワードのリスト）や元記事を対話的に参照することが可能である．

（3） ある時点で得られたクラスタ集合に対し，一つ前の時点で得られたクラスタ集合から，関連度の強さに応じてリンクを張ることで，隣接する時刻におけるクラスタ間の関連の把握を容易にする．

（4） ユーザインタフェース上に表示する時間軸の刻み幅をユーザの指定により調整可能とすることで，要求に合わせた詳細度で分析が行える．特に，時間軸の刻み幅を広くとり，トレンドを大まかにとらえる粗視化の機能が重要であり，これは，OLAP (On-Line Analytical Processing) におけるロールアップ（roll-up）の機能に対応づけることができる．

3.2 システムの概要

図 1 に，T-Scroll システムのインタフェースの概念図を示す．図は，10 月 1 日から 1 週間刻みで 10 月 15 日までのクラスタの流れを表示している様子を示している．例としては，国内のニュース記事を逐次クラスタリングする例を想定しているが，この図はあくまで例示のためであって，実際の表示例ではない．インタフェース上では左から右に時間が流れており，画面下部のスライドバーにより，前後の時点に移動することも可能である．画面上で同じ縦の点線上にある楕円は同じ時点で得られた

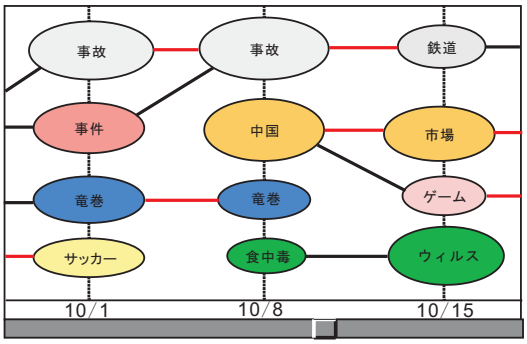

図 1　T-Scroll システムの概念

クラスタの集合を表している．

クラスタ上のラベルは，クラスタ中の文書に含まれる語で，スコアが最大のものを選択して表示する．いくつかのスコア付けを比較した結果，現在の実装では，クラスタ C_p における語 t_j のスコアを

$$score(t_j) = \sum_{d_i \in C_p} \Pr(d_i) tf_{ij} \qquad (5)$$

で求めている．つまり，クラスタ内の各文書について，語 t_j についての語頻度（term frequency） tf_{ij} を，その文書の重み $\Pr(d_i)$ と掛け合わせ，その総和をとっている．idf も含めたスコア計算法も試したが，上記方式の方が，ラベルとしてふさわしい一般性のある語が比較的得られるという経験が得られている．また，クラスタ上に複数の単語（たとえばスコアが上位 3 件の語）を並べて提示することも考えられるが，実システムで検討したところ，画面表示が煩雑になるため 1 語だけを選んでいる．詳しくは後述する．

クラスタ上に書かれた楕円の面積はクラスタに含まれる文書の数の量に対応しており，トピックの規模を示している．図で示されるように，一部のクラスタ間には左から右にリンクが張られている．これはクラスタ間の関連性の深さを示している．クラスタ間の関連度は

$$csim(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i|} \qquad (6)$$

という式により定義する．この式は，クラスタ C_i に含まれる文書がクラスタ C_j にどれだけ含まれているかを調べることにより関連性の深さを測っている．1 つのクラスタから 0 個以上のリンクが出ることを許し，トピックの消滅（0 個のリンクで表現）や分岐（複数個のリンクで表現）を表す．

また T-Scroll では，OLAP などでしばしば用いられるドリルダウン／ロールアップ機能をサポートする．分析する時間間隔を狭める場合（例：1 日単位）がドリルダウン，広げる場合がロールアップに相当する．

4. 実装システムの機能

T-Scroll の実装システムにおける各種機能について説明する．

4.1 インタフェース画面

T-Scroll を利用するユーザは，表示対象の期間をインタフェース上で指定する．たとえば半年分のクラスタリング結果がシステムに保持されている場合でも，ユーザが興味がある期間が「3ヶ月前から 1ヶ月前」という場合もありうる．期間を指定する機能を用いることで，処理がより軽量になるという利点もある．対象期間を指定した後，表示ボタンをクリックすることで，図 2 のようなインタフェース画面が得られる．

