E-028

テーマ指向性単語重み付け方式の提案と単文書要約への適用

Theme Characterized Term Weighting Method and Its Application to Single Document Summarization

渡辺修司† 木村 誠‡ 久光 徹‡ 絹川博之†
Watanabe Shuuji Kimura Makoto Hisamitu Toru Kinukawa Hiroshi

1. はじめに

近年，電子書籍，Web，ネットニュース，電子メールなど電子化されたテキスト情報が氾濫する状況となっている．また同時に，情報検索システムもその適用分野やデータベースサイズを急速に拡大しつつある．これに伴い，膨大な情報の概要や内容を短時間で把握することが必要となっており，これを補助する技術として，テキスト自動要約技術が注目されている[1].

本研究では，従来からの単語重み付け法の適用法を拡張し，要約対象文書中において，(1) 類似文書群を特徴付ける単語と，(2) 類似文書群との違いを特徴付ける単語の重みを，それ以外の単語より大きくする，テーマ指向性単語重み付け方式を提案し，単文書要約に適用する．提案方式は，要約対象文書には一般的にそれに類似した文書が複数存在することに着目したものであり，単語重み付けに複数の類似文書群の情報を用いることを特徴としている．

2. テーマ指向性単語重み付け方式

2.1 類似文書群と特徴単語の重み付け

語の出現頻度に基づいて，ある文書 d に含まれる単語 $\{w_1,w_2,...,w_n\}$ の重みを算出する場合，従来からの方法では，ある文書集合全体における w_i ($1 \leqq i \leqq n$) の出現頻度と，文書 d における w_i の出現頻度を用いた計算が行われる．この場合，算出された重みは，文書 d が扱うテーマや話題と全く関係のないものになる．

これに対して我々は，要約対象文書 d と類似した文書集合 D_S が存在することに着目し，d と D_S との共通点と相違点をそれぞれ特徴付ける単語に大きな重みを与えるテーマ指向性単語重み付け方式を新たに提案する．この類似文書集合 D_S を導入することにより，

D: 全文書集合
d: 要約対象文書
D_S: d の類似文書群

の3つの文書集合が定義され，$d \subset D_S \subset D$ となる．ここで，P をある集合，S を P の部分集合とすると，P と S の組み合わせとして以下の3種類が考えられる．

(A) P=D, S=d (B) P=D, S=Ds (C) P= Ds, S=d (図1参照)

(A) は D に対する d の特徴単語の重みを大きくする重み付け法となり，従来から用いられているものである．

(B) は D に対する D_S の特徴単語の重みを大きくする重み付け法．

(C) は D_S に対する d の特徴単語の重みを大きくする重み付け法．

(B) 及び (C) の方法で得られる単語重みを基に抽出した単語は，それぞれ，要約対象文書 d において，テーマ全体を特徴付ける単語と，要約対象文書 d に独自の話題や情報を特徴付ける単語であると考えることができる．

†東京電機大学大学院
‡日立製作所

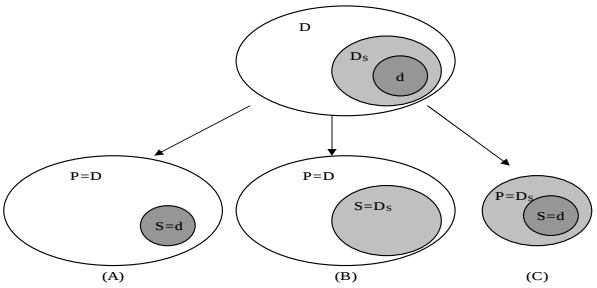

図 1 要約対象文書と類似文書群と全文書集合の関係

2.2 類似文書特徴に基づく単語重み付け方式

d の類似文書群 D_S の特徴を利用した単語重み付け方式を提案する．これらの重み付け方式は，テーマ全体に共通する話題，ある文書独自の話題や情報，を反映したものであると考えられる．また，類似文書群を用いない従来の重みを (4) に示す．

