T-Scroll：時系列文書のクラスタリングに基づくトレンド可視化システム

長谷川 幹根[†,☆] 石川 佳治[††]

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行われている．ここでは，ニュース記事のように，発行時刻などの情報をともない次々と配信される文書のことを時系列文書と呼ぶ．大量の時系列文書の中から，着目した時期における主要なトピックや，トピックの継続的なつながりにより出現するより大きなトレンドをとらえたいという要求がしばしば発生するが，一般には，文書データの閲覧・分析のために多大な時間と労力が必要となる．そこで本研究では，時系列文書集合中に含まれるトレンドをとらえるための可視化システムである T-Scroll（Trend/Topic-Scroll）の開発を行った．本システムは，下位の時系列文書を対象とするクラスタリングシステムが定期的に生成するクラスタリング結果をもとに，クラスタ間の関連を巻き物状に可視化して提示する．対話的な機能により，ユーザは文書集合中に含まれるトレンドや個々のトピックの詳細を把握することが可能となる．本論文では，システムのアイデア，機能，実現手法，評価について述べる．

T-Scroll: A Trend Visualization System Based on Clustering of a Time-series of Documents

Mikine Hasegawa[†,☆] and Yoshiharu Ishikawa[††]

On the Internet, a large number of documents such as news articles and online journals are delivered everyday. Documents continually delivered with timestamps such as issue dates are called a time-series of documents. We often need to review major topics and trends from a large time-series of documents, but it requires much time and effort to browse and analyze the target documents. We have therefore developed an information visualization system called T-Scroll (Trend/Topic-Scroll) to display the overall trends extracted from those documents. The system takes periodical outputs of the underlying clustering system for a time-series of documents then visualizes the relationships between clusters as a scroll. Using its interaction facility, users can grasp the trends and the details of the topics contained in the documents. This paper describes the idea, the functions, the implementation, and the evaluation of the T-Scroll system.

1. はじめに

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などのテキスト情報の配信がさかんに行われている．膨大なテキスト情報が日々継続して得られることから，文書クラスタリングや情報抽出などの技術が重要な研究課題となっている[1),9)]．それらを用いることにより，元のテキストデータを直接閲覧する場合に比べてユーザの労力を大幅に軽減することが可能となる．

しかし，配信された大量のニュース記事中から大まかなトレンドをとらえたいといったユーザの要求に対しては，必ずしもこれらの技術は有効であるとはいえない．たとえば，ある時期に配信されたニュース記事の集合をクラスタリングしたとする．クラスタ中の文書に出現する代表的な語を選択し表示する機能があれば，ユーザはクラスタの大まかな内容を把握できる．また，ユーザ自身がクラスタ内の文書を個別に閲覧することも考えられる．しかし，このようにして個々のクラスタの内容の把握は可能であっても，そのクラスタリング結果はある時点のスナップショットでしかなく，時間的なトピックの推移に関する情報が表現されていないため，ユーザがトレンドを把握することは容易ではない．トレンドの把握のためには，それぞれの

† 名古屋大学工学部電気電子・情報工学科（情報工学コース）
Department of Information Engineering, School of Engineering, Nagoya University
†† 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University
☆ 現在，株式会社日本製粉
Presently with Nihon Seifun Co. Ltd.

時点においてどのようなトピックが見られるか，また，時間の推移にともないトピックがどのように変化するかをユーザが容易に理解できるような機能が求められる．

このような背景を受けて，本研究では，配信された文書データのテキスト情報と時刻情報をもとに，ユーザがその大まかなトレンドを把握できるようにするための可視化インタフェースである **T-Scroll**（Topic/Trend-Scroll）の開発を行った[3]．T-Scroll は文書クラスタリングシステム上に構築されており，配信される時系列文書を定期的にクラスタリングした結果を視覚的に提示する．その特徴は，各時点で得られたクラスタリング結果を時間軸上に配置し，クラスタ間の関連性を表すリンクを提示して，トピックのつながりを巻き物（scroll）のような形で表す点にある．ユーザは T-Scroll を閲覧することで大まかなトレンドをとらえることができ，また，必要に応じて詳細な情報を選択表示することも可能である．

ここで，以降の議論を進めるにあたり，本論文で使用する用語について定義する．

- **時系列文書**：配信された日時などの時刻情報が付与された文書のことを総称して時系列文書と呼ぶ[14]．ニュース記事などがこれにあたり，新しい時刻情報が付与された文書が次々と配信されてくるという性質を持つ．
- **トピック**：特定の出来事や事件などに関する文書群を総称するカテゴリを意味する．後述する実験から例をあげると，「プレイステーション 3（PS3）の発売」というトピックには，PS3 の発売を報じるニュース記事群が対応する．なお，本論文では，ある時点において生成された各クラスタが，それぞれ 1 つのトピックにほぼ対応すると想定する．ただし，実際には，明確に 1 つのトピックに関連するとはいえないようなクラスタも存在する．
- **トレンド**：関連するトピックの継続的なつながりにより構成される，大きな話題のまとまりを指す．後述の例を再び取り上げると，「PS3 の発売」と「Wii の発売」という密接に関連するトピックが時間的に連続しているとき，これらを大きなまとまりとして「次世代ゲーム機」というトレンドでとらえるということがあげられる．

ただし，トピックとトレンドの境界は必ずしも厳密ではなく，区別が容易ではない場合も存在する．

本論文の貢献は，時系列文書のトレンドを可視化するインタフェースシステムを実際に構築し，システムとしての実用性および有用性について実践的な立場から評価を行っている点にある．より具体的には以下のようにまとめられる．

(1) 時系列文書集合におけるトピックやトレンドをユーザが把握できるようにするため，文書クラスタリングに基づく可視化インタフェースシステムの概念を提案し，具体的に実現した．
(2) トレンドおよびトピックの詳細をユーザが把握するために有用となる各種機能を提案し，その実装を行った．
(3) 実際に被験者によるシステムの評価を行い，有用性を評価した．

2 章で述べるように，時系列文書集合に含まれるトピックやトレンドの可視化技術，および，クラスタリング結果の分析技術などの関連研究は存在するが，同様の目的に関して実用的な見地からのシステム構築に焦点を当てた研究は，著者の知る限り存在しない．

以下，本論文の構成は以下のようになる．2 章では関連研究を紹介し本システムとの関連性について述べる．3 章では 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．4 章では T-Scroll システムの概要について，5 章ではその機能について説明する．6 章では実装方式について述べ，7 章ではシステムの評価を行う．最後に，8 章では結論と今後の課題を示す．

2. 関 連 研 究

2.1 時系列文書の可視化

時間依存で変化するデータの可視化に関しては，文献 11) に簡単なサーベイが存在するが，時系列的に取得される文書データに特化したシステムはあまり例が見られない．以下では 2 つの関連するシステムについて紹介する．

Havre らによる ThemeRiver は，話題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[4]．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．川の中にいく筋かの色分けされた流れが表示されており，これが 1 つ 1 つの話題（テーマと呼ばれる）に相当する．画面上には，それぞれの流れがどのようなテーマに対応するかを示すためのキーワードが表示される．川の流れの筋の幅は時間につれ変化し，各時点における対応する文書の数を表現する．

対象の文書集合からテーマをどのように選択するかについては，ThemeRiver では力点が置かれておらず，特段の工夫はない．文献 4) の実験では，事前の分析により対象の文書群（例：カストロ首相に関する

40 年間のさまざまな文書群）から 64 個のキーワード（kennedy，mexico，oil など）を選び，個々のキーワードをテーマと呼んでいる．各キーワードが 1 カ月ごとに何件の文書に出現したかをカウントし，この情報を川の流れの幅の計算に用いている．

左右にスクロールするインタフェースを用いるという点では 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．ThemeRiver は視覚的なインパクトを狙ったシステムであり，後述の T-Scroll が提供する，トピックの推移の表現や複数の時間間隔での表示などの機能は有していない．ThemeRiver は，大まかな傾向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閲覧するには，必ずしも強力なツールではない．後述の評価実験で示すように，ユーザが実際にトピックやトレンドを把握するには，文書のタイトルを表示したり，元文書にアクセスしたりする機能が有効である．このような機能も ThemeRiver では支援されておらず，ユーザの分析要求を支援するには不十分であるといえる．

Swan らは，時系列文書の集合をもとに，トピックの存在期間を表現するタイムライン（timeline）を抽出し表示する TimeMine システムを構築した[13]．指定された期間における時系列文書を分析して，近接して出現する同様なトピックの文書群を検出しその時区間を求める．トピックの検出には，クラスタリングではなく統計的指標を用いる．それをもとに，左から右に時間が流れている画面上の該当する箇所に，タイムラインを表す横長の長方形領域を表示する．また，タイムラインに対応するキーワードもあわせて表示する．検出されたトピックごとにタイムラインが提示されるため，ユーザは画面を眺めることでトピックがどの期間に出現したかを把握できる．TimeMine では主要なトピックとその期間を提示すること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，元文書へのアクセス機能などはない．探索的なブラウジングによるトレンドの把握に関しては，T-Scroll の方がより豊富な機能を有しているといえる．