図 2 では，10 月 1 日から 1 週間刻みで 12 月 31 日までの時間的トピックの推移を表示した例を示している。縦軸は時間軸であり，楕円はそれぞれのクラスタを表しており，それぞれ 20 個ずつにクラスタリングされている．なお，現在の T-Scroll では，クラスタ数を動的に変更する機能はいまのところない．これは，下位のクラスタリングモジュールが，継続的に k-means

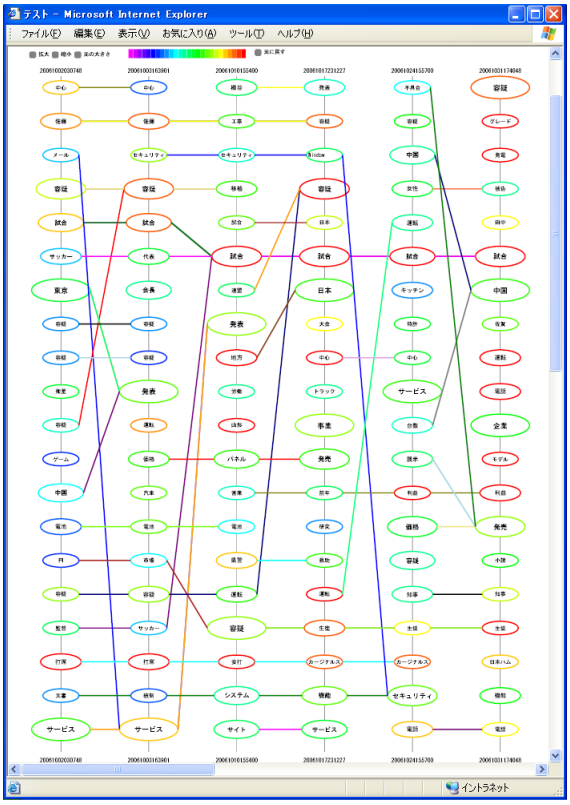

図 2　T-Scroll のインタフェース画面

法に基づくクラスタリングを行うためであり，k はその初期化段階において固定する必要があるためである．ただし，複数のクラスタ数の設定（例：$k = 10, 20, 30$）で複数のクラスタリング処理を並行して動作させることは可能であることから，複数個のクラスタ数の選択肢をユーザに提示し，選ばれた k の値に応じて表示を替えることは可能である．今後のシステム拡張の課題としたい．

楕円の大きさについては，クラスタに含まれる文書数に応じて数レベルのサイズの中から表示の大きさを選択する方式を採用している．これにより，ユーザは大まかなクラスタの大きさを知ることができ，クラスタリング結果における文書数の分布を大まかに捉えることが可能となる．ただし，クラスタのサイズが大きいことはホットなトピックの文書が集中して一見良いように思えるが，実際のクラスタリング結果を見ると必ずしもそうではなく，小さいクラスタの方がより互いのトピックが密接に関係している場合が多く見られる．これは，用いているクラスタリング手法 [6], [7] の特性による．一般に，ホットなトピックはある程度の数の文書があればクラスタを構成し，そのサイズは小規模から中規模である．一方，残りの文書（ホットでないトピックの文書や他と比較的類似していないトピックの文書）は，比較的大きなサイズのクラスタに吸収される．このような大きなサイズのクラスタは，平均的なトピックに対応すると考えられ，特定のトピックにはあまり対応せず，トレンドを把握するという意味ではあまり有効ではない．

クラスタのサイズだけでなく，クラスタの質の良さも容易に把握できるようにするため，T-Scroll ではクラスタの質の高さ

を色分けして表示する．具体的には，楕円の輪郭の線の色により，クラスタの質の良さを表現する．可視光線のスペクトル分解を参考にし，赤に近いほどクラスタの質が高く，紫に近いほどクラスタの質が低いことを意味する．クラスタ C について，その品質のスコアは

$$quality(C) = |C| \cdot avg\_sim(C) \quad (7)$$

$$avg\_sim(C) = \frac{1}{|C|(|C|-1)} \sum_{d_i,d_j \in C, d_i \neq d_j} sim(d_i, d_j) \quad (8)$$

と与えられる [7]．ここで $|C|$ はクラスタ C 中の文書数を表し，$avg\_sim(C)$ はクラスタ内の文書の平均類似度を表している．すなわち，$quality(C)$ は，文書数が多いだけでなく，クラスタ内の文書が互いに似ている場合に大きい値をとるようなスコアとなっている．[7] のクラスタリング処理では，クラスタリングの結果生じるクラスタ集合において，それらの品質の総和が最大となることを目標としてクラスタリングを行うことから，[7] のクラスタリング手法におけるクラスタ品質の考え方をそのまま導入しているといえる．

