電子新聞に対する動的クラスタリング手法の提案

2V－1

北川 結香子，中嶋 卓雄，中村 良三

熊本大学 工学部

1　はじめに

ネットワークの発展によりWebなどにおいても，日々内容が変化する情報の量が増加している．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．

情報検索の分野で広く使われてきた分類では，まず，単語の重みをtf*idf手法によって計算し，その値に基づき文書間の関連度を計算しクラスタリングされている．

しかし，tf*idf手法は，膨大な文書集合に基づく検索には適していると考えられるが、1日の新聞の記事集合のように話題が変化し，中心となる話題が新聞全体に反映するような文書集合には適するとは限らない．

本稿では，単語の重みを評価する関数を提案するとともに，記事の関連度を求め，電子新聞に対する動的なクラスタリング手法を提案する．

2　記事中の単語の重み

2.1　tf*idf手法

tf*idf手法では，記事中の単語は多く出現するほど記事の内容を表わすと考えて大きな重みを与え，さらに記事全般に頻繁に出現する単語は一般的な単語と見なして小さな重み与える．

まず，TF法では，記事中に多く出現する単語ほどその文書の内容をよく表わすと考えて、そのような単語に大きな重みを与える．記事d中の単語t_0の出現回数$F(t_0)$による重み付け関数$TF(d,t_0)$は次のように定義できる．

$$TF(d,t_0) = F(d,t_0) \tag{1}$$

また，単語の1日の記事全体にわたる出現頻度に対応して，単語t_0の重み付け関数$IDF(t_0)$を次のように定義する．

$$IDF(t_0) = log(M/df(t_0)) \tag{2}$$

M：　1日の記事中に含まれる記事の総数

$df(t_0)$：　1日の記事中で単語t_0が出現する記事総数

tf*idf法において，単語の重み付け関数$V(d,t_0)$を2つの関数の積として次のように定義する．

$$V(d,t_0) = TF(d,t_0) \times IDF(t_0) \tag{3}$$

2.2　提案する重み付け関数

2.2.1　1つの記事中の単語の重み

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．

Proposal of Dynamic Clustering Method for Electronic Newspaper

Yukako Kitagawa,Takuo Nakashima,Ryozo Nakamura

Faculty of Engineering, Kumamoto University,

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 $L(d,t_0)$ を次のように定義する．

$$L(d,t_0)=\sum_{x=1}^{X} f(x)*TFL(d,x,t_0) \qquad (4)$$

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

$f(l)$: 位置 l における重み関数（減少関数）

2.2.2 重み付け関数

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かと思われる．

そこで，日付 $a-\alpha$ から a において，次の単語集合 $D(a,\alpha)$ に単語が含まれる場合には，一般的な単語を表わしていると考える．

$D(a,\alpha)=$ $\{t_0|$ 日付 $a-\alpha$ から a までの期間において $\frac{TF_{all}(t_0)}{TFA} \geq threshold\}$

$TF_{all}(t)$: 1日のすべての記事における単語 t の出現頻度

TFA : 1日のすべての記事のすべての単語の出現頻度の総和

本研究では，位置情報による重みと混合させ，記事 d 中の単語 t_0 の重み付け関数 $V(d,t_0)$ を次のように定義する．

1. $t_0 \in D(a,\alpha)$ の時，

$$V(d,t_0)=L(d,t_0)*IDF(t_0) \qquad (5)$$

2. $t_0 \notin D(a,\alpha)$ の時

$$V(d,t_0)=L(d,t_0)*W(t_0) \qquad (6)$$

ここで，$W(t)$ は単語 t の出現頻度の増加によって増加する関数とする．

3 記事の関連度

2つの記事の関連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 d_x，d_y 間の関連度 R を次のように定義する．

$$\mathrm{R}(\mathrm{d_x},\mathrm{d_y})=\frac{\sum_{t_{x\cap y}} \mathrm{V}(\mathrm{d_x},\mathrm{t_{x\cap y}})}{\sum_{t_x} \mathrm{V}(\mathrm{d_x},\mathrm{t_x})} * \frac{\sum_{t_{x\cap y}} \mathrm{V}(\mathrm{d_y},\mathrm{t_{x\cap y}})}{\sum_{t_x} \mathrm{V}(\mathrm{d_y},\mathrm{t_x})} \qquad (7)$$

4 クラスタリング

記事間の関連度を計算し，クラスタ間の距離（類似度）を記事の関連度とし，クラスター分析によって記事をクラスター化する．クラスター分析する手法としては，最長距離法および群平均法を考えている．

最長距離法を利用する場合，p クラスタと q クラスタを統合して新しく r クラスタを作るとき，それと別の s クラスタとの関連度 R_{rs} を，

$$R_{rs}=max(R_{ps},R_{qs}) \qquad (8)$$

により定義する．

5 おわりに

本稿では，日々変化する新聞記事を対象とした動的クラスタリング手法を提案した．部分的な評価実験により，単語集合 $D(a,\alpha)$ は，$\alpha>2$ の場合において，要素の総数がゆるやかに減少することが判明しているので，$\alpha=2$ または 3 における，評価により有効性が求まると考えている．

今後は，評価実験により，妥当な関数の構成を求めていきたい．

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).