2.2 時間変化するクラスタの分析

可視化を対象としてはいないが，時間的に変化するクラスタを追跡・分析するためのアプローチがいくつか提案されている．Mei らは，時系列文書の集合の中から主要なテーマを発見するための統計的なアプローチを提案した[10]．この手法では，テーマはある時区間における語の確率分布としてとらえられ，一種のクラスタとして表現される．連続する時刻におけるテーマ間の関連性は，確率的な指標に基づいて判定される．導かれたテーマの推移を表すグラフは，T-Scroll におけるクラスタの関連を表すグラフと似た構成となる．文献 10) では，さらにそのグラフを大局的に分析してパターンを発見する手法が提案されている．ユーザインタフェースや可視化を目的とはしていないが，対象データとアイデアの一部には T-Scroll と共通するものがある．

Spiliopoulou らによる MONIC では，クラスタの変化をさまざまなパターンを用いて把握するためのアプローチが提案されている[12]．T-Scroll と同様，対象とするクラスタの集合は時間的に変化し，要素の追加・削除が発生する．クラスタの関連性は T-Scroll と同様，集合演算をもとに定義されている．MONIC では，クラスタのスナップショットの履歴情報をもとに，クラスタの分岐・融合・消滅やサイズの変化などのイベントを発見する．

時間的に推移・変化するクラスタの関連を，クラスタどうしの集合演算をもとに検出するアプローチは，移動するオブジェクトの集合からの移動クラスタの検出などでも用いられている[6]．

2.3 最新のトレンドの検出

T-Scroll は，過去に取得された時系列文書に対し，後の時点から遡及的にトレンドの分析・把握を行う場合に特に有効であると考えられる．一方で，最新のトレンドを把握したいというユーザの要求も存在する．このような要求に対して，Yahoo!トレンドワード☆では，過去 24 時間にブログやニュースで話題になったキーワードを選択表示している．また，kizasi.jp☆☆においても，ブログから最新の話題のキーワードを抽出して表示している．トレンドをとらえたいという点に関しては T-Scroll と関連しているが，対象とするデータやユーザの要求は大きく異なっている．

3. 新規性に基づく時系列文書のクラスタリング手法

本研究が基礎とするのは，文献 5)，7)，8) において提案されている，新規性に基づく時系列文書のクラスタリング手法である．これは時系列文書の特性を考慮して工夫された手法であり，以下の特徴を有する．

(1) 類似度計算において，単に文書の内容のみを考

☆ http://trendword.yahoo.co.jp/
☆☆ http://kizasi.jp/

慮して類似度を与えるのではなく，文書の新規性も考慮に入れた類似度を導入し，新規性の高い文書をより重視したクラスタリングを行う．

(2) 新たに文書が配信された場合には，最新のクラスタリングを得るため再クラスタリングが必要となるが，その処理を効率化するために低コストのインクリメンタルな処理を行う．

(3) 新規性の高い文書をより重視するという性質により，文書は古くなるほどクラスタリング結果に与える影響が小さくなり，次第に外れ値（outlier）となる．そのため，十分古くなった文書は自動的にクラスタリングの対象から削除し，クラスタリングの処理における無駄を省く．

このようなアプローチにより，時系列的に配信されてくる文書データを，新規な文書に重点をおいて定期的にクラスタリングすることで，最近の主要なトピックを中心に情報を集約することを可能としている．

文書類似度に関して詳しく説明する．時系列的に配信されるニュースなどの時系列文書においては，文書の価値はそれが入手された時点を基準として，時間が経過するにつれて一般に低下していくと考えられる．新規性に基づく文書クラスタリング手法では，得られた文書データの影響力が時間の経過とともに徐々に逓減するような**影響力の逓減モデル**を提案し，そのモデルに基づく文書間の類似度を導いている．

影響力の逓減モデルでは，文書の価値（重み）が時間の経過に従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．

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

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ータである．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

$$tw_d = \sum_{l=1}^{n} w_{d_l} \tag{2}$$

で与え，文書 d_i の文書集合中での生起確率を

$$\Pr(d_i) = \frac{w_{d_i}}{tw_d} \tag{3}$$

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[5),7),8)]．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り，外れ値となっていくという性質を有する．この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．

実際に用いられているクラスタリングのアルゴリズムは，k-means 法[2)] の拡張である．新たな文書群が取得されたときには，それを反映するため再クラスタリングを行う必要があるが，一からクラスタリングを行うコストが高いため，前回のクラスタリング結果における k 個のクラスタ代表を初期的なクラスタ代表として採用する．これにより，クラスタリングが早期に収束するだけでなく，クラスタリングの質自体も向上することが分かっている[8)]．また，通常の k-means 法と異なり，外れ値の文書を検出し，クラスタリングの対象から自動的に外す機能も有している．これにより，外れ値の存在によるクラスタリングの質の悪化を防いでいる．

以上のアプローチに基づき，新規性に基づく文書クラスタリング手法では，時系列的に配信されてくる文書集合に対し定期的にインクリメンタルなクラスタリングを行い，そのつど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主要な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開発を行ったのが，以下で述べる T-Scroll システムである．

4. T-Scroll システムの概要

4.1 システムの特徴

本研究で開発を行った **T-Scroll**（Topic/Trend-Scroll）システムの特徴は，主に以下のようになる．

(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．

☆ 各文書ベクトル $\boldsymbol{d}_i$ は，基本的には $tf \cdot idf$ 方式で重み付けされたベクトルである．ただし，通常の $tf \cdot idf$ 方式の拡張となっており，逆文書頻度（inverse document frequency）idf は時間的に変化するという性質を持つ．詳細は省略する．

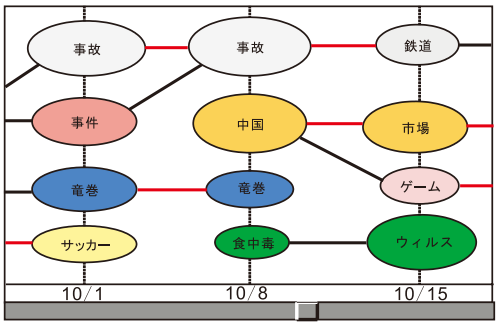

図 1 T-Scroll システムの概念
Fig. 1 Concept of T-Scroll system.

(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（関連する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，1 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，隣接する時刻間で関連の深いクラスタのペアを把握可能とする．

(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．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ことは粗視化であり，その逆は詳細化にあたる．これは，OLAP（On-Line Analytical Processing）におけるロールアップ（roll-up）とドリルダウン（drill-down）の機能[2] に対応づけることができる．

以上のような機能により，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（scroll）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．

4.2 基本的なアイデア

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，2006 年の 10 月の記事に対し，1 週間ごとにクラスタを表示している．なお，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クラスタの集合を表している．

クラスタ上のラベルとして，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

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

で求めている．つまり，クラスタ内の各文書 d_i 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．式 (5) 以外にも，idf を含む計算式 $score(t_j) = \sum_{d_i \in C_p} \Pr(d_i) \cdot tf_{ij} \cdot idf_j$ など，他の複数の候補に対して予備的な実験を行ったが，著者らによる主観的な比較では，ラベルとしてふさわしい一般性のある語が選ばれやすいという点で，式 (5) が最も優れていた．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検討したが，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 語だけを選択表示している．

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．ある時刻におけるクラスタ C_i とその次の時刻のクラスタ C_j 間の関連度を

$$\Pr(d \in C_j | d \in C_i) = \frac{|C_i \cap C_j|}{|C_i|} \qquad (6)$$

という確率により定義する．この式は，クラスタ C_i に含まれるある文書 d がクラスタ C_j にも含まれる確率であり，C_i 中の文書がどれだけ C_j に移動したかを表す．C_i，C_j に対して非対称であり，$C_i \to C_j$ というリンクの方向に対応している．システムは，ある閾値以上の関連度についてリンクを作成する．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．なお，7 章で述べる実験では，予備的な実験をもとに閾値 0.5 以上の場合にリンクを生成している．

また T-Scroll では，ドリルダウン/ロールアップ機能をサポートする．たとえば図 1 の例では，表示の時間間隔は 1 週間単位となっているが，より詳細化して 3 日単位で表示したり，逆に 2 週間単位に伸ばしたりするなどが可能である．なお，実際の実装では定期的にクラスタリングを行い，そのつど式 (6) により，どのクラスタのペアの間にリンクを作成するかを求め，その情報をグラフ構造として保存しておく．T-Scroll インタフェースで表示を行う場合には，保存されたグラフ構造から表示に必要な情報を抜粋して表示する．

たとえば 1 週間単位で表示する場合には，まずクラスタリング結果から 1 週間ごとにクラスタ集合を選択する．クラスタ間のリンクを表示するのは，保存されているグラフ構造において，そのクラスタのペアの間に経路が存在する場合である．このようなリンクの提

