マルチメディア通信と分散処理　100－6
（２０００．１１．９）

電子新聞に対する情報トラッキング

北川 結香子, 中嶋 卓雄

熊本大学 工学部

〒862-8555　熊本市黒髪 2-39-1
E-mail: {kity@st.,taku@}cs.kumamoto-u.ac.jp

新聞購読者が新聞を読む場合，キーワードを入力する検索システムのような情報の探し方はせず，またカテゴリー化された記事を中心にトップダウンに記事を読むこともしない．本研究では，購読者が新聞を読みながらある特定の内容について興味を持ち，その関連情報を効率良く辿ることができる，トラッキング手法について考察する．トラッキング手法の中で，購読者の情報要求に対して注目する視点をモデル化し注目の移動と日々の情報の変化とを対応させて情報要求の変化を考察する．また，トラッキングを実現するための種々の指標の定量化アルゴリズムを示す．

A Information Tracking for Electronic Newspaper

Yukako Kitagawa, Takuo Nakashima

Faculty of Engineering, Kumamoto University

2-39-1 Kurokami Kumamoto-shi ,860-8555 ,Japan

In the situation of normal daily newspaper reading, electronic news subscribers of Internet focus to some article and spread the attention to the same field of the past articles. We assume that this movement of intention is main information needs of people. In this paper, we show the total tracking algorithm of electronic newspaper based on the user movement of intention, and formalize the expression of evaluation parameters.

1 はじめに

ネットワークの発展，普及によりWebなどにおいても日々内容が変化し，発信される情報量が急激に増加している．特に，日常的な出版物である新聞，雑誌が電子化されるにつれて，その情報に対する分類・整理が必要とされている．

日常的に更新され，さらに新しいカテゴリーが出現する新聞記事に対しては，適切な(1)記事の分類手法，(2)記事に対する検索手法を考慮する必要がある．

まず，文章分類に関しては，単語の抽出に既存の辞書を利用しながら，(1)既存のカテゴリー情報に基づき分類する手法，(2)分類のためのカテゴリー情報を人手で抽出しシステムに学習させ，その情報に基づき分類する手法がある．しかし，新聞などのように新しい言葉が次々生まれ，そのカテゴリーの階層関係が変動する文書には適当な手法と言えない．また，(2)では，精度の高い分類が得られているが，本質的に人手を介することにより，人間の持つ知識が前提となる手法であり，内容の変動が多い大量の文章の自動分類にとって適当な手法と言えない．したがって，日毎の記事のみから特徴となる単語を抽出するするクラスタリングを考慮する必要がある．

一方，記事に対する検索に関しては，(1)キーワードを入力することによる検索，(2)すでに階層的にカテゴリー化された情報からトップダウンに検索する，ことが考えられる．しかし，キーワード検索の場合，ユーザが特定の情報に対する明確な要求を持つことが事前に要求される．したがって，ユーザは情報に対して既知でなくてはならず，既知度により検索の結果が左右される．また，大量の記事を組織化し，ユーザに提示することにより，情報の把握やユーザによる検索を容易にさせるサービスがあり，ディレクトリーサービスなどはこの考え方に基づくサービスである．これらの分類においては固定した分類項目により記事が組織化されているため記事における話題の変遷に適応させることが難しい．

本研究では，購読者が新聞を読みながらある特定の内容について興味を持ち，その関連情報を効率良く辿ることができる，トラッキング手法について考察する．トラッキング手法の中で，購読者の情報要求に対して注目する視点をモデル化し注目の移動と日々の情報の変化とを対応させて情報要求の変化を考察する．

まず，トラッキングの対象となる記事中の単語の重み付けアルゴリズムを定義する．その情報から記事間の関連度によりクラスタリングするアルゴリズムを提案する．クラスタリングの結果を利用して，(1)主題の抽出を行い，(2)クラスタの時系列による関連性を抽出し，(3)クラスタ時系列の代表単語を抽出する．(4)さらに基準単語によりクラスタを再構成し，(5)基準単語と連想単語による記事・クラスタの構造化を行う，ことによってトラッキングを形成する．

