情報利得比に基づく語の重要度とMMRの統合による複数文書要約

佐々木拓郎 † 野澤正憲 ‡ 森辰則 ††

† 横浜国立大学 大学院 環境情報学府 ‡ 横浜国立大学 工学部 †† 横浜国立大学 大学院 環境情報研究院

E-mail: {takuro,nozawa,mori}@forest.eis.ynu.ac.jp

1 はじめに

電子化された文書が溢れ，なおも増え続ける現在，そこから利用者が必要とする情報を効率よく入手する事は困難である．確かに，WWW 検索エンジンに代表される情報検索システムのように文書の組織化を行なうことで，検索要求に対応する文書を容易に得る事が出来るようになった．しかし，その絞り込まれた文書集合から利用者が真に必要とする情報を入手する為には，原文書を読まなければならないのが現状である．一方で，利用者の読むべき文書量を削減する手法として，自動文書要約も注目されている．組織化された文書集合から情報を獲得する事を考えると，要約対象を単一文書とするよりも複数文書とした方がより効率が上がる．

そこで，本稿では，文書分類をされた後の複数文書を対象とし，要約を提示する一手法を提案する．複数文書から内容を網羅した要約を作成する為に，情報利得比を用いた語の重要度と MMR の統合による重要文抽出を行なう．また，抽出した重要文の間の結束性を高めるために，ハニング窓関数を用いる．そして，本手法を内容の網羅性と可読性の観点から検討する．

2 情報利得比に基づく語の重要度と MMR の統合による重要文抽出

要約対象が複数文書の場合には，単一文書要約とは別に考慮すべき要素がある．本稿では，要約対象は利用者により既に選択されている検索結果としての複数文書を前提とし，それらは与えられるものと考える．この状況下では，複数文書要約を行なうために，以下の事柄が必要と考える．

1. 複数文書からの検索要求を考慮した重要箇所抽出
2. 文書間の冗長な箇所を削除する事
3. 文書間の相違点をまとめる事

本節では生成される要約文書における内容の網羅性を高める為に，以下の 2 つを統合した重要文抽出手法を提案する．

- 情報利得比に基づく文の重要度の導出 (1，3)
- MMR に基づく要約文中の冗長性の制御 (2)

そして，抽出した重要文を話題ごとに分類し，話題ごとに説明のためのキーワードを付与して出力する．

2.1 情報利得比に基づく語の重要度

我々は検索結果文書の各々を要約する手法として，情報利得比に基づく語の重み付けを用いた重要文抽出手法を提案している [森 02]．この手法では，複数の検索文書の間に存在する類似性構造を階層的クラスタリングにより抽出し，その構造に則した語に高い重みをつける．文書間の類似性構造を語の重みに写像する方法として，我々は，各クラスタ内での語の確率分布に注目し，情報利得比 (Information Gain Ratio,IGR)[Qui93] と呼ばれる尺度を用いる．さて，C_i を C の部分クラスタとするとクラスタ C における単語 w の情報利得比 $gain\_r(w, C)$ は次のように求められる．

$$
\begin{aligned}
gain\_r(w,C) &= \frac{gain(w,C)}{split\_info(C)} \\
gain(w,C) &= info(w,C) - info_{div}(w,C) \\
info(w,C) &= -p(w|C)\log_2 p(w|C) \\
&\quad - (1-p(w|C))\log_2(1-p(w|C)) \\
p(w|C) &= \frac{freq(C,w)}{morph(C)} \\
info_{div}(w,C) &= \sum_i \frac{morph(C_i)}{morph(C)} info(w,C_i) \\
split\_info(C) &= -\sum_i \frac{morph(C_i)}{morph(C)} \log \frac{morph(C_i)}{morph(C)} \\
freq(w,C) &= \text{クラスタ内の語}w\text{の頻度} \\
morph(C) &= \text{クラスタ内の形態素数}
\end{aligned}
$$

ここで次の二点に注意しなければならない．

1. 対象文書群が情報検索結果であれば，それらと検索されなかった残りの文書群との対比により得られる情報が重要である．このため，図 1の最上部に示す通り，検索文書集合から得られたクラスタ構造の根の上にもう一つ仮想的なクラスタを設ける．このクラスタには検索文書の属する部分クラスタとそれ以外の文書が属する部分クラスタが存在する．このクラスタにより，対象文書群全体に関連する語に高い重みが与えられるので，検索のトピックに関する語が暗に重みづけられる．

2. 階層的なクラスタリングを考える場合，図 1に示すとおり，各クラスタにおいて情報利得比による語の重みが得られる．各階層でのクラスタ分割に関しての語の重要度を考慮するためには，これらを統合する必要がある．本稿では，各文書の所属するすべてのクラスタにおける語の重みの和を採用する．これを IGR_sum と呼ぶ．

