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

2X−05　石川佳治† 梶並千春†† 北川博之†
†筑波大学 電子情報・工学系　††筑波大学 第三学群 情報学類

1. はじめに

近年，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，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，オンライン情報処理の重要性がさらに増してきている．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研究はほとんど見られず，新たな技術の開発が求められている．

そのため，本研究グループでは，オンラインニュース記事のような時系列文書を継続的にかつ効率的にクラスタリングするための文書クラスタリング手法の開発を進めてきた [1, 2]．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く．
2. 再クラスタリングのための低コストのアルゴリズムの使用，およびインクリメンタルな更新により，文書データが追加された際の更新コストを削減する．
3. 十分古くなった文書を自動的にクラスタリングの対象から削除する．

本稿では，上記アプローチに基づく文書クラスタリングモジュール上に構築を行っている視覚的インタフェース **ClusterFlow** について，その概要を述べる．このシステムは，継続的に時系列文書をクラスタリングした結果を視覚的に提示し，ユーザによるトピックの分析や追跡を容易にすることを目的としている．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示すことで，トピック（各クラスタで表現される）の流れを表す点にある．

以下ではまず，この **ClusterFlow** システムの基盤となるクラスタリング手法について説明を行う．次いで，**ClusterFlow** システムの機能およびその設計内容について述べる．

2. 新規性を考慮した文書クラスタリング手法について

ニュース記事のような時系列的な文書データを考えると，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．我々の研究グループで研究を進めてきた文書クラスタリング手法 [1, 2] 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．

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．

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

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

$$tdw = \sum_{i=1}^{n} dw_i$$

で与え，文書 d_i の文書集合中での生起確率を

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

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 [1, 2]．その一般形は

$$sim(d_i, d_j) = \Pr(d_i)\Pr(d_j)\frac{\vec{d}_i \cdot \vec{d}_j}{len_i len_j}$$

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．

1節で述べたように，時系列的な文書のクラスタリングには，新たに文書が追加された際のクラスタリング結果の更新コストが小さいことが求められる．[1, 2] で提案した手法は，文書追加の際の統計情報（文書類似度の計算に利用する）の更新コストが文書数，総単語数について線形オーダである点，そして，再クラスタリング処理が Scatter/Gather アルゴリズム [3] をベースとしており，

ClusterFlow: A Visual Interface for Topic Tracking of On-Line Documents
Yoshiharu Ishikawa†, Chiharu Kajinami‡, and Hiroyuki Kitagawa†,
† Inst. of Inf. Sci. and Elec., Univ. of Tsukuba
‡ College of Inf. Sci., Univ. of Tsukuba

文書数に対し線形であるという利点を有している．

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進めているのが，以下で述べるClusterFlowシステムである．

3．ClusterFlowシステムの概要

本研究で開発を進めているClusterFlowシステムの特徴は，主に以下の3点である．

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. 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表示することで，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．
3. 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．これは，OLAP (On-Line Analytical Processing) におけるドリルダウン／ロールアップ機能に対応づけることができる．

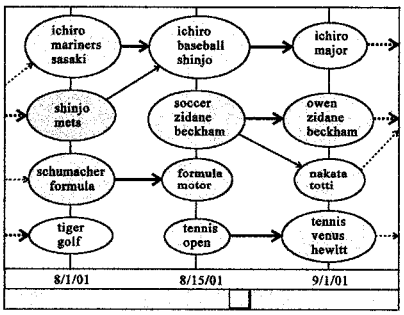

図1　ClusterFlowのGUIの概念図

図1に，ClusterFlowのインタフェースの概念図を示す．図は，2001年の8月1日から2週間刻みで9月1日までのクラスタの流れを表示している箇所を示している．例としては，スポーツ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．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は，クラスタ代表に含まれる単語で$tf \times idf$に相当する重み付け値が大きいものを選択して表示する．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重み (dw_i) の総和に対応しており，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

$$csim(C_i, C_j) = \frac{\sum_{d_l \in C_i \cap C_j} dw_l}{\sum_{d_l \in C_i \cup C_j} dw_l}$$

という式により定義する．この式は，文書類似度の定義などで用いられるJaccard係数 [4] を，クラスタ中の文書の重みを考慮して拡張したものである．この定義により，新規性の高い文書を共通して含むようなクラスタほど類似度が高くなることになる．クラスタ間の類似度が大きいものはより太いリンクで表現し，関連の深さを表現する．また，一つのクラスタから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ClusterFlow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．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ClusterFlowシステムの概要について述べた．現在，この実装と機能の詳細化を進めている段階である．

謝辞

本研究の一部は日本学術振興会科学研究費基盤研究(B)(12480067)，奨励研究(A)(12780183)，および文部科学省科学研究費特定領域研究(C)(13224008)による．

参考文献

[1] 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 Digital Libraries (ECDL 2001)*, LNCS 2163, pp. 325–339, Sept. 2001

[2] 石川佳治，北川博之：忘却の概念に基づくインクリメンタルな文書クラスタリング手法，情報処理学会研究報告，2001–DBS–125(I), pp. 313–320, 2001年7月．

[3] D.R. Cutting et. al: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, in *Proc. of ACM SIGIR Conf.*, pp. 318–329, Jun. 1992.

[4] C.J. van Rijsbergen, *Information Retrieval*, Butterworths, London, 1979.