話題の変化を捉えるニュース記事クラスタリング手法

小島 寛樹　　亀井 清華　　藤田　聡
（広島大学 大学院工学研究科 情報工学専攻）

概要　ストリーミング配信されるニュース記事を，話題の変化に応じて動的にクラスタリングする手法を開発した．この手法を用いることで，たとえば事件発生から数ヶ月間の動向をカテゴリの変化として俯瞰的に捉えることなどが可能となる．

◎ 背景と研究目的

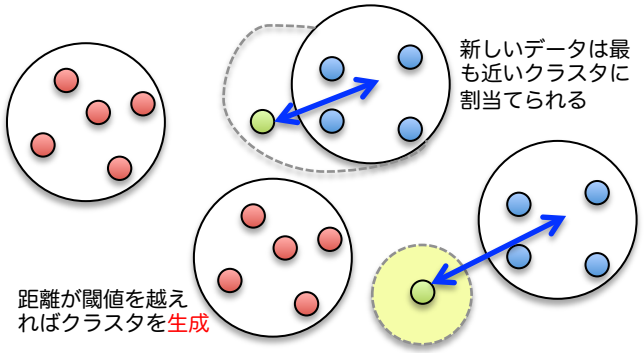
興味のあるカテゴリのニュースのみを読みたい．
e.g., Gunosy, SmartNewsなどの記事配信サービス

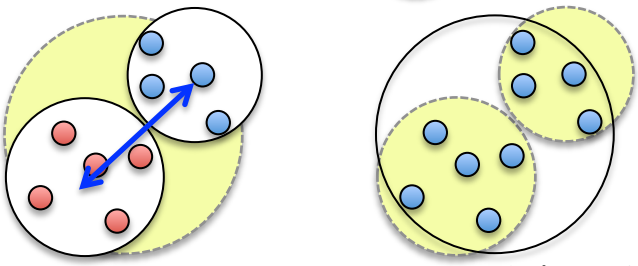
・カテゴリ分けそのものを話題の流れにあわせて変化させたい．
・ある重大事件のニュースが時間経過に伴ってどのように枝分かれしたのかを俯瞰したい．

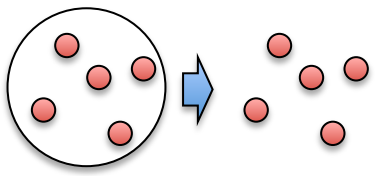
「ストリームクラスタリング」の手法を応用することでニューストリームの適応的なクラスタリングがおこなえないか．

◎ KomkritらのE-stream (2007)

・身長や体重などの数値データが対象
・数値データはストリームとして与えられるものと仮定







各データには経過時間とともに減少する重みが与えられており，重みの総和が閾値を下回った時点でクラスタは消滅

◎ 提案手法の概要

クラスタの更新：各クラスタの中心や重みをクラスタに含まれているニュース記事の特徴ベクトルなどから再計算．

クラスタの消滅：E-streamに準じる．

クラスタの分割：閾値を超えるサイズのクラスタに対してk=2のk-clusteringを実行→得られたクラスタの中心間の距離が閾値を超えていれば分割．

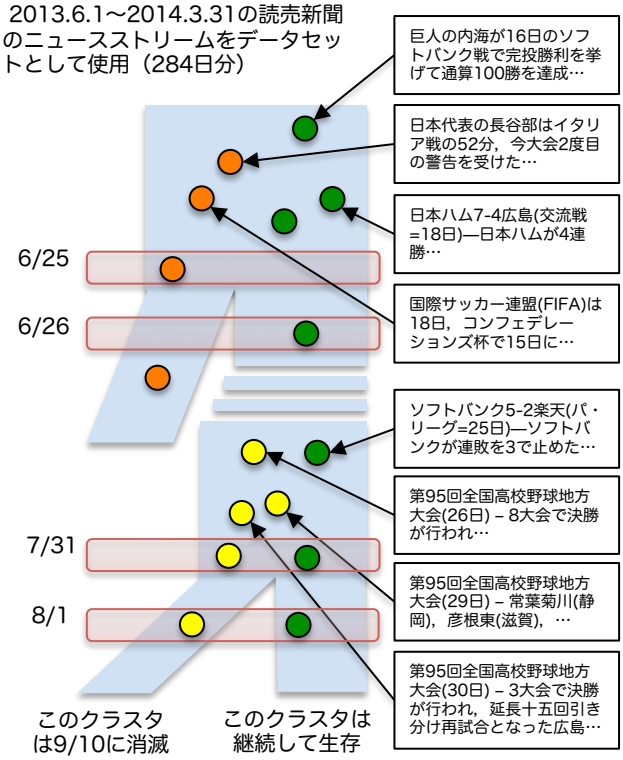
クラスタの併合：中心間の距離が閾値を下回ったとき実行．

非アクティブクラスタの特定: 長い間記事の割当てがないものは非アクティブとし，記事割当ての対象からはずす．

クラスタへの記事の割り当て：E-streamに準じる．

◎ 実行結果の一部

2013.6.1～2014.3.31の読売新聞のニューストリームをデータセットとして使用（284日分）



◎ 今後の展開

ニューストリームのクラスタリング結果を可視化するブラウザを開発し，βテストをおこなう．ブラウザには個人化機能を組み込み，各ユーザの閲覧履歴に応じて興味のある特定の分野に関連するクラスタを細分化するなどの改良を加えていく．