示方式をとる理由は，表示間隔を大きくとった場合，式 (6) の確率の値が一般に 0 に近い値となることによる．時間的に離れたクラスタのペア間には，共通の文書があまり含まれないためである．

以上が T-Scroll の根幹となるアイデアである．実装システムでは，他にも工夫を行ってユーザへの便宜を図っている．それらについては次章で説明する．

5. 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．

5.1 インタフェース画面

図 2 は，ネットワーク上のニュースサイトから入手した記事データに対し，2006 年 10 月 1 日から 1 週間刻みで 12 月 31 日まで表示を行った例を示している．この実験データについては，6.1 節で詳しく説明する．各時点において $k = 20$ 個のクラスタが生成されている．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 法に基づくクラスタリングを行うため，クラスタ数 k はその初期化時に固定する必要があることによる．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行う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切り替えることは技術的に可能である．

クラスタ上に書かれた楕円の大きさはクラスタに含まれる文書数に対応しており，トピックの規模を示している．具体的には，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サイズ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とらえることが可能となる．

クラスタの質の良さについても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，クラスタの輪郭の線の色により，そ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とする．具体的には，クラスタ C の品質のスコアを以下のように与える[8]．

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

$|C|$ はクラスタ C 中の文書数を表す．$avg\_sim(C)$ はクラスタ内の文書の平均類似度を表し，以下のように定義される．

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

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指標となっている．本システムが利用しているクラスタリングのアルゴリズムでは，k-means 法の繰り返し処理における目標を，k 個のクラスタの品質のスコアの総和の最大化においている[8]．よって，上記の品質の提示手法は，下位のクラスタリングシステムにおけるクラスタ品質の考え方と整合しているといえる．

楕円の中には，そのクラスタを最も適切に表すようなキーワード☆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．

4.2 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す．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，見やすさを考慮して，1 つ前の時点におけるリンクの配色とできるだけ整合するように色を選択している．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，後述の SVG による実装では，ブラウザの制限により，指定した太さには必ずしも表示されないという問題があったため，すべて同じ太さで表示している．

5.2 詳細情報の表示機能

T-Scroll では，ユーザが各クラスタの詳細を調べることができるための機能を提供している．

上述のとおり，図 2 のように，クラスタに対するラベルとしてクラスタを代表するキーワードを提示するが，これだけではクラスタの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閲覧可能とする機能を提供している．クラスタを表す楕円上にマウスカーソルが乗ると，そのクラスタに関連するキーワードのリストを表示する．実行した様子を**図 3** に示す．これは左上の「中心」とラベル付けされたクラスタ上にマウスカーソルを移動した場合を示しており，クラスタ内の単語のうち，スコアが上位 20 位のものが順に表示されている．

上記のようなキーワード表示機能によってクラスタの大まかな内容は把握できるが，実際にどのような文書がクラスタに含まれるかは分からない．よって本システムでは，クラスタ上でクリックすることで，クラ

☆ 実際には，日本語を対象とした実装では形態素，英語を対象とした実装では語幹である．

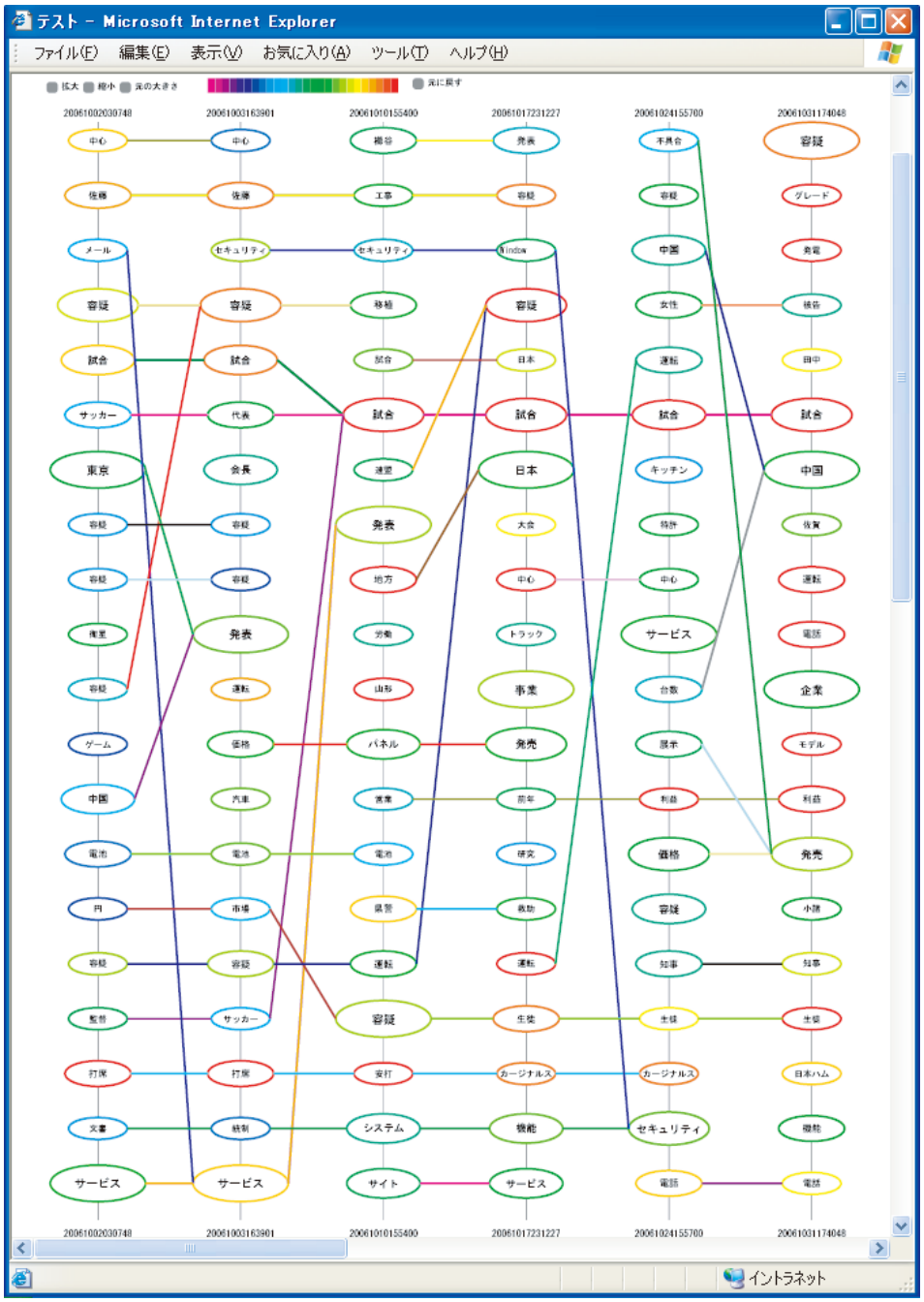

図 2 T-Scroll のスクリーンショット
Fig. 2 Screenshot of T-Scroll.

スタに含まれる文書を一覧表示する機能も実現している．実行の様子を**図 4** に示す．クラスタに含まれる文書のうち，発行日時が新しいものから上位 10 位のタイトルを表示している．また，図は省略するが，タイトルをクリックすることで文書の内容を表示することも可能である．さらに，図 4 に表示されている「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のタイトルを表示することもできる．

5.3 その他の機能

T-Scroll は，そのほかにも以下のようなユーザの便宜を図る機能を提供している．

- **キーワードによるクラスタの強調表示**：キーワー

図 3 クラスタのキーワードリストの表示
Fig. 3 Displaying keyword list for a cluster.

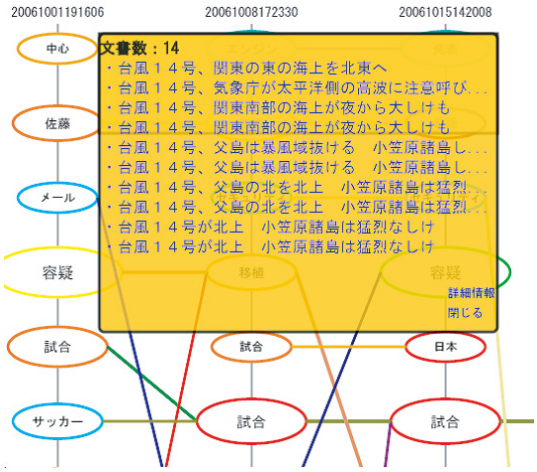

図 4 クラスタ内の文書のタイトルの表示
Fig. 4 Displaying document titles in a cluster.