2 ユーザの視点

対象を電子化されWebとして公開されている新聞記事データとする．ユーザが目で追いながら選択する情報の単位としては，(1)関心を持っているキーワードが含まれる記事，(2)特定の記事，(3)特定の面に関連した記事全体，が考えられる．(1)では，従来のキーワード検索によりユーザの要求に答えることは簡単であり，(3)のような傾向の場合，記事データが階層的にカテゴリー化されていれば，その階層に沿ってユーザがリンクをたどればよい．しかし，通常記事データをアクセスする場合には，事前に関心を持つキーワードが存在せず，また既存の階層化された記事集合の中に分散した特定の記事に興味を持つ場合が多いと考えられる．

したがって，ある特定の記事を中心にして，そこから過去へのトラッキングを考える．しかし，記事は特定のクラスタに含まれているとするが，そのクラスタと他のクラスタとの関係は動的に変化することを前提とする．

3 記事の基本要素と関連度

トラッキングする記事の最小単位となる単語の重み，および記事間の関連度を定義する．

3.1 単語の重み

記事の特徴を表わす単語の重みが高くなるように重み付け関数を定義する．本研究では，単語の出現位置を考慮した重み付け手法を提案する．

新谷 [1] らも提案しているように，重要な単語ほど記事の先頭に出現する．また，一般的に新聞記者は読者が記事の途中で読む作業を終了しても意味が把握できるように文の順序にそって内容を記載している．

新谷らは，単語が最初に出現した文の位置に注目し重み関数を定義しているが，

- 重要と考えられる単語は、繰り返し文中に記述される可能性があるが、それが考慮されていない．
- 同じ文に複数回出現する場合の重みも考慮されていない．

ことを考慮して，本研究では，記事 d 中の単語 t_0の位置に依存した重み付け関数 $L_p(d,t_0)$ を次のように定義する．

$$L_p(d,t_0) = TFL(d,0,t_0) + \sum_{x_{t_0}=1}^{X} \frac{1.0}{x_{t_0}} * TFL(d,x,t_0) \quad (1)$$

$TFL(d,l,t_0)$: 記事 d の l番目の文章における名詞 t_0の出現回数

x_{t_0}: 名詞 t_0の出現する位置

X: 記事 d 中の本文中の文の数

だたし，見出しは位置が 0 の文と見なし，本文の位置は 1 から増加するものとする．

3.2 単語の変化量の定義

単語の場合，複数の分岐および結合の場合は存在しない．以下では同じ単語間の変化量を定義する．

文献などと同様に新聞記事においても一般的な単語は記事全般にわたって出現する．例えば,「町」とか「市」のような単語や，地方新聞の場合には，その地方都市の名前などの単語である．また,「経済」や「政府」のような単語の出現頻度も高い．

しかし，ある事件が発生した場合やイベントが開催される場合などにおいては，関連する単語が一定期間記事全体に出現するが，それらは一般的な単語と見なすのではなく，記事の特徴を表わす単語として扱うのが適当である．

記事中に頻繁に出現する単語は，(1) 記事を記述する一般的な単語，(2) 事件，イベントなどにより一定期間頻繁に出現する単語 のどちらかである．しかし，1 日の新聞記事を母集団として，頻繁に出現する一般的な単語の重みを下げる idf 手法を使えば，(1) 事件，イベントなどの重要な単語も重みが下がり，(2) あまり重要でない単語に逆に大きな重みをつける可能性がある．

事件，イベントなどに出現する単語は，(1) 一定量の記事中に集中して出現する場合が多く，(2) その頻度も時期的に増加しながら出現する傾向にある．

そこで，日々更新される新聞を 1 日を単位にして，単語の集中度の時期的な変化率を定義する．日付 d における単語 t_0の前向き変化率 $\Delta_w(d,t_0)$ を次のように定義する．

$$\Delta_w(d,t_0) = \frac{TFD(d,t_0)}{DFD(d,t_0)} - \frac{TFD(d-1,t_0)}{DFD(d-1,t_0)} \quad (2)$$