ある単語重み付け法 Method による，S の P における特徴を表す単語 w の重みを，$W_{Method}(w|S|P)$ と表すものとする．

(1) テーマ共通特徴重み

要約対象文書 d において，類似文書群 D_S を共通的に特徴付ける単語 w の重み $W^C{}_{Method}(w)$ を(式 1)で定義し，以降テーマ共通特徴重みと呼ぶ．

$$W^{C}{}_{Method}(w) = W_{Method}(w \mid D_S \mid D) \qquad \text{(式 1)}$$

(2) テーマ個別特徴重み

要約対象文書 d と類似文書群 D_S との違い，すなわち要約対象文書 d 独自の話題や情報を特徴付ける単語 w の重み $W^D{}_{Method}(w)$ を (式 2) で定義し，以降テーマ個別特徴重みと呼ぶ．

$$W^{D}{}_{Method}(w) = W_{Method}(w \mid d \mid D_S) \qquad \text{(式 2)}$$

(3) テーマ混成特徴重み

テーマ共通特徴重みとテーマ個別特徴重みとの双方を併せ持つ重み $W^{C+D}{}_{Method}(w)$ を (式 3) で定義し，以降テーマ混成特徴重みと呼ぶ．

$$W^{C+D}{}_{Method}(w) = W^{C}{}_{Method}(w) + W^{D}{}_{Method}(w) \qquad \text{(式 3)}$$

テーマ混成特徴重みは，d の類似文書群 D_S 全体を特徴付ける度合いと，d と D_S との違いを特徴付ける度合いをともに含んでいると考えられる．

(4) 類似文書集合を用いない重み

類似文書群を用いずに，要約対象文書 d において，全文書集合 D との違いを特徴付ける単語 w の重み $W^U{}_{Method}(w)$ を (式 4) で定義する．

$$W^{U}{}_{Method}(w) = W_{Method}(w \mid d \mid D) \qquad \text{(式 4)}$$

3. 単語重み付け法のテーマ指向性への拡張

2 章で提案したテーマ指向性重み付け基本方式に基づき，4 種類の単語重み付け法を拡張する．

3.1 tf-idf 法

Salton らにより提案された手法で，より少ない文書に偏って出現する単語が多く出現するときに大きな重みを与える方法である．(式 5) のように拡張する．

$$W_{tf-idf}(w|S|P) = tf(w|S) \times \log\left(\frac{N(P)}{N(P|w)}\right) \quad \text{(式 5)}$$

ここで，tf(w|S) は S に含まれる w の数，N(P) は P に含まれる全文書数，N(P|w) は P のうち w を含む文書数である．

3.2 tf/TF 法

w の，S における出現確率と，P における出現確率の比である．(式 6)のように拡張する．

$$W_{tf/TF}(w|S|P) = \frac{tf(w|S)}{tf(w|P)} \quad \text{(式 6)}$$

3.3 SMART 法

Singhal らにより提案された手法．tf-idf 法に文書長による正規化を施し，精緻化した方式である．(式 7) のように拡張する．

$$W_{SMART}(w|S|P) = \left\{\sum_{t\in S} \frac{\log(tf(w|t))+1}{Ave_{u\in t}\{\log(tf(u|t))+1\}}\right\} \times \log\left(\frac{N(P)}{N(w|P)}\right) \quad \text{(式 7)}$$

※Ave{・}：{ }内の要素の平均をとるオペレータ

3.4 HGS 法

久光らが提案した手法．超幾何分布を応用した確率計算に基づく方式であり，高頻度語や低頻度語に偏らない公正な重み付けが高速に行える．(式 8)で定義し，パラメータ定義を拡張する．

$$W_{HGS}(w,S,P) = -\log\left(\sum_{l\geq k} hg(N,K,n,l)\right)$$

$$hg(N,K,n,l) = \frac{C(k,l)C(N-K,n-l)}{C(N,n)} = \frac{n!K!(N-K)!(N-n)!}{N!l!(n-l)!(K-l)!(N-K-n+l)!} \quad \text{(式 8)}$$