ド入力フィールドに検索語を与えることで，キーワード検索を行うことが可能である．検索語がスコアの上位 20 位までのキーワード集合に含まれる場合，そのクラスタを強調表示する．実行の様子を**図 5** に示す．これは図 2 に対して，キーワード「サッカー」を含むクラスタを強調した結果である．濃く塗られたものが強調されているクラスタである．また，検索語は複数与えることも可能である．その場合は，与えられた検索語のどれか 1 つがクラスタのキーワード集合に含まれていれば強調表示される．この機能により，特定のトピックの推移をより容易に把握することが可能になる．

- **質によるクラスタの強調表示**：前述のように，本システムでは，クラスタを表す楕円の輪郭の色によりクラスタの質の良さを表している．しかし，画面全体を見渡したとき，どのクラスタの質が良いかが即座には分からないこともある．よって，本システムでは，クラスタの質を指定することによって，指定された質を持つクラスタを強調表示する機能も提供している．図 2 の上部に表示されているスペクトル分解の色表示上で，色（例：赤）をクリックすると，その色を輪郭線に持つクラスタが強調表示される（図については省略する）．この機能は，前述のキーワードによる強調表示機能と併用できるため，指定されたキーワードを持ち，質が高いクラスタを選択表示することが可能である．

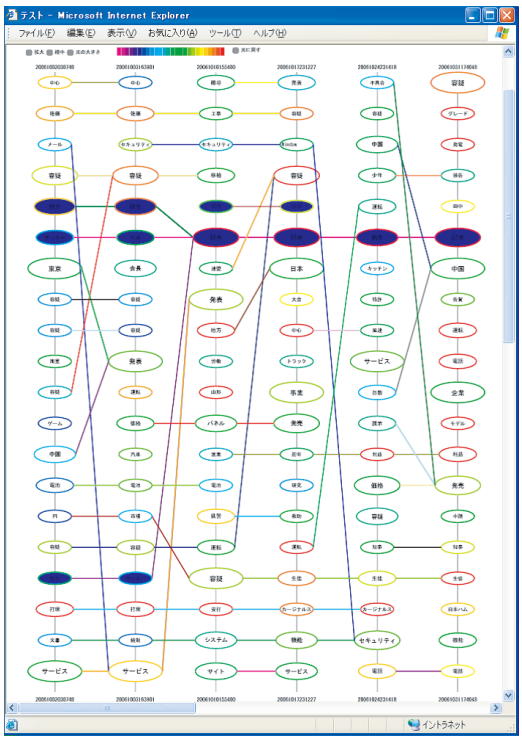

図 5 キーワードによるクラスタの強調表示
Fig. 5 Emphasized cluster display based on keywords.

6. システムの実装

T-Scroll システムの実装について述べる．**図 6** に T-Scroll システムの構成を示す．以下では，各構成要素について説明する．

6.1 実験対象の情報源

今回の実験システム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．情報収集を 2 時間おきに行い，各サイトについて，前回の情報収集時から更新された記事について，本文などの情報を抽出した．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．入手したテキスト情報を形態素分析

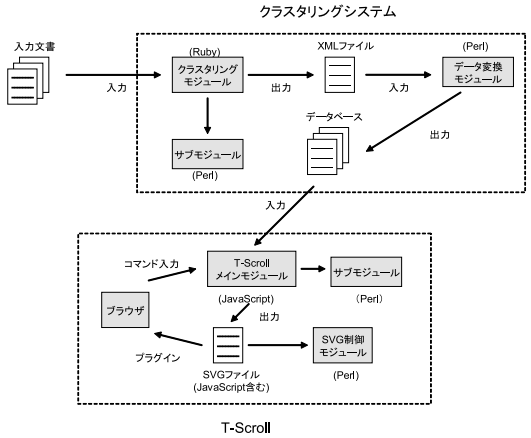

図 6 T-Scroll システムの構成
Fig. 6 System organization of T-Scroll system.

ソフトである茶筌☆により処理し，名詞あるいは未知語とされた形態素を抽出している．茶筌の辞書に登録されていない外国語や新語などに対応するため，未知語も特徴語として利用する．

6.2 文書クラスタリングシステムとの連携

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[7),8)] 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に対し，各時点で取得した新たな文書集合を与えることで，その時点の最新のクラスタリング結果が得られる．クラスタリング結果は XML 形式のファイルとして出力される．

実験システムでは，クラスタリングは 6 時間ごとに実施しており，$k = 20$ 個のクラスタを毎回生成している．また，式 (1) におけるパラメータ λ は $\lambda = 0.958$ と設定した．これは，1 週間程度で式 (1) に示す文書の重みが 1/2 となるような設定である．また，10 日過ぎた文書はクラスタリングの対象から外すことにした．これにより，毎回のクラスタリング結果には過去 10 日以内の記事が含まれることになる．

6.3 インタフェースの実装

T-Scroll は，クラスタリングシステムによって出力された XML 形式のファイルを入力とする．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．

☆ http://chasen.naist.jp/hiki/ChaSen/

T-Scroll は，ユーザから対象の期間やクラスタ表示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．その際には，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．このサブモジュールは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呼び出される．このような仕組みにより，先に述べた対話的システムの機能を実装している．

7. 実験結果およびシステムの評価

T-Scroll システムを用いて行った実証実験の結果，およびシステムの評価について述べる．

7.1 著者らによる主観的評価

まず，実際にシステムを開発・利用した著者らにより得られた知見を報告する．

7.1.1 表示対象の期間

T-Scroll では，表示対象の期間（例：2006 年 10 月 1 日〜12 月 31 日）を任意に設定可能である．しかし，それを長期（例：3 カ月以上）に設定することはあまり有効とはいえなかった．トレンドを観測するには 1〜2 カ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ことと，長期の場合には表示が煩雑になりインタフェースの動作が重くなることが理由である．

7.1.2 表示する時間間隔の設定

表示する時間間隔の設定については，実装システムでは「1 日」「3 日」「1 週間」「2 週間」を選択できる．他の時間間隔に対応することもシステムの軽微な修正で対応可能である．先に示した図 2 は 1 週間間隔で表示した場合を示しているが，同じ 2006 年 10 月初頭の記事を 1 日ごとに表示すると，**図 7** のように比較的単調な表示となる．1 日程度では大きなトピックの変化がないため，表示が冗長であるという印象を受ける．大まかなトレンドを把握するという目的に関しては，図 2 のように 1 週間ごとに表示する方が，より適切な表示であると感じられた．図 2 の方が，クラスタ間のリンクの交差などが見られ，視覚的にも面白いものとなっている．ただし，1 週間間隔で表示した場合には，リンクが張られている隣接するクラスタ間でトピックが異なる場合が見られた．3 日間隔の表示は，両者の中間の傾向を示した．著者らは，今回の対

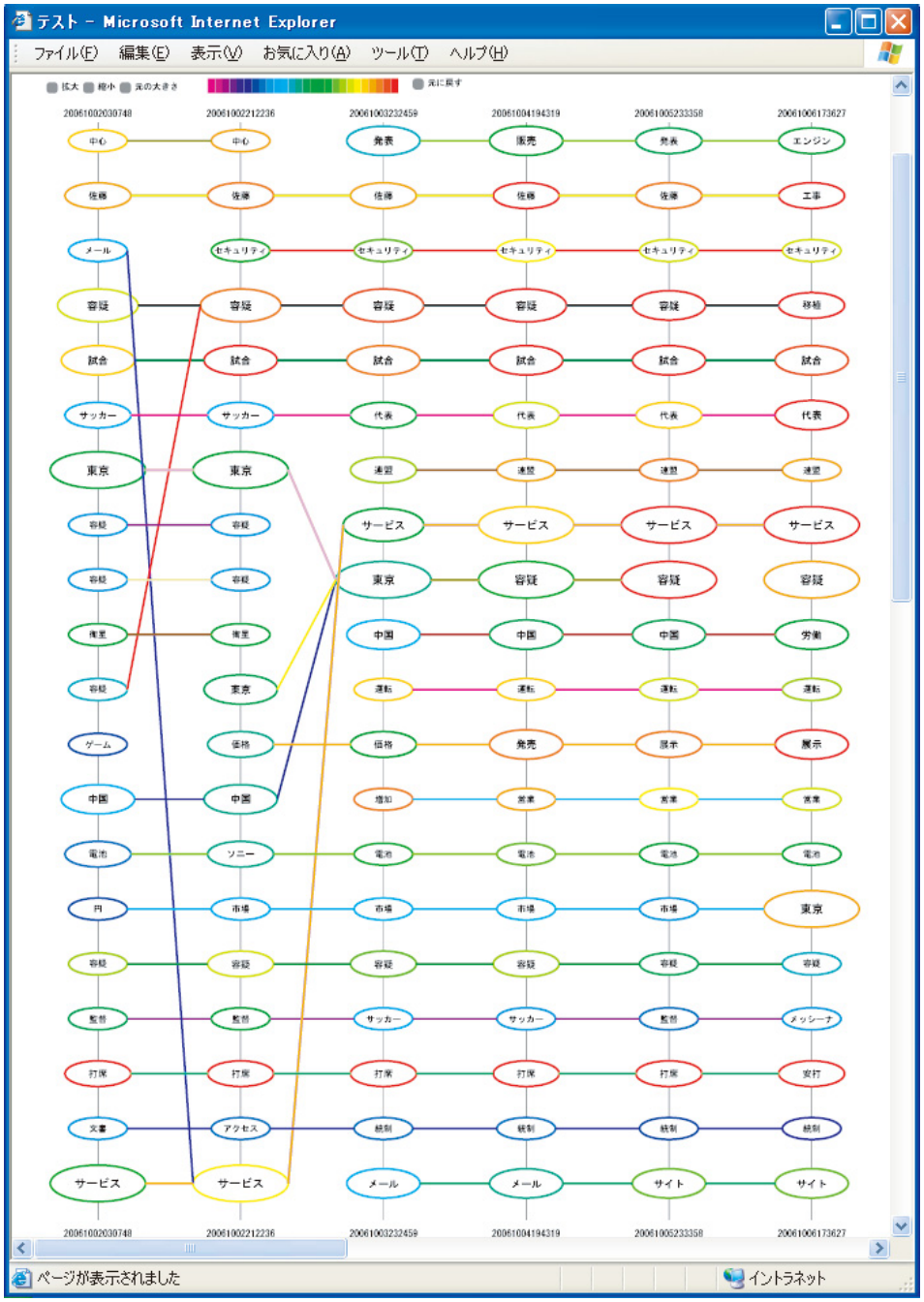

図 7 T-Scroll のスクリーンショット（1 日刻み）
Fig. 7 Screenshot of T-Scroll (daily basis).