ここで，

$TFD(d,t_0)$: 日付 d における，単語 t_0の出現回数の移動平均値

$DFD(d,t_0)$: 日付 d における，単語 t_0の出現する記事数の移動平均値

とする．

3.3 記事間の関連度

tf·idf 法は記事に対して頻出単語を一般的な単語として扱う傾向があり，トレンドに関係し

た単語も一般的な単語として扱う傾向にある．

そこで，変化率を位置情報による単語の重みと混合させ，変化による記事 a 中の単語 t_0 の重み付け関数 $V(a,d,t_0)$ を次のように定義する．

$$V(a,d,t_0) = L_p(a,t_0) * \alpha^{\Delta_w(d,t_0)} \quad (3)$$

ここで，$\alpha(>1)$ はトレンドを表わす変化への重みである．

前述した重み関数が単語の重みを表すことになるので，その値を用いて，記事間の関連性を計算する．

2つの記事に関連があるか否かを記事中で共起する単語の重みのづけの総和により定義する．記事に含まれる単語数と単語の重みが異なるので，全単語数の重みの総和により正規化して，記事 d_x，d_y 間の関連度 R を次のように定義する．

$$R(d_x,d_y) = \frac{\sum_{t_{x\cap y}} V(d_x,t_{x\cap y})}{\sum_{t_x} V(d_x,t_x)} * \frac{\sum_{t_{x\cap y}} V(d_y,t_{x\cap y})}{\sum_{t_x} V(d_y,t_x)} \quad (4)$$

$$\begin{aligned} t_x &= \quad t \in s_x \\ t_y &= \quad t \in s_y \\ t_{x\cap y} &= \quad t \in (s_x \cap s_y) \end{aligned}$$

4 トラッキングアルゴリズム

4.1 クラスタリング

本研究では，記事の分類項目名の決定法および付与の仕方など，記事編集者によって変化する可能性もあり，さらに，突発的な事件・事故も多く，今までの分類には該当しない記事も多いと考え，一般的な分類項目が存在せず，その項目自身も記事から抽出しながらにクラスタリングする手法について考察する．

一般的なクラスタリング手法では記事間の関連度を平均化しながらボトムアップ的にクラスタを作成していく重心法などが利用されるが，一日の記事集合を対象としたクラスタリングにおいては、あまりおおきなクラスタにならなく，記事数が1つのクラスタが数多く生成される傾向がある．このような傾向を持つため，平均化操作が関係したクラスタリングは精度が落ちるので，特徴単語を中心にしてクラスタが進む傾向にある最長距離法によるクラスタリングを行う．

4.2 主題の抽出

1つのクラスタからクラスタの特徴を表す数個の代表単語を抽出する．特徴を表す代表単語を新聞記事に対して考えた場合，固有名詞，特に新しく出現する固有名詞が表していると考えられる場合が多い．そこで本研究では，自動的に固有名詞を抽出しその名詞に注目して代表単語を選択する．

4.3 クラスタの時系列

日単位に生成されたクラスタの時系列を抽出する．クラスタの代表単語の集合からクラスタが構成されると考え，記事間の関連度を求めた手法と同様の手法を用いる．

クラスタに含まれる単語数と単語の重みが異なるので，全単語数の重みの総和により正規化して，クラスタ c_x，c_y 間の関連度 R_c を次のように定義する．

$$R_c(c_x,c_y) = \frac{\sum_{t_{x\cap y}} V(c_x,t_{x\cap y})}{\sum_{t_x} V(c_x,t_x)} * \frac{\sum_{t_{x\cap y}} V(c_y,t_{x\cap y})}{\sum_{t_x} V(c_y,t_x)} \quad (5)$$

$$\begin{aligned} t_x &= \quad t \in s_x \\ t_y &= \quad t \in s_y \\ t_{x\cap y} &= \quad t \in (s_x \cap s_y) \end{aligned}$$

この関連度からクラスタの時系列を月単位に抽出する．一般には次に示すようなクラスタの変化のパターンが存在すると考えられる．以下のパターンを考慮して抽出を行う．

発生 :前日の情報とはまったく関連のない情報が生成された場合．

消滅 :前日の情報と関連のある情報が存在しない場合．

分岐 :前日の１つの情報と関連のある情報が複数存在する場合.

結合 :前日の複数の情報と関連度のある情報が１つ存在する場合.

