様式 C-19

科学研究費補助金研究成果報告書

平成　21 年　6 月　1 日現在

研究種目：若手研究（B）
研究期間：　2007　～　2008
課題番号：19700143
研究課題名（和文）　WWW における話題の分岐収束過程の抽出と可視化に関する研究

研究課題名（英文）　Tracing and Visualizing Topic Forks and Mergers on the WWW

研究代表者　森　幹彦　（MORI MIKIHIKO）
京都大学・学術情報メディアセンター・助教

研究者番号：70362423

研究成果の概要：
これまで注目していなかった話題について新たに興味を持った場合，どのような経緯をたどったのかを知りたくなることも多い．本研究は，常に新規の情報が提供される WWW サイトのニュース記事を収集し，それを話題ごとに自動分類できる方法を提案した．このとき，日々少しずつ変化していく話題の遷移を自動的に追跡するだけでなく，以前は複数の話題として語られていた内容が統一的な問題として語られる場合や，逆にひとつの小さな話題がいくつもの大きな話題に分かれた場合の認識も可能である．最後に，これらの話題の遷移を時系列で表示する可視化法を提案した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	2,000,000	0	2,000,000
2008 年度	1,300,000	390,000	1,690,000
年度			
年度			
年度			
総　計	3,300,000	390,000	3,690,000

研究分野：情報検索，人工知能
科研費の分科・細目：情報学・知能情報学
キーワード：文書クラスタリング，テキストマイニング，時系列可視化，トピック抽出

１．研究開始当初の背景

World Wide Web (以降，Web と呼ぶ)の利用者も情報取得の手段として Web への依存が深まっている．そのような利用者は，様々な事件やイベントに関して調べるとき，これまでの経緯，現在の状況，今後の展開を Web から抽出したいという要求を持つ．

従来，事件などの全体像を知りたいという利用者の要求に対して，Google などの検索エンジンにキーワードを入力して提示される Web 文書群から前後関係を読み取り，手作業または頭の中で関連づけの作業を利用者が行う方法しかなかった．したがって，特定の事件に途中から興味を持った者にとって，大きな事件の全体像を掴むことは難しく，事件の初期から注目している者にとっても，後から系統的に思い起こすのが困難である．

このような目的で Web 検索をする利用者には次のような要求ということができる：

・特定の話題の時間的変遷：特定の話題に注目して，時間遷移による話題の内容の変遷を知りたい．
・話題の分岐・収束：内容の変遷として話題の分岐・収束の様子を知りたい．

すなわち，以下のようなシステムによる情報提示が求められている：

・自動的に話題ごとに Web ページを分類する．ただし，逐次最新情報が追加される環境を想定するため，時系列を無視した一括での分類はしない．
・話題の内容が新たな情報が追加されるなどによって時間とともに変化しても同一の話題として認識できなければならない．また，ひとつの話題がいくつかの話題に広がっていく分岐過程や，複数の話題のように見られていたものがひとつの話題にまとまる収束過程をも発見したい．
・上述のような話題の変遷について，話題の内容やその変化が一覧できるような可視化による表示がほしい．

２．研究の目的

本研究は，話題を文書のクラスタと考えて，以下のような要件を満たす文書クラスタリング手法を提案する．

・話題の時間的な変化に左右されない．
・話題が分岐や収束したとき，それを検知し分岐前後の話題のつながりを維持できる．
・現実の話題の数は変化するため，動的に話題数を決められる．
・文書が新たに追加され続けるため，追加されてもその分だけに適用できる．

さらに，本クラスタリング手法によって生成されたクラスタを表示するため，以下のような要件を満たす可視化法を提案する．

・話題が変化したときにそれが以前と異なることがわかる．
・話題が分岐や収束をするときに，分岐や収束の前後の話題の関係がわかる．
・話題に含まれる文書が一覧できる．また，その文書の内容が見られる．