象データに関しては 3 日もしくは 1 週間の時間間隔が適切であるという印象を受けた．

なお，特に図 7 で顕著であるが，トピックに変化がない場合に水平方向のリンクが現れる理由について述べておく．3 章で述べたように，本システムが利用している新規性に基づく文書クラスタリング手法[7),8)] では，k-means 法に基づくクラスタリングをインクリメンタルな処理に拡張している．新たなクラスタリング処理は，前回のクラスタリング結果における k 個のクラスタ代表と，前回の文書のクラスタへの所属情報を初期状態として再利用する．ただし，新規に入手された文書の情報と，今回削除された文書の情報を反映

してクラスタリングを行う．その結果，トピックに大幅な変化がなければ，前回のクラスタリング結果とクラスタ番号も含め同様の構成となり，水平的なリンクが支配的になる．単調な表示の箇所は，トピックに大きな変化がなく継続している状況であると理解できるため，この性質はユーザの理解しやすさのために有用であると考える．

7.1.3 クラスタのサイズ

5.1 節で述べたように，本システムではクラスタ中の文書数をクラスタの表示サイズに反映させている．クラスタのサイズが大きいことは，そのクラスタがその時点における主要なトピックに対応しており，結果としてクラスタ中の文書数が多くなるのではないかという仮説を考えることができる．この考え方によれば，大きいクラスタほど，より主要なトピックに対応することになる．しかし，実際のクラスタリング結果を見ると必ずしもその仮説は成り立たず，サイズが特に大きいクラスタには，雑多なトピックの文書が含まれる傾向が見られた．以下に述べるように，これは用いているクラスタリング手法[7),8)] の特性による．

一般に，新規性が高くある程度の文書数があるようなトピックは個別にクラスタを構成し，そのサイズは小規模から中規模となる．一方，残りの文書（古いトピックの文書や他とあまり類似していない文書）は，比較的大きなサイズのクラスタ（通常 1 個か 2 個出現する）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化された一般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特に対応せず，トレンドを把握するためにはあまり有効ではないと考えられる．T-Scroll ではクラスタの質の良さをクラスタの輪郭線の色で表現していることから，クラスタのサイズだけでは分からないクラスタの質をうまく表していると考えられる．

7.1.4 システム利用で得られた観測結果

本システムで，2006 年 9 月から 12 月に取得した文書データをもとにした表示について，実際に観測できた内容を以下にまとめる．まず，全体的な傾向について述べる．

- 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・事故に関する記事が含まれる．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分かった．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，それだけで単独にクラスタを構成し，継続的に存在する：鉄道関係の記事には，それ特有の語の出現があり，比較的継続して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 1 つのクラスタを構成する：今回，sportsnavi.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，非常に質の良いクラスタである場合が一般的であった．一方，野球については，提供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 技術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ものが多く含まれるということを反映している．

以上は，実験対象のサイトのニュース記事の内容を考えると，妥当な結果であるといえる．

一方，今回対象とした期間では以下のようなトレンドが観測された．ただし，継続したクラスタをどの場合にトレンドと見なすかは，著者の主観に基づいている．

- 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 年の後半に知事の汚職事件が連続したことが反映されている．クラスタのラベルとしては「佐藤」（知事の名前），「容疑」，「知事」，「汚職」などが代表的であった．
- いじめや教育は 2006 年の後半には重要な社会問題となったが，これもトレンドとして出現した．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 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，火事に関するトレンドが出現した．
- 2006 年は，ノロウイルスの大流行にともない，12

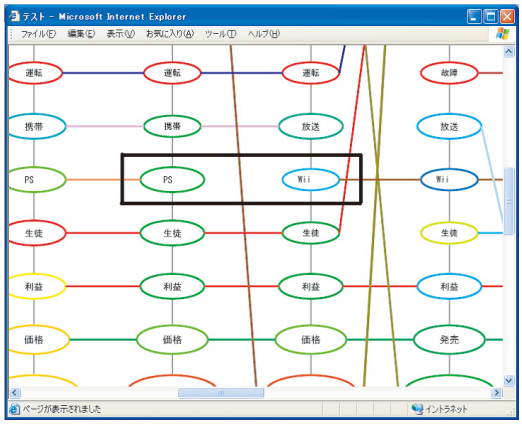

図 8 PS3 から Wii へのトピックの変化
Fig. 8 Topic transition from “PS3” to “Wii”.

月を過ぎたころからノロウイルスに関するトレンドが出現した．

- 2006 年の 12 月下旬からバラバラ殺人が連続して起こったこともあり，関連するトレンドが出現した．

特に興味深い事例として，次世代ゲーム機の発売に関するトレンドを取り上げる．2006 年後半には PS3, Wii という次世代ゲーム機が相次いで発売されたが，**図 8** に示すような関連するクラスタが観測できた．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」のラベルが見られるようになり，大きなトレンドとしては「次世代ゲーム機」が継続しているが，トピック自体は PS3 から Wii へと移行していることが分かる．図 8 において，「PS」とラベル付けされたクラスタが「Wii」というラベルのクラスタに移る箇所が，この移行の時点に該当する．

「PS」と「Wii」の間にはリンクが存在しないが，これは両クラスタ間には閾値を満たすほどには共通する文書が存在していないことを意味する．この理由について説明する．7.1.2 項で述べたように，新規性に基づくクラスタリング手法では，新たなクラスタリング処理の際に前回のクラスタリングにおける k 個のクラスタ代表を初期値として活用する．この事例においては，前回の PS3 に関するクラスタ代表が利用されるため，次世代ゲーム機に関する新たな文書の多くがこのクラスタ代表に類似文書として割り当てられる．しかし，その中には Wii に関する文書が多数含まれているため，k-means 法の繰り返しの過程で Wii のトピックにより類似した方向にクラスタ代表がシフトし，その結果，一部の PS3 に関する記事はこのクラスタから外れることになる．結果的に，次世代ゲーム機のトレンド自体は維持されるが，リンク自体は表示されず，トピックの変化が生じたことが観測できることになる．

その他，システムを利用して気づいた点として以下のものがあった．

- 地震や台風などの大規模自然災害に関する記事は，通常それ単独でクラスタとして出現する．
- 時期が過ぎた話題のクラスタは，事件関係のクラスタに統合されていく傾向がみられた．これにともない，事件関係のクラスタは文書数が多い傾向がみられた．
- 2006 年 10 月には北朝鮮の核実験があり，ニュースで注目されていたが，今回の実験では単独のクラスタとしてはほとんど出現しなかった．「政府」や「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．

なお，突発的に発生した事件などのトピックに対しては，最初の記事が出た時点ではなく，ある程度記事数が出現した時点（通常 2，3 日後）でクラスタが出現する．すなわち，そのトピックに関する文書がある程度増えるまでは画面上には出現しない．これは，T-Scroll システムがクラスタリングシステムを基盤としていることによる性質である．

7.2 被験者による評価

以下では，被験者による評価実験の結果について述べる．

7.2.1 全般的な印象

まず，ユーザインタフェースとしての T-Scroll の全般的な評価について尋ねた．被験者は情報分野の学部生 10 名であり，事前にシステムの説明を行ってから実際に利用してもらった．5 段階の絶対評価とし，3 が「どちらともいえない」を表し，1 が「非常に悪い」，5 が「非常によい」であり，2，4 はそれらの中間である．評価項目は，使いやすさ，分かりやすさ，有用性，デザインの 4 つである．

評価結果を**図 9** に示す．平均値に加え標準誤差をプロットしている．有用性については平均 3.7 というよいスコアを得ているが，使いやすさに関しては 2.5 という値であり，さらに改善が必要であることが示されている．分かりやすさ，デザインはそれぞれ 3.1 と 2.8 のスコア値であり 3 に近い値となっているが，デザインに関してはばらつきが大きいことが分かる．

被験者からは以下のコメントが得られた．

(1) 表示を更新すると着目していた領域がリセット

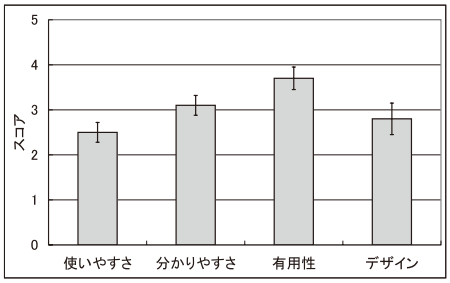

図 9 全般的な評価スコア
Fig. 9 Overall evaluation scores.