4.4 クラスタ時系列の代表単語の抽出

時系列として得られたクラスタ集合から，１，２個の単語を抽出する．クラスタの時系列は少ない話題から多方面に議論が展開されていると考え，この単語数は，クラスタの主題を抽出した単語数より少ない単語数とする．この単語を基準単語とする．

4.5 基準単語に基づくクラスタの再構成

基準単語からクラスタを再構成する．この段階で単にクラスタ間の関連度により同じ時系列として扱われていたクラスタを排除する．さらに，基準単語と共起関係にあり、なおかつそのクラスタの主題として抽出された単語をそのクラスタの連想単語とする．記事に関しても同様に基準単語に基づいた解析が可能であるが、同じ日には同じ傾向を持つ，すなわち同じ連想単語となる記事がクラスタ化される可能性が高いので，記事を単位とするのではなく，クラスタを単位として構造化する．

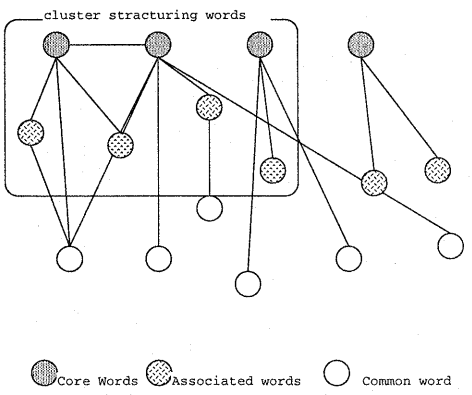

図１ クラスタ時系列による基準・連想単語の関連概念図

4.6 基準単語と連想単語による記事・クラスタの構造化

基準単語および連想単語に近い単語を，単語の出現位置に基づく意味的な近さから，階層的な単語クラスを構成する．さらにその記事の構造に基づきクラスタの構造を階層的な単語クラスにより再構成する．

4.7 クラスタ中の単語クラス構造の差分

クラスタ中の単語クラス構造の差分をとることにより，意味構造の変化を表す．その結果から，基準単語および連想単語間の時間的な有効性を表し，共起単語間の時間的な繋がりの強さを表すものとする．

5 特徴抽出実験

クラスタ時系列中の特徴単語を抽出するアルゴリズムを考えるため，簡易的な特徴抽出実験を行った．

まず，実験は 72 日間の記事データから，すでに辞書に存在する「コソボ」をキーワードとして，一旦クラスタデータをとして抽出した．この記事集合をクラスタを構成するクラスタ時系列として考えた．1，２週間の記事ではクラスタ時系列を構成はできるが，時系列から話が展開するなどの動きに対応できないため，２ヵ月以上のデータから収集することにした．

その結果，228 記事が抽出でき，その記事集合から重みをつけた単語を抽出し，時系列的に並べると次のような特徴が見つかった．

- 記事の連想単語に相当する，特徴的な単語は，連続した日に現れ，さらに期間全体に渡っても現れる傾向がある．
- 「周辺」,「同日」などの時間，場所に関係する一部の名詞も同様の傾向があるが，このような名詞は特に事件に関連した記事の場合，その事件に集中して現れる傾向が大きく，idf 法などでは排除が困難である．ここでは，基準単語の前後に現

れる可能性が大きいと考え，基準単語の連接単語としてフィルターする処理を考えた．

- 一方,「今後」などの単語も連続して出現するが，これらはidf法などで排除可能であると考えた．

このような傾向と対策から基準単語と連想単語の抽出が可能であると考える．

6 おわりに

本研究では，新聞記事に対してユーザが自然発生的に興味を起こし，それを契機に記事を追う興味の動きを考慮にいれ，情報トラッキングのアルゴリズムを考えた．さらに一部単語の時系列的な抽出と頻度，重み情報の解析により，クラスタを特徴を表し，トラッキングの対象となる，基準単語，連想単語と抽出のためのアルゴリズムの詳細化を行った．

今回は特定のテーマの記事にしぼった実験だったので，一般的なテーマとなるイベント，事件に関係したトラッキング手法を実験的に評価していき，提案したアルゴリズムの適性を評価したい．

参考文献

[1] 新谷研，角田達彦，大石巧，長尾真：単語の共起頻度と出現位置による新聞の関連記事の検索手法，情報処理学会論文誌,Vol.38,No.4,pp.855-862(1997).

[2] 巖寺俊哲，菊井玄一朗：トレンド・トラッキング型テキスト自動分類の試み，情処研報，NL119-4，pp.19-24(1997).

[3] 杉崎正之，井上孝史，大久保雅且，田中一男：情報潮流抽出のための分類精度の改善手法について，第56回情処全大会，分冊3，pp.98-99(1998).

[4] 湯浅夏樹，上田徹，外川文雄：大量文章データ中の単語間共起を利用した文書分類，情報処理学会論文誌,Vol.36,No.8,pp.1819-1827(1995).