$$(\min\{0, N+K-n\} \leq l \leq \max\{n, K\})$$

N = P の単語数　　K = w の P 中での頻度

n = S の単語数　　k = w の S 中での頻度

4. 重要文抽出型単文書要約への適用

提案方式を用いて，新聞記事の重要文抽出型単文書要約システムを試作した．以下に，処理手順を示す．

(1) 重み付け対象単語の選出

d の見出しと本文を Juman Version３．６１で形態素解析し，品詞の選定により重み付けの対象として，名詞，動詞，形容詞，未定義語を選出した．

(2) Ds 作成

Ds は，初めに d と D の見出しの重み付け対象単語により類似度を求め上位１０００件に絞る．次に d と絞られた１０００件の本文の重み付け対象単語により類似度を求め上位 300 件を Ds とした．

(3) 単語重み付け

重み付け対象単語 w に対して $W^{U}_{Method}(w)$，$W^{C}_{Method}(w)$，$W^{D}_{Method}(w)$，$W^{C+D}_{Method}(w)$ を算出する．Method として 3 章で述べた 4 方式の各々を適用し，4 種の実験を可能とした．

(4) 文重みの計算

ある文 s に含まれる重み付け対象単語 $\{w_0, w_1, \ldots, w_n\}$ の重みの合計を文重みとした．

(5) 重要文抽出

文重みの上位から，指定された要約率を超えるまで抽出を行い，出現順に接続して結果として出力する．

5. 評価結果と考察

NTCIR Workshop2 TSC1 の重要文抽出型要約タスク (Task A1) の Formal run データにより評価した(表１)．HGS 法の w^{C} が全ての要約率において最高値をしめした．要約率 10%，50%では NTCIR の最高値を上回り，要約率 30%では NTCIR の２番目に良い結果となった．

6. おわりに

本研究では類似文書群の情報を用いる単語重み付け方式を提案し，その効果を確認した．今後は最も有効な Ds の文書数を実験により求めていく．

参考文献

[1] 奥村 学, 難波 英嗣, "テキスト自動要約に関する研究動向(巻頭言に代えて)," 自然言語処理, Vol.6, No.6, pp.1-26, 1999.

[2] 久光 徹, 丹羽 芳樹, "組み合わせ的確率モデルに基づく特徴単語選択方法 -超幾何分布の応用-," 自然言語処理, 140-12, pp.85-90, 2000.

[3] 木村 誠, 絹川 博之, "新聞記事からの特徴単語抽出方式とその評価," 情報処理学会第63回全国大会講演論文集, Vol.2, No.1Q-2, pp.149-150, 2001.

[4] 木村 誠, 絹川 博之, "新聞記事のテーマ指向性要約における各種単語重み付け方式の定量的評価,"情報科学技術フォーラム一般講演論文集, Vol.2, No.E-4, pp.89-90, 2002.

表 1　NTCIR-2 TSC1 TaskA1 Formal run データによる単語重み付け方式の評価

要約率	tf-idf				tf/TF				SMART				HGS				NTCIR
	W^{C}	W^{D}	W^{C+D}	W^{U}	W^{C}	W^{D}	W^{C+D}	W^{U}	W^{C}	W^{D}	W^{C+D}	W^{U}	W^{C}	W^{D}	W^{C+D}	W^{U}	
10%	0.308	0.346	0.323	0.279	0.284	0.257	0.264	0.239	0.350	0.237	0.350	0.282	0.368	0.272	0.306	0.267	0.363
30%	0.444	0.449	0.447	0.450	0.425	0.390	0.404	0.348	0.456	0.395	0.456	0.439	0.469	0.426	0.454	0.447	0.483
50%	0.598	0.607	0.607	0.598	0.581	0.564	0.558	0.565	0.617	0.554	0.617	0.582	0.617	0.582	0.616	0.591	0.612
平均	0.442	0.468	0.457	0.442	0.430	0.404	0.409	0.384	0.474	0.396	0.474	0.434	0.485	0.426	0.458	0.435	0.467

※網掛け：最も性能の高かった重み付け方式