されてしまうので，表示位置を保持してほしい．

(2) キーワード指定によりクラスタを強調したときに，クラスタのラベルが見にくくなる．

(3) 応答が遅い．

(4) 単語の並びだけでは直感的に内容が把握しにくい．1 つのラベルだけではクラスタの内容が分かりにくい．

(5) リンクがなぜつながっているかの根拠が分かりにくい．

(6) キーワードの抽出，ラベル付けの精度の向上が必要．

(1)，(2) に関してはシステムの改良で対応できると考える．(3) の応答速度が遅い理由は，現在の実装では，クラスタリング結果の XML ファイルを実行時に多数読み込んでいることによる．前処理を行い，表示期間と表示の時間間隔の組合せごとに必要な情報のみを抜粋したファイルを事前に作成しておき，T-Scroll の実行時にはそれを選択し読み込むようにすれば，余分な記憶領域が必要ではあるものの，処理速度の改善が図れると考える．(4) の問題については著者らも問題点を感じており，図 3 および図 4 のように，各クラスタの代表的なキーワードや文書タイトルを簡単に閲覧する機能を提供している．

(5) に関しては，時間間隔が長期（例：2 週間）であり，リンクされている前後のクラスタ間に共有する文書が少なく，かつ，トピックが時間の経過とともに変化している場合が主に該当する．これは指摘のとおりであり，今回の対象データについては，2 週間という長期間の時間間隔の設定は，関連が把握しにくいという点であまり有効ではないといえる．

(6) で指摘されたキーワードの抽出とラベル付けに関しては，改善の余地がある．今回，茶筌システムをそのまま利用したため，茶筌の辞書に登録されていない語（特に新語や固有名詞）については誤った抽出が行われる場合が見られた．これに対しては，辞書のチューニングを行えばある程度対応できると考える．また，ラベルとして必ずしも直感的に分かりやすい語が選ばれないという問題に関しては，今後さらに検討が必要である．ラベルとして利用可能な語彙のリストを用いて統制することも考えられるが，語彙リストを利用する方式では，先の例に示した「PS」と「Wii」のような固有名詞や新語に関するラベルを与えることが必ずしも可能ではないという問題点が存在する．

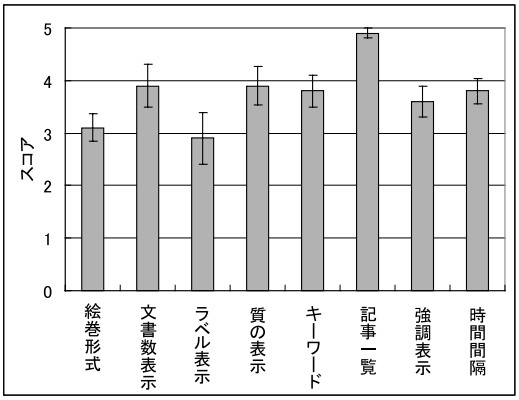

図 10 機能ごとの評価スコア
Fig. 10 Evaluation score for each function.

7.2.2 機能別の評価

次に，本システムが提供する各種機能に対する個別の評価の結果を**図 10** に示す．図の左から順に，以下のような項目について評価を求めた結果である．なお，被験者は先の実験と同じである．

- 絵巻形式：絵巻形式で可視化するアプローチに関する評価
- 文書数表示：クラスタの文書数をクラスタの表示サイズに対応付ける方式の妥当性についての評価
- ラベル表示：ラベル表示機能についての評価
- 質の表示：クラスタの質を輪郭線の色で表す方式の評価
- キーワード：キーワードリストの表示機能（図 3）の評価
- 記事一覧：記事のタイトル一覧を表示する機能（図 4）の評価
- 強調表示：キーワードによるクラスタの強調表示機能（図 5）の評価
- 時間間隔：表示の時間間隔の設定機能に関する評価

「ラベル表示」を除き，平均するとスコア 3 以上を得ており，個別の機能については比較的良い評価が得られているといえる．「絵巻形式」については平均スコアが 3.1 である．平均スコアがあまり高くなかった

理由は，この実験では T-Scroll に対する比較対象がなかったことも一因としてあると考えられる．一方，「ラベル表示」に関しては，前項でも触れた理由が存在すると考えられる．「記事一覧」の機能については，非常に有用という評価が得られた．クラスタの内容を具体的に確認するためには記事タイトルの一覧表示がやはり有効であることを意味している．

各機能についての改善のための重要なコメントとして，以下のようなものがあった．

(1) クラスタの文書数の大小は気にならない．
(2) 色によるクラスタの質の良さの表示が妥当であるかが，クラスタ内を実際に見てもよく分からない．
(3) キーワード検索の機能は利用するのが面倒である．
(4) 時間間隔を変えた場合の変化が分かりにくいことがある．
(5) 画面上のクラスタの配置にも工夫がほしい．

(1) は，被験者によっては，クラスタの文書数にはまったく興味がない者がいたことを意味している．(2) については，まず，色以外の表現方式を検討することが課題として考えられる．表示された色が，実際にクラスタ内を閲覧したときに持つクラスタの質についての印象と一致するかという点については，明らかに一致するとはいえないというのが実情である．人手によるクラスタの質の評価は容易ではなく，また，隣り合う色（例：赤とオレンジ）程度では質の差が明確ではない．しかし，著者らによる印象では，最良に近い色（赤）と最悪に近い色（紫）で輪郭が表示された 2 つのクラスタ間には明らかな質の違いが感じられる．特に，記事数が少なく輪郭が赤であるクラスタは，その中の記事のトピックが一致しており，他のクラスタと明確に区別できることが一般的であった．(4) の時間間隔については，確かに「3 日ごと」と「1 週間ごと」などでは，表示指定を切り替えても，見た目は大幅には変わらない．しかし，「1 日ごと」と「1 週間ごと」などでは違いが顕著になるため，機能としては有用ではないかと考える．

なお，本実験では，被験者に対し「1 日ごと，3 日ごと，1 週間ごと，2 週間ごとのうち，今回のデータについて時間間隔の設定はどれが適切と感じたか」という質問も行った．結果は「3 日ごと」と「1 週間ごと」が 5 人ずつであった．

7.2.3 トレンドの観測可能性に関する検証（1）

次に，2006 年 11 月において継続的に大きく報道され，トレンドといえる話題について，被験者が実際にそれらをインタフェース上で確認できたかの検証を行った．実際には以下のトレンドを対象とした．

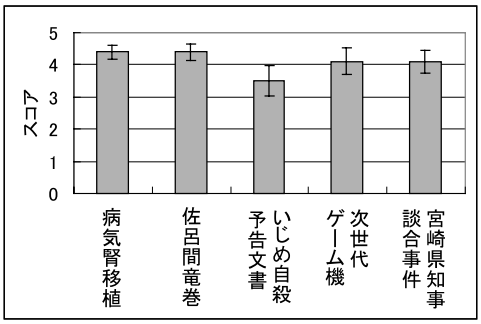

図 11 トレンドの観測可能性
Fig. 11 Observability of trends.