楕円の中には，そのクラスタを最も適切に表すような単語（実際には形態素）を選んで表示している．そのクラスタについて式 (5) に基づいて各語のスコアを計算し，その値が最大のものを選んで提示している．

3 節で述べたように，クラスタ間のリンクは，クラスタ間の関連度が大きいことを表し，ある閾値以上の関連度についてリンクを作成している．リンクの色については，関連度を反映する方式も検討したが，見やすさを考慮し，一つ前の時点におけるリンクの配色と整合性の高い色を次の時点に採用するようにしている．リンクの線の太さを関連度に応じて変更することも検討したが，SVG による実装（後述）では，表示の段階で，指定した太さに必ずしも表示されないなどの問題があり，現在はすべて同じ太さで表示している．

4.2 クラスタの詳細情報

クラスタに関しては，ユーザがさらに詳細を調べることができるための機能を実現している．

図 2 のように，クラスタに対するラベルとして 1 つのキーワードを与えるだけでは，クラスタ内容を判断するのが困難な場合もある．そこで本システムでは，クラスタの内容を容易にブラウズできる機能も提供している．クラスタ上（楕円上）にマウスカーソルが乗ると，そのクラスタに関連の深い複数のキーワードが表示される．実行した様子を図 3 に示す．クラスタ内の単語のうち，スコアが上位 20 位のものを順に表示している．

上記のようなキーワード表示機能によってクラスタの内容はわかるが，実際にクラスタに含まれる文書はわからない．よって，本システムでは更に，クラスタの上をクリックすることでクラスタに含まれる文書を表示する機能も実現している．実行の様子を図 4 に示す．図 4 では，クラスタに含まれる文書のうち発行日時が新しいもの上位 10 位のもののタイトルを表示している．文書の内容はタイトルをクリックすることによって表示される．また，詳細情報をクリックすることにより，クラスタに含まれるすべての文書を表示する機能も実装している．

図 3　クラスタのキーワードリストの表示

図 4　クラスタ内の文書の表示

5. システムの実装

本節では T-Scroll システムの実装について述べる．

5.1 システムの構成

本システムは以下の図 5 のような構成をしている．本システムは，新規性に基づく時系列文書のクラスタリングのプログラム [6], [7] と連携し，その出力を利用する形で構築している．文書クラスタリングのプログラムは Ruby 言語で書かれており，各時点で取得された新たな文書集合をバッチ的に与えることで，その時点の最新のクラスタリング結果を出力する．クラスタリングの結果は XML 形式のファイルとして出力される．

対象とする文書集合に応じて，文書のフォーマットやクラスタリングに用いる語の抽出法は異なりうる．また，式 (1) で示した文書の重みの逓減パラメータ λ の値をどのように設定するかも，対象に依存する．さまざまな入力文書の設定に対応するため，文書クラスタリングのプログラムでは，入力文書を取得して特徴としての語のリストを抽出するモジュールについて拡張性を与えている（具体的には，デザインパターンにおける Factory Method パターン [2] が用いられている）．そこで，以下に述べるような拡張を行っている．

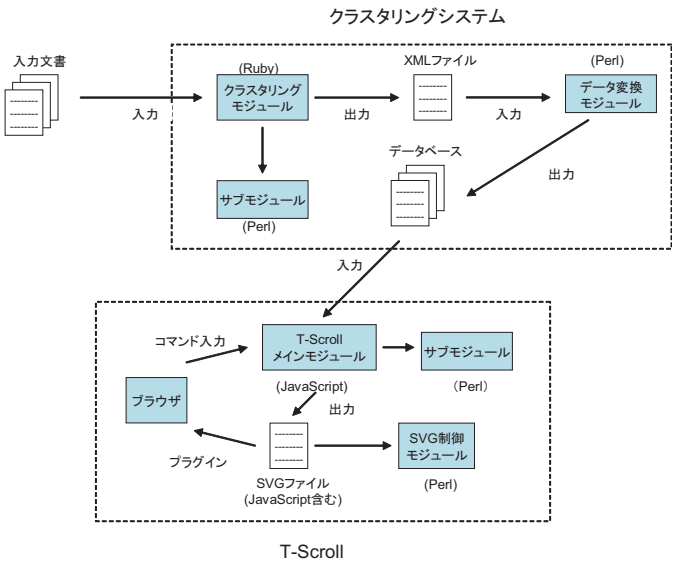

図 5　システム構成図

5.2　実験対象の情報源

本実験において対象とした情報源は，RSS データを提供しているニュースサイトである nikkeibp.net, asahi.com, sportsnavi.com（サッカー・野球）の 4 つのサイトである．情報収集は 2 時間おきに行っている．それぞれの RSS サイトにアクセスし，前回情報収集した時から更新された情報について，リンク先などの必要な情報を取得する．次いで，取得したリンク先情報をもとに，サイトにアクセスしウェブページから記事の本文を抽出する．抽出方式は以下のとおりである．