３．研究の方法

(1) 動的クラスタリング法

本研究では，逐次追加更新される Web 文書としてオンラインニュースサイトのニュース記事を対象にして，そのニュース記事で記述された内容をもとにしたクラスタリング手法を開発した．クラスタリングのために，文書間の類似度と文書間の距離の 2 つの評価基準を用いた．

次に述べる評価値と操作手順を用いてクラスタの分割と合併を行い，その過程を追跡することにより話題の分岐と収束などの変化の過程を追跡することとした．

① 文書のベクトル化

文書を単語の集合として扱い，ある文書 d_i は，そこに含まれる名詞を用いて文書ベクトル $d_i=(w_{i1},...,w_{ij},...,w_{in})$ として表す．ここで，w_{ij} は i 番目の記事における j 番目の単語の頻度を利用し，d_i が単位ベクトルになるよう正規化した．

② 文書の類似度

発表時期の離れた文書間の類似度は，近い時期に書かれた文書同士の類似度よりも小さくなるとする．そこで，期間をまたがる類似度は忘却を考慮する．

文書群を一定期間ごとに分割し，それぞれの期間を t とする． i 番目の文書が期間 t に現れ，j 番目の文書が期間 $t+a$ に現れたとき，i 番目の文書と k 番目の文書の類似度 $sim(d_i,d_k)$ は，文書ベクトルの内積を正規化した次の式で算出する．

$$sim(d_i,d_k)=\lambda^a \frac{d_i \cdot d_k}{\sqrt{|d_i||d_k|}}$$

ここで，λ は忘却係数を表し，$0<\lambda\leq 1$ である．

③ クラスタの重心

話題クラスタ D_i の重心 C_i は，D_i の文書数を $|D_i|$ として次のように表せる．

$$C_i=\frac{1}{D_i}\sum_{j=1}^{D_i} d_{ij}$$

これは後述する k-means 法における定義と同様である．

④ クラスタの形状

クラスタの大きさは様々であるため，単純にクラスタの重心間の類似度によってクラスタを分割や併合の対象とするかの評価はできない．そこで，クラスタの分散を用いることとした．

文書ベクトルの次元をもとに，あるクラスタに対して重心を平均とみなし，各次元について重心からの距離を共変動とみなして共分散を用いる．

⑤ クラスタの分割条件

2 つの話題クラスタ D_x と D_y の分割条件は次の式で表される．

$$\sigma_{Dx} + \sigma_{Dy} < \frac{|C_x - D_y|}{b}$$

ここで，σ_{Dx}，σ_{Dy}は2つのクラスタの重心間の方向における分散の平方根であり，重心間の方向における偏差とみなせる．また，bは任意の係数で，文書群の偏在性に依存して手動で決定する．

⑥ 話題クラスタの生成

本クラスタリング法では，文書が追加されるごとに既存クラスタの類似性をもとに逐次的に処理をする．

まず，文書を時系列に直列に並べた文書列から古い順に文書を取り出す．i番目の話題クラスタD_iにおけるj番目の文書をd_{ij}とする．文書d_{ij}が期間tの文書であるなら，$j < k$であるd_{ik}は期間tもしくはそれ以降の期間の文書とする．いま新たに取り出した文書を$dnew$とすると，$dnew$の所属を決めるために各D_iのd_{ij}に対して前述の類似度を求める．もし，$s(dnew, d_{ij}) > \theta$が成り立つなら，そのd_{ij}が所属するD_iに$dnew$も所属するとする．ここで，θは閾値で$0 < \theta \leqq 1$の範囲とする．ある文書が2つの話題クラスタD_xとD_yのどちらに対してもθより大きいなら，D_xとD_yを併合する．もし，どのクラスタに対してもθ以下であったなら，その文書のみを持つ新たなクラスタを生成する．

文書が追加されるたびに分割を試み，分割条件を満たす場合にはそのクラスタを分割し別のクラスタとする．

文書群の分割には様々な手法が提案されているが，本研究では，k-means法を適用することにする．分割数は2とし，分割後のクラスタに対して併合の式で評価する．

最後に，長い期間，すなわち期間tから期間$t+\tau$（τは任意の定数）までに新しい文書が追加されない話題クラスタを終了した話題とみなす．忘却を考慮した類似度の計算によって，話題の終了を定義しなくても新規の文書は追加されないため，実質的には問題にならない．