(1) 病気腎移植：愛媛県の病院で，多くの病気腎移植が行われたことが発覚し，11 月 3 日ごろから一連のニュースとなった．
(2) 佐呂間竜巻：11 月 7 日に北海道の佐呂間町で大規模な竜巻が発生し，死傷者が出た．
(3) いじめ自殺予告文書：2006 年後半にはいじめに関する話題が継続して現れたが，この時期に文科省にいじめ自殺予告の文書が送られ，文部科学大臣による対応が行われた．いじめ問題自体はこの時期以前にも大きな話題であったが，11 月 7 日に自殺予告文書が送付されたため，新たな議論が巻き起こった．
(4) 次世代ゲーム機：11 月半ばに PS3，Wii という次世代ゲーム機が相次いで発売され，話題となった．
(5) 宮崎県知事談合事件：2006 年後半には知事の汚職事件が相次いだが，この時期には宮崎県の知事談合事件が大きく報道された．11 月半ばから大きな問題となり，12 月に知事が逮捕された．

先の実験と同じ 10 名の被験者による評価結果を**図 11** に示す．それぞれのトレンドが実際にインタフェースを用いて把握できたかどうかを，1 から 5 の段階で回答してもらった．

図に示されているように「いじめ自殺予告文書」以外のトレンドについては，いずれも平均 4 以上のスコアが得られており，十分トレンドが把握できていることが確認できた．「いじめ自殺予告文書」に関しては，スコアを 2 とした者が 5 名，5 とした者が 5 名であり，判断が大きく分かれた．この理由であるが，いじめ問題は自殺予告文書が送られる以前からトレンドを構成していたことによる．いじめ自殺予告文書につい

ては，あるときにはいじめ問題全般とは独立して別個のクラスタの連鎖が現れるが，あるときには本体に合流するなどの振舞いが見られ，その結果，スコアを 2 とした被験者は，いじめ自殺予告文書に関するトレンドが観測しにくかったと考えたと思われる．一方，スコアを 5 とした被験者は，いじめ問題全体のトレンドといじめ自殺予告文書のトレンドを総合的に判断したものと思われる．

大きなトレンドに対し，話題が包含されるような部分的なトレンドが派生して，それが別個の流れを作ることはあまり見られるものではない．しかし，利用者にとってはそのような流れをつかむことは必ずしも容易でないと考えられるため，派生的なトレンドをどのように表現するかは今後の課題としたい．

7.2.4 トレンドの観測可能性に関する検証（2）

前項の実験ではトレンドの候補を事前に提示していたが，この実験では，候補を事前に知らされない被験者が，実際にトレンドを観測できるかについて検証を行う．前項と同様，2006 年 11 月を対象として，11 名の被験者を用いて実験を行った．なお，今回の被験者には，前項までの実験における被験者との重複はない．

実験においては，インタフェースの説明や，トピック，トレンドなどの概念の簡単な説明を行った後，主要なトレンドとそれらがどの程度はっきり観測できたかを報告してもらった．報告するトレンドの個数については，0 個以上の任意の数と指示した．観測できたトレンドには，はっきり確認できた場合を 3，観測できたがあまり明確でない場合を 1，その中間の場合を 2 としてスコアを与えてもらった．なお，「事件一般」，「サッカー関係」など，定常的に話題が見られるものについては，報告しなくてよいものとした．

実験結果を**表 1** に示す．報告されたトレンドごとに，それが観測できたと回答した被験者数，スコアの平均および標準誤差を示す．表 1 のエントリは，回答者数の降順に，また，同じ回答者数の場合はスコア平均の降順にソートしている．なお，1 名のみが回答したトレンドが 6 件あるが，それらは掲載していない．

この結果を見ると，半数近くの被験者が，「病気腎移植」，「次世代ゲーム機」，「いじめ問題」，「PC・携帯新モデル」について，トレンドが観測できたと報告している．「病気腎移植」についてのスコアが低い理由としては，「病気腎移植」に対応するクラスタは質は高いが比較的小規模なクラスタであったため，被験者は「小規模である」ということを「トレンドがそれほど顕著でない」と判断したのではないかと想像される．著者らは，被験者がクラスタの質の高さとクラスタの大きさの双方を考慮してトレンドが顕著であると判定することを期待したが，実験結果を見ると，クラスタの大きさの方がより重視されているように感じられた．被験者に対し実験を説明する際に，どのような場合によいトレンドと見なすかをより詳細に説明すると，多少傾向が変わるのではないかと考えられる．

表 1 観測されたトレンド（2006 年 11 月）
Table 1 Observed trends (November 2006).

トレンド	回答者数	スコア平均（標準誤差）
病気腎移植	8	1.88 ± 0.30
次世代ゲーム機	7	2.57 ± 0.20
いじめ問題	6	2.67 ± 0.33
PC・携帯新モデル	5	2.80 ± 0.20
携帯の MNP	3	2.33 ± 0.33
知事談合	3	2.00 ± 0.58
日ハム日本一	3	1.67 ± 0.33
JR 運転見合わせ	2	2.50 ± 0.50
児童虐待	2	2.00 ± 0.00
耐震偽装判決	2	1.50 ± 0.50

7.2.3 項の実験で用いた，筆者らが選択したこの時期の主要な 5 つのトレンドと比較すると，表 1 では「PC・携帯新モデル」が現れ，「知事談合」が比較的下位にあり，「佐呂間竜巻」は出現していない．筆者らは，PC や携帯のニュースは常時継続的に観測できるため特筆すべきトレンドとして考えていなかったが，被験者の一部は「秋冬の新モデルの発表」というトレンドがあるととらえたことによる．「知事談合」が低い理由は，質は良いが小規模であったこと，また，クラスタのラベルが「佐藤」など，より詳しい表示を見ないと分かりにくかったことが理由として考えられる．「佐呂間竜巻」が出現しなかったことについては，小規模であったこと，また，定常的な気象関係のニュースの一部と見なされたことなどによると思われる．

今回の実験については，被験者がその場で初めてシステムを利用したこと，平均して十数分程度の利用での回答であること，また，実験を行った時点（2007 年 5 月）が対象の期間の半年後であり，事件などの記憶が薄れていることなどを考えると，条件としてはあまり良い状況ではなかった．しかし，「病気腎移植」，「次世代ゲーム機」などのトレンドに関しては，特に事前の知識が与えられなくても，ユーザによる検出が可能であったことは，本システムの有効性を裏付けるものであるといえる．

7.2.5 他の可視化方式との比較

2 章で述べたように，提案した T-Scroll システムとまったく同一の目的や機能を持つ他のシステムは存在していないが，ある側面を取り出して比較することは可能である．特に本研究では，トピックやトレンドに

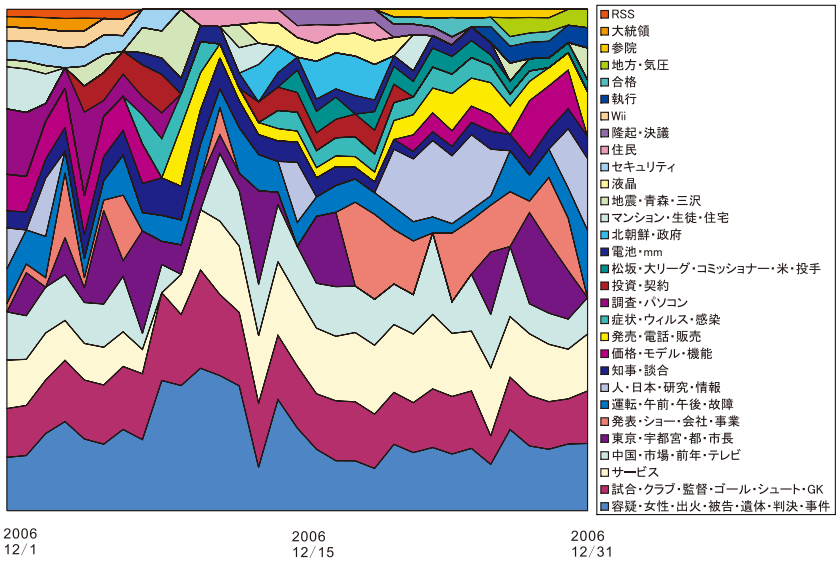

図 12 代替的な手法による可視化の例
Fig. 12 Visualization example using alternative method.

関する情報をインタフェース上に可視化する点に着目していることから，時系列文書データの可視化システムである ThemeRiver [4] における可視化方式が比較の候補となる．ただし，ThemeRiver は公開されていないため，実システムどうしの比較は可能ではない．そこで本項では，ThemeRiver のアプローチをヒントとした可視化手法の代替案を考え，T-Scroll の可視化方式と比較する．

この実験では，2006 年 12 月におけるクラスタリング結果を，T-Scroll で可視化した場合と代替案の方式で可視化した場合の比較を行う．まず，2006 年 12 月の記事集合に対する T-Scroll の表示で利用したデータもとに，以下の手順で**図 12** に示すような代替案の表示例を作成した．

(1) まず，それぞれの日における各クラスタについて，表示されているラベルとクラスタ内の文書数を調べる．同一のラベルが複数ある場合は文書数を合計する．

(2) あるトピックに関連が深いラベルが複数ある場合（例：「容疑」，「事件」），それらを統合してラベルのリストとする．関連が深いラベルの統合においては，著者らによりラベルを統合してよいかの判断を行った．この結果，30 個のラベルのリストを構築した．

(3) 上記の結果を，Excel のグラフ機能を用いて表示した．見やすさのために試行錯誤を行った結果，各時点について，個々のラベルのリストに対応する文書数の平方根の値の割合に基づいて表示の幅を決めた．平方根をとる理由は，小規模なトピックもインタフェース上で確認できるようにするためである．

なお，上記のステップ (2) では，人手による多大な労力と判断が必要であるため，可視化処理を実際に実現するには困難がともなうと考えられる．

表 2 質問 1 の結果
Table 2 Result of question 1.