（1） RSS データから取得したリンク先のウェブページの内容は，同一サイトについては本文を除けば大部分が同じ内容である可能性が高いので，前回アクセスしたウェブページと diff コマンドによって比較し，一致した行は削除する．

（2） ウェブページのうち本文が書かれているのは TD・SPAN・DIV タグの中に限られると考えられるのでそれらのタグに囲まれている部分を抜き出し本文候補に加える．また，抜き出した部分の中に更に TD・SPAN・DIV が含まれていたらそのタグの中身を取り出し新しい本文候補に加え，元の本文候補から抜き出した部分を削除する．

（3） それぞれの本文候補の中の残っているタグや改行などは削除し文章だけにする。

（4） 抽出された本文候補のうち最も文字数が多いものを本文とする．

入手したテキスト情報を形態素分析ソフトである茶筌により形態素分析し，分析結果が「未知語（記号除く）」,「名詞–一般」,「名詞–固有名詞」,「名詞–サ変接続」,「名詞–副詞可能」,「名詞–形容動詞語幹」になった形態素を語としてクラスタリングのための入力情報として用いる．茶筌は半角のアルファベットに完全には対応しておらず,「MLB」などの連続したアルファベットは 1 単語として扱われず，アルファベットの M と L と B に分割されてしまう．そこで，連続したアルファベットが続いた場合は 1 単語と見なし未知語としている．なお，全角の連続したアルファベットに対しては未知語として単語認識される．

実装においては，未知語もクラスタリングのための語として抽出している．この理由は，茶筌の辞書に登録されていない最新の語である可能性がある場合があることと，茶筌では半角全角共に連続したアルファベットは未知語と見なされるためである．未知語と判断された語が，抽出ミスなどによるノイズである場合も考えられるが，そのようなノイズ語については，クラスタリングの際の重み付け（$tf \cdot idf$ 方式に似た方式が用いられる [6], [7]）により，その悪影響を軽減できると考えられる．

今回は，先に述べたような 4 つのサイトの記事を対象としたが，同様な処理により他の情報源を対象とすることも容易である．

5.3　T-Scroll メインモジュールの実装

T-Scroll は，新規性に基づく時系列文書のクラスタリングのプログラム [6], [7] によって出力された XML 形式のファイルを入力とする．この XML ファイルには，ある時点におけるクラスタリングの結果であるクラスタの内容が納められている．各クラスタごとに，クラスタの品質を表すスコア，クラスタを代表となる語のリスト，クラスタの要素である文書のリスト（クラスタ代表との類似度値によりランク付けされている）などが含まれている．クラスタリングモジュールは定期的にクラスタリングを行い，最新のクラスタリング結果を出力するため，このような XML ファイルが多数存在することになる．T-Scroll は，ユーザから指定された対象の期間に応じて，必要な XML ファイルを適宜読み込んで利用する．

T-Scroll のメインモジュールは JavaScript で記述されており，Web ブラウザ内に読み込まれ動作する．ユーザインターフェースに関する一部の処理は JavaScript および AJAX の機能を用いて実現している．

ユーザから対象の期間や分析の時間間隔の入力を受けた後でインタフェース画面を表示するが，そのためには，メインモジュールから Perl で作成されたサブモジュールを呼び出すことになる．実際にはこのサブモジュールがクラスタリング結果の XML ファイルを読み込み，ユーザの指定に応じて内容を解析し，インタフェース画面に表示するための SVG 形式のファイルを作成する．作成された SVG ファイルはブラウザに即座に読み込まれ，図 2 に示したインタフェース画面が表示される．SVG ファイル中には JavaScript のコードが埋め込まれており，その中から必要に応じて Perl により記述されたモジュールが実行される．このような仕組みにより，先に述べたシステムの機能を実装している．