(2) 可視化法

① クラスタの表示

本可視化法では，2次元平面上に横軸を時間軸としてとる．期間を1ずつ進めたときのそれぞれの期間を右方向に表示していく．期間tにおける話題クラスタは，縦方向に表示する．可視化表示上でクラスタは，楕円として描かれ，内部にクラスタ番号とそのクラスタを代表する2単語を表記する．表示する2単語はクラスタに入った文書の単語頻度を算出して最も頻度の高いものとした．

継続的に文書の加わっているクラスタと文書の追加がなくなったクラスタを区別するため，期間tで新たに作成されたクラスタと期間tで文書が加わったクラスタのみを期間tの位置に表示する．すなわち，$t-1$から継続している話題クラスタで，別の期間に新たな文書が加わったなら，再度その期間にも表示する．

期間の異なる同一のクラスタ間を線で結ぶ．また，分割や併合によって分かれた別のクラスタに対しても同様に線で結ぶ．つまり，期間tで1つだったクラスタが期間$t+1$で2つのクラスタに分割された場合，期間tのクラスタから2本の線で期間$t+1$のクラスタをそれぞれ結ぶ．

ここで注意すべきは，クラスタの分割のタイミングである．ある文書によっていくつかのクラスタが併合されても次のステップで分割が試みられる．この段階で要素の異なる別のクラスタに分割されることが起こりうる．もちろん，パラメタの設定によっては同じクラスタに戻ることもある．したがって，可視化表示上では，隣り合う期間にあるクラスタ同士を相互結合するように線がつながることが起こり得る．

② クラスタの内容表示と文書表示

クラスタ間の関係だけでなく各クラスタの内容や実際の記事を表示するため，画面上を分割して表示する．

４．研究成果

(1) 可視化表示システム

本クラスタリング手法を用いた結果の可視化の表示画面の例を図1に示す．本表示例ではYahooニュースの2009年2月1日から2月7日までの3563件ある記事を用いている．

Webブラウザ上で閲覧できる．左側に表示されたグラフにおいて各ノードはクラスタを表す．縦に一列に表示されているノードは同じ期間のクラスタである．ノード間が線で接続されているということは，左側のクラスタが分割や併合，文書の追加によってクラスタが変更されたことを表す．ノードをクリックすることにより，そのクラスタに含まれる文書のタイトル一覧が右上に表示される．さらにそのタイトルをクリックすることによって文書の内容を見ることができる．

他のノードと接続のないクラスタを表示しないことにより，再現率は落ちるが閲覧は容易になることがわかった．今後は，あるグラフ上のノード数による基準で表示するノードを減らす工夫が必要であろう．

(2) パラメタの性質

本クラスタリング手法では，類似度しきい値θ，忘却係数λ，クラスタ間距離係数b，話

題終了定数 τ の任意で決められる 4 種類のパラメタがある．これらのパラメタの特徴は定性的に次のようなものであった．

① 類似度しきい値 θ

小さくするほどクラスタを併合させる作用がある．b の大きさによってはすぐに分割されるが，その結果として多くのクラスタとつながりを持たせ，可視化表示上で多数の線を表示させることになる．

可視化表示の見易さの観点からすると，θ は最も効果を及ぼす．少しだけ類似していても線を結ぶためにクラスタ間の関係が読み取れなくなる．

② クラスタ間距離係数 b

大きくするほどクラスタを併合させる作用がある．しかし，非常に類似性の高い密な分布をしている場合を除き，分割を引き止める作用しかならない．

③ 忘却係数 λ

大きくするほどクラスタを併合させる作用がある．新規の文書を比較的適切なクラスタに日も付けることにより，1 文書のみの新規クラスタの生成を抑える．

④ 話題終了定数 τ

大きくするほど過去のクラスタと接続できるように探索される．ただし，あくまでも接続されている場合にそれを接続として扱うかを決める値である．すなわち，λ で算出できる時間内に新規の文書がクラスタに追加されない場合は，そのクラスタが存在し続けず接続していないため，τ を大きくとっても変化はない．$\lambda^a \geq \tau$ のときに有効である．