分類	人数
T-Scroll の方が良い	2
T-Scroll の方がやや良い	3
どちらともいえない	3
代替案の方がやや良い	3
代替案の方が良い	0

7.2.4 項の実験に参加した 11 名の被験者に対し，図 12 を提示し，以下の 2 つの質問を行った．

7.2.5.1 質 問 1

この質問では，大まかなトレンドをとらえるためにはどちらの表示方式が有効と思われるかを尋ねた．ただし，比較は表示の仕方のみに限定し，キーワードリストの表示など，T-Scroll の付加機能については考慮しないものとした．

表 2 に結果を示す．おおむね T-Scroll の方が良いという結果が得られた．なぜそのように回答したかという理由を尋ねたところ，代替案の方はぱっと見た印象は良いものの，あまりに大まかすぎて情報が把握できない，色が煩雑であり分かりにくいといったコメントが得られた．ただし，両者とも見にくいという意見

表 3 質問 2 の結果
Table 3 Result of question 2.

分類	人数
よく確認できる	0
ある程度確認できる	3
あまり確認できない	4
まったく確認できない	4

もあり，情報の可視化の手法としてはどちらも改善の余地があるといえる．

7.2.5.2 質 問 2

2006 年 12 月の文書群に対し，T-Scroll を用いて筆者らが観測できたトレンドである「火事相次ぐ」，「松坂大リーグ移籍」，「ノロウィルス蔓延」，「知事汚職問題」，「次世代ゲーム機」のトレンドが，図 12 の表示において観測できるかを被験者に尋ねた．ただし，個々のトレンドが観測できたかを個別には尋ねず，全体としての感想を尋ねた．

結果を**表 3** に示す．「あまり確認できない」，「まったく確認できない」が半数以上を占めた．評価が低い理由の 1 つとして，図の右に示された凡例に「Wii」，「松坂・大リーグ・コミッショナー・投手」，「症状・ウィルス・感染」という関連するエントリが見られるが，図中でその領域を見出すことが困難であるという意見が複数あった．図の下の方に表示されている面積が大きい領域は，継続的に見られ文書数が多い，事件などのトピックに対応しており，一方で，「松坂大リーグ移籍」など，ある時点のトレンドを表す領域はわずかな面積しか占めず，観測することは容易ではない．すなわち，トレンドを構成する文書群が文書全体に対する割合は一般に低いため，大量に存在する他の文書群に埋もれてしまうことになる．

図 12 のような表示は，観測したい少数のトレンドをあらかじめ吟味して設定し，それらのトレンドに合った文書群を選定して表示に用いる場合には適用可能であると考えられる．しかし，事前の人手による作業などを特に行わず，大量の時系列文書の中からトレンドを把握したいという本研究の目的に対しては，有効なアプローチではないといえる．ThemeRiver [4] も，大まかには図 12 と同様の特徴を有していることから，同様な欠点を有するといえる．

比較のため，T-Scroll を用いて，表 1 と同じ方法で，対象期間のみを 2006 年 12 月に変えて行った実験結果を**表 4** に示す．1 件のみ報告されたトレンドについては省略している．なお，この実験は上記の質問 2 を実施する前に行ったものであり，被験者には事前の情報は与えていない．これは T-Scroll の機能を十分に活用した場合の結果であるため，質問 2 のアンケート結果とは直接比較はできないが，T-Scroll においてはトレンドの観測がより容易であることを裏付けるものといえる．なお，トレンド「次世代ゲーム機」については，表には現れていないが 1 件の報告があった．「火事相次ぐ」については特に報告はなかった．事件一般の話題の一部であると判断されたと考えられる．

表 4 観測されたトレンド（2006 年 12 月）
Table 4 Observed trends (December 2006).

トレンド	回答者数	スコア平均（標準誤差）
ノロウィルス蔓延	7	2.00 ± 0.22
松坂大リーグ移籍	6	2.00 ± 0.26
知事談合	6	1.33 ± 0.21
携帯電池破裂事故	3	2.00 ± 0.58
車両脱線事故	2	2.50 ± 0.5
次世代ゲーム機	3	2.50 ± 0.5
青島幸男死去	2	2.00 ± 1.00
住基ネット判決	2	1.50 ± 0.50
東京モノレール故障	2	1.00 ± 0.00

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し，トレンドを把握するためのインタフェースである T-Scroll システムの概要，アイデア，実現手法，評価について述べた．本システムは新規性に基づく文書クラスタリング手法に基づいており，クラスタリングの出力を利用することで可視化情報を作成する．T-Scroll では対話的処理も支援しており，キーワードリストの表示や元文書へのアクセスなど，さまざまな機能を提供している．

ユーザによる評価では，実際に発生したニュースのトレンドが，T-Scroll インタフェース上で観測できたことが示された．これにより，システムの目的である，時系列文書におけるトレンドをとらえる機能については実現できたものと考える．

ただし，システムにはいまだ改善すべき点が多い．実験の被験者からのコメントにも見られたが，インタフェースの使いやすさ，分かりやすさについてはさらに改善が必要である．特に，クラスタ上に表示するキーワードが必ずしも適切なものでないという問題点がある．これについては，形態素解析に用いる辞書の拡充などでもある程度対応できるが，良質のキーワード抽出にはさらなる工夫が必要であると考えられる．

謝辞 貴重なご意見をいただいた査読者および担当編集委員の皆様に深く感謝いたします．また，実験を補助していただいた名古屋大学データベースグループの飯島裕一氏，飯田卓也氏の両名にも厚く感謝いたします．本研究の一部は，日本学術振興会科学研究費

（19300027）および放送文化基金の助成による．

参 考 文 献

1) Allan, J. (Ed.): *Topic Detection and Tracking: Event-based Information Organization*, Kluwer (2002).
2) Han, J. and Kamber, M.: *Data Mining: Concepts and Techniques*, 2nd edition, Morgan Kaufmann (2005).
3) 長谷川幹根，石川佳治：T-Scroll：時間的トピックの推移をとらえる可視化システム，電子情報通信学会第 18 回データ工学ワークショップ（DEWS2007）(2007).
4) Havre, S., Hetzler, E., Whitney, P. and Nowell, L.: ThemeRiver: Visualizing Thematic Challenges in Large Document Collections, *IEEE Trans. Visualization and Computer Graphics*, Vol.8, No.1, pp.9–20 (2002).
5) Ishikawa, Y., Chen, Y. and Kitagawa, H.: An On-Line Document Clustering Method Based on Forgetting Factors, *Proc. 5th European Conference on Research and Advanced Technology for Digital Libraries* (*ECDL 2001*), LNCS, Vol.2163, pp.332–339 (2001).
6) Kalnis, P., Mamoulis, N. and Bakiras, S.: On Discovering Moving Clusters in Spatio-temporal Data, *Proc. Int'l. Symp. on Spatial and Temporal Databases* (*SSTD'05*), LNCS, Vol.3633, pp.364–381 (2005).
7) Khy, S., Ishikawa, Y. and Kitagawa, H.: Novelty-based Incremental Document Clustering for On-line Documents, *Proc. Int'l. Workshop on Challenges in Web Information Retrieval and Integration* (*WIRI 2006*) (2006).
8) Khy, S., Ishikawa, Y. and Kitagawa, H.: A Novelty-based Clustering Method for On-line Documents, *World Wide Web Journal* (2007). (to appear).
9) Kontostathis, A., Galitsky, L.M., Pottenger, W.M., Roy, S. and Phelps, D.J.: A Survey of Emerging Trend Detection in Textual Data Mining, *Survey of Text Mining: Clustering, Classification, and Retrieval*, Berry, M.W. (Ed.), chapter 9, pp.185–224, Springer-Verlag (2003).
10) Mei, Q. and Zhai, C.: Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining, *Proc. ACM KDD*, pp.198–207 (2005).
11) Müller, W. and Schumann, H.: Visualization Methods for Time-dependent Data: An Overview, *Proc. 2003 Winter Simulation Conf.*, pp.737–745 (2003).
12) Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y. and Schult, R.: MONIC: Modeling and Monitoring Cluster Transitions, *Proc. ACM KDD*, pp.706–711 (2006).
13) Swan, R. and Allan, J.: Automatic Genration of Overview Timelines, *Proc. ACM SIGIR*, pp.49–56 (2000).
14) 角谷和俊，松本好市，高橋美乃梨，上原邦昭：マルチチャ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，情報処理学会論文誌：データベース，Vol.43, No.SIG 5(TOD 14), pp.87–97 (2002).

(平成 19 年 3 月 20 日受付)
(平成 19 年 7 月 6 日採録)

(担当編集委員　池田 哲夫)

長谷川幹根

2007 年名古屋大学工学部電気電子・情報工学科情報工学コース卒業．在学時は情報検索のためのインタフェースに関する研究を行う．現在，日本製粉（株）に勤務．

石川　佳治（正会員）

1989 年筑波大学第三学群情報学類卒業．1994 年同大学大学院博士課程工学研究科単位取得退学．同年奈良先端科学技術大学院大学助手．1999 年筑波大学電子・情報工学系講師．2004 年同助教授．2006 年名古屋大学情報連携基盤センター教授．博士（工学）（筑波大学）．データベース，データ工学，情報検索等に興味を持つ．日本データベース学会，電子情報通信学会，人工知能学会，ACM，IEEE CS 各会員．