6.　システムの評価

6.1　システム利用による評価

まず，実際にシステムを利用した筆者により得られた知見を報告する．今回は，5.2 節で述べた 4 つのサイトからの記事を対象としており，1 日あたり平均しておよそ 100 件のニュース記事が取得されている．設定により，各時点において 20 点のクラスタが作成され表示されている．表示の対象とする期間については，長期（例：3ヶ月以上）に設定することはあまり有効

とはいえなかった．トピックの推移は 1〜2ヶ月程度ぐらいの範囲でとらえる方が分かりやすいという点と，長期の場合には表示が煩雑になり，また，インタフェースの動作が重くなるためである．

時間間隔の設定については，1 日刻みで表示した場合には比較的単調な表示となる．その様子を図 6 に示す．この図は，12 月 1 日から 1 日刻みで 12 月 10 日までトピックの推移を表示している．利用した印象としては表示が冗長であるという感触を得た．これは，1 日程度では大きなトピックの変化がないためである．一方，1 週間刻みで表示した場合 (図 2 参照) には，トレンドを把握するという意味ではより適切な表示であると感じられた．インタフェースの表示においても，クラスタ間のリンクの交差などが見られ，視覚的には面白いものとなっている．ただし，たまにリンクが張られている隣接するクラスタでトピックがずれていること，すなわちトピックドリフトが見られた．

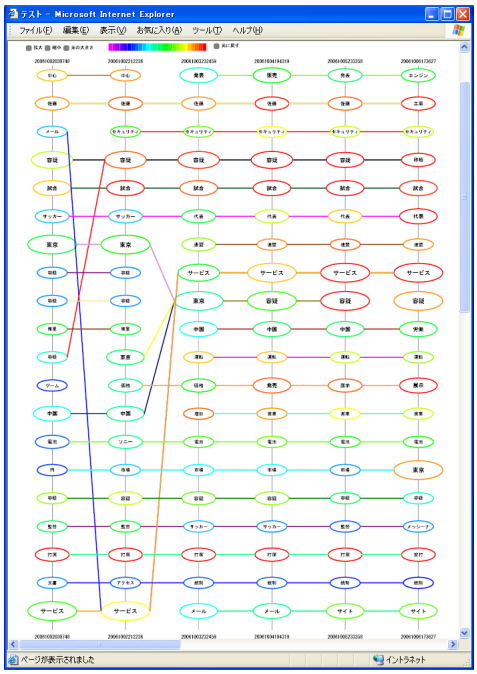

図 6　T-Scroll 全体図 (1 日刻み)

本システムで 2006 年度の後半に取得した文書データをもとにした表示について，観測できた結果を以下にまとめる．

- 事件・事故に関するクラスタが継続的に存在する：毎日のニュース記事には必ずといっていいほど事件・事故のニュースが含まれる．そのため，継続的にこの種のトピックに関するクラスタが見られる．T-Scroll のインタフェース上では適切にクラスタ間のリンクが表示されており，関連がうまく表現されていることがわかった．クラスタに付与されるラベルとしては「容疑」などが一般的であった．
- 鉄道に関する事故や遅延情報などは，それだけで別に単一のクラスタを構成し，継続的に存在する：鉄道関係の記事は，特有の単語の出現パターンを示しており，比較的継続的に出現するためであると考えられる．クラスタに付与されるラベルとしては「運転」などが一般的であった．
- サッカーに関する記事が，継続的に一つのクラスタを構成する：今回，sprtsnavi.com から野球とサッカーに関するニュース記事を取得したが，サッカーの記事は数も多く，継続的に記事が見られるため，サッカー記事のみのクラスタがどの時点でも見られた．クラスタ内の文書はサッカー記事のみから構成され，非常に質のよいクラスタである場合が一般的であった．クラスタのラベルとしては「ロ」,「ナウ」など，あまり適切でないラベルが選ばれる場合が多くみられた．これは,「ロナウジーニョ」のような有名選手の名前が茶筅の辞書に登録されていないためである．一方，野球については取得される記事数が比較的少ないことから，固有のクラスタとなる場合はあまり見られなかった．
- コンピュータ・IT 関係に関するクラスタも継続的に見られた．ただし，ゲーム関係のクラスタなどに分岐したり，統合したりする場合が見られた．ラベルとしては「セキュリティ」などが一般的であった．
- 経済関係の記事についてのクラスタも継続的に存在する．ラベルとしては経済を表す「市場」なども見られるが,「中国」が選ばれることも多くみられた．経済関係の記事には中国に関する記事が多く含まれるということを反映している．

