6T-1

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の開発

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

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など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いる [1]. 大量のテキスト情報を内容に基づきグループ化する手法として文書クラスタリング手法が存在するが，時々刻々と配信される時系列的な文書データに適したクラスタリング手法についての情報の要約と提示に関する研究に関しては，新たな技術の開発が求められている．

本稿では，文書クラスタリングシステム上に構築を行っている，トピックの推移を可視化するインタフェース **T-Scroll**(Topic/Trend-Scroll）について，その概要を述べる．

2　新規性に基づく時系列文書のクラスタリング手法

ニュース記事のような時系列的な文書（以下では**時系列文書**と呼ぶ）を考えると，文書の価値は，それが入手された時点から時間が経過するにつれ，一般には低下していくと考えられる．よって，時系列的な文書データを対象としたクラスタリングでは，新規性の高い文書データの影響力をより重視してクラスタリング結果を生成するようなクラスタリング手法が有用であると考えられる．[2] において提案・開発された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計算を行っている．これにより，文書が古くなればなるほど，それがクラスタリングに与える影響が小さくなり，その結果として新規のトピックを中心としたクラスタリングが行われることになる．

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，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進めているのが，以下で述べる T-Scroll システムである．

Development of a Trend Visualization System Based on Time-series Document Clustering
Mikine Hasegawa†, Yoshiharu Ishikawa††
† Department of Infomation Engineering, School of Engineering, Nagoya University
†† Infomation Technology Center, Nagoya University

3　T-Scroll 実装システムの概要

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) システムには主として以下の特徴がある．

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．
2. 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．

3.2　実装システムの概要

図 1 に，T-Scroll システムのインタフェースを示す．図は，10 月 1 日から 1 週間刻みで 10 月 20 日までのクラスタの流れの一部を表示している．

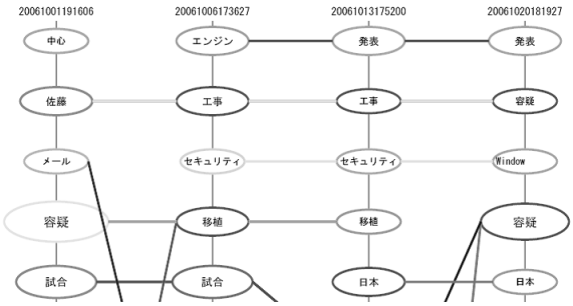

図 1: T-Scroll インタフェース画面

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．スコアは，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．また，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．さらに，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．

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

$$csim(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i|} \quad (1)$$

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．

3.3 実装システムの機能

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1のようなインタフェース画面を得られる。

図1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．また，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．以上の機能の実行の様子を図2に示す．

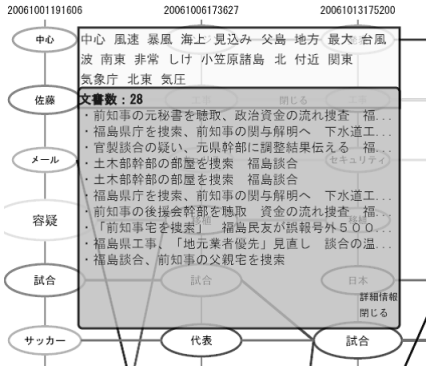

図 2: T-Scroll のインタフェース画面

4 システムの実装

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 [2] 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 XML ファイルとして出力する．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1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．

5 まとめと今後の課題

本稿では，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 システムの概要について述べた．現在，システムの評価を進めている段階であり，今後の評価をもとにその有効性をさらに明らかにしていきたいと考えている．

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．

参考文献

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

[2] 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.