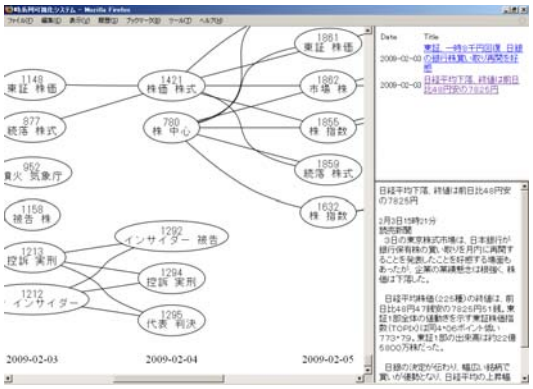

図 1 文書群の話題変化の可視化表示

(3) 提案クラスタリング手法

本研究のクラスタリング手法は従来の文書クラスタリングの手法と比較して次のような特長がある．

① クラスタ数を動的に変えられる．

② 時間情報を持つ文書が逐次投入されるという前提を生かして，差分のクラスタリングが可能である．

③ 内容が少しずつ変化した場合にも 1 つのクラスタとして追跡することができる．

④ クラスタの変化はクラスタ自体に起きるため，隣接期間での同一クラスタの推定が必要ない．すなわち，期間ごとのクラスタリング結果からクラスタ間の類似度等を求めるような手法でない．

本クラスタリング手法は，クラスタの分割と併合が起きやすい．そのため，非常に密接な関係を持つクラスタ群が平行して存在し続けることがあった．たとえば，図 1 では左下のあたりに描かれた関係である．長期にわたっていくつかのクラスタが交差した状態で続いていることがあった．本クラスタリング手法の特性から，クラスタ同士が併合することはない．したがって，新規の類似した記事がコンスタントに間を取り持つようにしてクラスタの併合と分割を繰り返させることになったと考えられる．

このような密結合のグラフが描かれるときは，クラスタの分割併合は話題の分岐収束とは異なる状態であることも多い．しかし，クラスタ間の関係として読み取れるため，クラスタ間の関係を類似度などのみで測る手法と比べると，異なるクラスタであると認識することは少ない．さらに，クラスタ間のグラフを大きな視点で見ると大枠の流れを総覧することはできる．今後は，表示上は分かれていても大きなクラスタと読み取ることができる．今後は，密結合のノードを仮想的に一つのノードとして表示できるような改善が考えられる．

各クラスタに含まれる文書の適合率は，従来研究で行われた k-means 法を用いたものと大きな違いはなかった．前述のパラメタの特性を用いてクラスタの大きさを小さくするように設定すると適合率は向上するが，本来は 1 つの話題のものが複数のクラスタに分散してしまい再現率を下げた．

過剰に併合・分割を繰り返さないような θ に設定すると，生成されるクラスタの多くが文書を 1 件しか持たない孤立したノードになった．グラフの見易さとの兼ね合いではあるが，θ を低くした方が意味を読み取れるクラスタとその関係を表示であると判断できた．

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① 森幹彦，ニュース記事の話題分岐を時系列で追跡可能な可視化法，情報処理学会第 71 回全国大会講演論文集，査読無し，2009.

② 森幹彦，ニュース記事における時間変化する話題の抽出，人工知能学会第 22 回全国

大会予稿集，査読無し，2008.
③ 森幹彦，文書群からの時間的変化する話題の抽出，情報処理学会第 70 回全国大会講演論文集，査読無し，2008.

６．研究組織

(1)研究代表者

森　幹彦（MORI MIKIHIKO）
京都大学・学術情報メディアセンター・助教
研究者番号：70362423