以上は，今回の対象であるニュース記事のサイトの内容を考えると，継続的に出現することが妥当であるクラスタであるといえる．一方，今回対象とした期間に含まれるホットなトピックを表すような，以下のような観測も得られた．

- 知事と汚職に関するクラスタが 9 月頃から継続して存在した．2006 年の後半に汚職関係の記事が継続して出ていたことが反映されている．図 7 に知事汚職に関するクラスタのトレンドを示す．この図は，10 月 1 日から 1ヶ月刻みで 12 月 31 日までを表示しており，クラスタ内が青くなっているところが知事汚職に関するクラスタである．

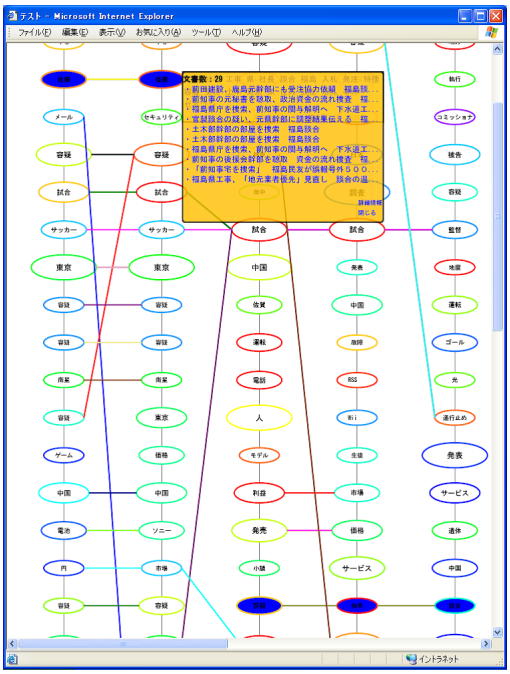

図 7　知事汚職に関するクラスタの流れ

- いじめや教育に関する問題も 2006 年の後半には重要なトピックとなったが，これもクラスタの流れに反映されている．10 月中旬〜11 月中旬頃の時期において，クラスタの連鎖として表現されている．
- 2006 年後半には，PS3 と Wii のゲーム機の発売に関連するクラスタも見られた．興味深いことに，当初は PS3 を表す「PS」というラベルが表示されることが多かったが，Wii の発売直前から「Wii」などのラベルが見られるようになり，トレンドの流れをうまく表現しているといえる．
- 2006 年 12 月になると，12 月が最も火事が多い時期ということを反映して，火事に関するクラスタの流れが出現した．
- 2006 年は，ノロウイルスの大流行に伴い，12 月を過ぎたころからノロウイルスに関するラベルのクラスタの流れが出現した．

その他クラスタの流れをみていて気づいた点として以下のものがあった．

- 地震や台風などの自然災害の記事を含むクラスタは，ホットなトピックであり，台風なら台風関連の記事だけで構成されていた．
- 時期が過ぎた話題のクラスタはラベルが「容疑」のものに流れていく傾向がみられた．また，これにともないクラスタのラベルが「容疑」のものは文書数が多い傾向がみられた．
- キーワードに「中国」を含むクラスタは毎週のように存在するが隣接するクラスタとほとんどリンクされていなかった．
- 10 月には北朝鮮の核実験などがありニュースで注目されていたが，今回の実験結果には，キーワードが「北朝鮮」,「核」となるクラスタはあまりなかった.「政府」「内閣」などのキーワードが共通することから，他の政治関係の記事などと一緒になって大きなクラスタを構成していたと考えられる．

6.2 クラスタのトレンド評価

本節では，実際に起きた出来事の流れとクラスタのトレンドを比較し，T-Scroll のクラスタのトレンドの正確性を評価する．

評価にあたり，各クラスタの内容判断は，クラスタに含まれる文書のうち発行日時が新しいもの上位 10 件までを対象とし，上位 10 件までに対象とする出来事に対する記事がどれくらいの割合で含まれているかを評価の値 (**トレンド値**と呼ぶ) として用い，クラスタのトレンドとしてグラフに表し評価する．

2006 年 10 月 1 日から 2006 年 12 月 31 日まで (10 月 21 日，22 日を除く) の様々な出来事に対するクラスタのトレンドの評価を行ったが，ここでは例として「知事談合」に関するクラスタのトレンドの評価を示す.「知事談合」に関する主要な出来事は，以下のようになっている．

9 月 27 日頃：福島県知事談合問題発生
10 月 7 日頃：和歌山県知事談合問題発生
10 月 23 日：福島県知事逮捕
11 月 15 日：和歌山県知事逮捕
11 月 16 日頃：宮崎県知事談合事件発生
12 月 8 日：宮崎県知事逮捕

図 8 に「知事談合」に関するクラスタのトレンドを示す．10 月 1 日から 12 月 30 日まで 3 日ごとのクラスタのトレンドを示している．図 8 と「知事談合」に関する主要な出来事を比較した結果を以下に述べる．

- 知事談合に関する記事はほとんど 1 つのクラスタに集中しているので，どれか 1 つの談合のトレンド値が高くなると他の談合に関するトレンド値は落ちてしまっている．
- 福島県知事の談合問題の事件については，発覚が 9 月 27 日と今回の対象期間より早いため，10 月に入ったころには事件の進展情報もなく徐々に下がっていると思われる．また，事件が進展した福島県知事が逮捕された 10 月 23 日以降から急激にクラスタのトレンド値が上昇し，しばらくすると一気に下降している．
- 和歌山県知事の談合問題は 10 月 7 日頃から発生し，それに合わせるにトレンド値が徐々に増加し，10 月 23 日に和歌山県知事が逮捕されるとしばらくしてトレンド値が急激に落ちている．
- 宮崎県知事の談合問題は 11 月 16 日頃から発生したが，先に福島，和歌山で談合事件がおき，他の地域でも談合の疑いがあったためそれらに関する記事に邪魔される形で，初めはあまりトレンド値が高くないが，12 月 8 日に宮崎県知事逮捕されるとトレンド値は急激に高くなっている．逮捕後も，そのまんま東氏が知事選に出馬するなどの出来事があったためトレンド値は高い値を維持し続けている．

以上の結果より,「知事談合」に関するクラスタは実際の出来事の流れをかなり正確に表していると考えられる．

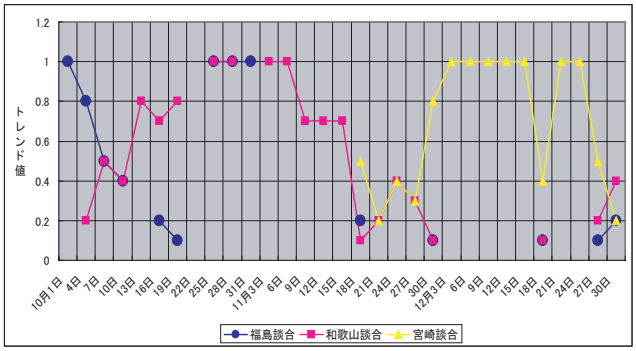

図 8 「知事談合」に関するトレンド

以上の 3ヶ月のクラスタのトレンドの評価より得た知見を以下に示す．

- スポーツや自然災害などのクラスタのトレンドは時期が去っても高いトレンド値を維持することが多い．これは，スポーツや自然災害などが他の種の記事とあまり類似度が高くないため，ある文書の重みが小さくなっても程度の記事が消滅するまで残ってしまうと考えられる．
- 海外の事件，事故に関するクラスタのトレンドが全くといっていいほど現れない．これは，今回利用したニュースサイトが海外に関する記事が少ないことが原因であったと思われる．
- 政治に関するクラスタのトレンドがほとんど現れない．これは，先に述べた通り今回利用したニュースサイトが政治に

関する記事が少なかったことと政治に関する記事は1つのクラスタに集まりやすいためだと推測される．

- 裁判の判決など事前に起こる時期が分かっている出来事は，発生よりも前から低いトレンド値でクラスタのトレンドが現れることが多い．また，地震や事件など先に予測しておくことが出来ない出来事は，急にクラスタのトレンドが現れることが多い．
- 多くのクラスタのトレンドが事件などの発生時期よりも遅れる．これは，ある1つのことに関するクラスタとして現れるためにはそれなりの記事の量が必要であるためであると考えれる．

よって，今回の T-Scroll のクラスタのトレンドは事件などの発生や時期が過ぎた後に正確でないトレンド値を記録することがあるが，最もホットな時期にはクラスタのトレンドの中で最高値を記録することがほとんどであるので，T-Scroll は大まかなトピックのトレンドをとらえるのには有効であると評価できる．また，海外や政治などのクラスタについては，それに関する記事を増やしたり，クラスタの数を増やすことによってクラスタのトレンドが現れやすくなると思われる．

7. 関連研究

時系列的に取得されるニュース記事などの文書データをもとに可視化を行うシステムはあまり見られない．以下では2つの関連するシステムについて紹介する．

ThemeRiver [3] は，トピックの流れを川に見立てて表示を行う可視化システムである．川が画面の左から右に流れるような表示を用いるが，これは左から右への時間の推移に対応している．川の中にいくつかの色分けされた流れが表示されており，これが一つ一つのトピック（テーマ）に対応している．画面上には，それぞれの流れがどのようなトピックに対応するかを示すためのフレーズが表示される．また，川の幅は各時点における記事の量を表している．トピックの流れを左右にスクロールするインタフェースで表現するという点では，ThemeRiver は T-Scroll と共通しているが，クラスタリングを用いているわけではない．視覚的なインパクトはあるものの，トピックの推移は表現できず，複数の時間間隔での表示なども可能でない．大まかなトレンドの把握には利用可能であるが，実際に時系列的な文書データを分析的にブラウズするには，必ずしも強力なツールではない．

Swan と Allan [8] は，トピックを表現する timeline を表示するインタフェースである．指定された期間における時系列的な文書を分析して，継続して出現するトピックを検出し，画面上に時区間を表す棒状の表示（timeline）を提示する．また，timeline には併せてキーワードが表示される．検出されたトピックごとに timeline が提示されるため，ユーザは画面を眺めることでトピックがどの期間に見られるかを把握できる．クラスタリングではなく，統計的指標を用いてトピックの検出を行っている．主要なトピックとその期間を提示する，に焦点を当てており，その点に関しては T-Scroll より優れている面もあるが，トピック（クラスタ）間の関連や，複数の時間間隔による分析機能はない．探索的なブラウジングにおいては T-Scroll の方がより豊富な機能を有しているといえる．

8. まとめと今後の課題

本論文では，時系列的な大量のオンライン文書のトピックの変遷・推移を対話的に分析するためのインタフェースである T-Scroll システムの特徴，機能，構成，そしてその評価について述べた．

今回開発した T-Scroll は，以下のような利点がある．

（1） 全ての機能においてユーザはボタンをマウスでクリックする程度の作業で操作することができ，パソコンに初心者なユーザでも簡単に利用できる．

（2） クラスタリング結果を時系列的に可視化し，クラスタの質を色で表現することによってある時期のトレンドが即座に理解できる．

（3） SVG を採用したことにより，ブラウザ上での対話的操作が行える．

今回，時系列文書のクラスタリング結果のトレンドを可視化するにあたって SVG を利用した．これは，SVG が XML ファイル形式で書くことができ，比較的簡単に作成できることと，JavaScript をファイル内に埋め込むことによりマウスのクリックなど簡単な操作で制御できるからである．

今後の課題としては，日本語以外の記事への対応，および，マルチユーザ環境への対応（クライアント-サーバアーキテクチャの採用など）が考えられる．

謝　辞

本研究の一部は，日本学術振興会科学研究費（16500048），柏森情報科学振興財団，セコム科学技術振興財団の助成による．

文　献

[1] J. Allan ed. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer, 2002.

[2] エリック・ガンマ他. オブジェクト指向における再利用のためのデザインパターン. ソフトバンククリエイティブ, 1999.

[3] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing thematic challenges in large document collections. *IEEE Trans. on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[4] Y. Ishikawa, Y. Chen, and H. Kitagawa. An on-line document clustering method based on forgetting factors. In *Proc. of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, Vol. 2163 of *LNCS*, pp. 332–339, 2001.

[5] 石川佳治, 梶並千春, 北川博之. Clusterflow: 時系列文書のトピック追跡のための視覚的インタフェース. 情報処理学会第64回全国大会, 2002. 2X-5.

[6] S. Khy, Y. Ishikawa, and H. Kitagawa. Novelty-based incremental document clustering for on-line documents. In *Proc. of International Workshop on Challenges in Web Information Retrieval and Integration (WIRI 2006)*, 2006.

[7] S. Khy, Y. Ishikawa, and H. Kitagawa. A novelty-based clustering method for on-line documents. *World Wide Web Journal*, 2007. (to appear).

[8] R. Swan and J. Allan. Automatic genration of overview timelines. In *Proc. ACM SIGIR*, pp. 49–56, 2000.

[9] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999.