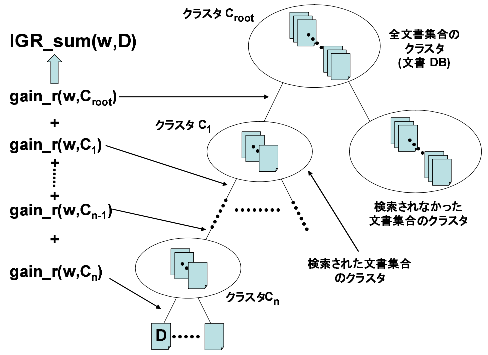

図 1: 情報利得比に基づく語の重み IGR_sum

そして，この重みと文書内単語頻度 (TF 値) や文書頻度の逆数 (IDF 値) など従来より提案されている他の重みづけを組み合わせることにより，最終的な語の重みとする．TF，IDF，IGR_sum はそれぞれ各文書，全文書集合，文書間の類似性構造に基づき決まる語の重みであり独立な要素であるので，全ての要素が独立に寄与する積 ($TF{\cdot}IDF{\cdot}IGR\_sum$) を用いる．各文の重要度はそこに含まれている名詞の重要度の総和を文の長さ (単語単位) により正規化したものである．

2.2 MMR

Carbonell ら [CG98] は，MMR(Maximal Marginal Relevance) と呼ばれる手法でパッセージ間の冗長性に対処するための再順位付けを行なっている．MMR は式 (1) により定義され，検索質問 Q が与えられた時に，関連する文書集合 R から次に選択すべき文書を与えるものである．すなわち，検索質問に対するパッセージの関連度とパッセージ間の類似度の両方を考慮して，共通箇所の検出 (冗長性削除) と相違点の検出 (内容の網羅) を同時に行なう事が出来る．

$$MMR(R,A) \stackrel{\mathrm{def}}{=} Arg \max_{D_i \in R \backslash A} [\lambda Sim_1(D_i,Q) - (1-\lambda) \max_{D_j \in A} Sim_2(D_i,D_j)] \quad (1)$$

ここで，A は既に選択された上位の文書集合，Sim_1，Sim_2 は，それぞれ，文書と検索質問の間の類似度，文書間の類似度である．A を空集合に初期設定し，λ に適切な値を設定してから式 (1) を繰返し適用すると，冗長性を考慮した文書の順位づけが行なえる．式 (1) において，右辺第一項と第二項はそれぞれ，検索質問に関連する箇所を抽出する項，冗長性削除と相違点検出を行なう項と考えることが出来る．この両項の効果を調節する働きをするのが変数 λ である．

2.3 MMI-MS

ここで情報利得比に基づく重要文抽出手法において，MMR と同種の冗長性制御機構を導入することを考える．MMR は，本来，文書もしくはパッセージを単位として，順位づけを行なうものであり，初期順位 (Sim_1) は，検索質問に対する文書の関連性 (Relevance) を用いていた．これを文を単位とし，初期順位を文の重要度により与えるように変更すれば，重要文抽出の枠組で MMR と同等のことが行なえる．本稿ではこれを MMI-MS(Maximal Marginal Importance – Multi-Sentence) と呼び，式 (2) のように定義する．

$$MMI\text{-}MS(SS,A) \stackrel{\mathrm{def}}{=} Arg \max_{S_i \in SS \backslash A} [\lambda Imp(S_i) - (1-\lambda) \max_{S_j \in A} Sim_s(S_i,S_j)] \quad (2)$$

ここで，SS は要約対象となるすべての文の集合，$Imp(S_i)$ は文 S_i の重要度，Sim_s は文間の類似度を表す尺度である．また，複数文書における全ての文を等しく扱っていることに注意されたい．

本稿では，$Imp(S_i)$ と Sim_s について，それぞれ，情報利得比に基づく文の重要度と文ベクトルの cosine 類似度を採用する．

2.4 ハニング窓関数による文重要度平滑化

これまでに述べた手法でシステムを作成し，NTCIR3 TSC2 課題 B による評価実験を行なった結果，対象文書数が多い時には，多くの文書から少しずつ重要文を抽出し，文間の結束性が低下する傾向が見られた．対象文書数が多い場合にも対応可能なシステムを構築する為には，文間の結束性を高める事が必要である．そこで，重要な文がほぼ収まるような文の数を設定し，その範囲内で重要度が滑らかに変化するように重要度の平滑化を行なうことを提案する．具体的には，各文の重要度を，ハニング窓関数を用いる事で，周囲のある範囲の文の重要度も考慮しつつ再計算する．ハニング窓関数は，範囲の中心付近の重要度を重視し，中心から離れるにしたがって重みを軽くするという特徴を持つ．

窓の幅 (重みを与える範囲) を W，窓の中心位置を l とすると，ハニング窓関数 $h_l(i)$ は式 (3) により与えられる．

$$h_l(i) = \frac{1}{2}(1 + cos 2\pi \frac{i-l}{W}) \qquad (|i-l| \leq \frac{W}{2}) \quad (3)$$

黒橋ら [黒橋 96] はハニング窓関数を用いて，語のテキスト中での出現密度分布を調べ，その高密度な出現位置を取り出す事によって，その語の重要説明箇所を特定する手法を提案している．栗山ら [栗山 02] は，文書中から複数の語を選定し，ハニング窓関数を用いて複数の語の出現密度分布を調べ，その高密度な出現位置を重要文として抽出し，抄録を作成する手法を提案している．これらの手法では，ハニング窓関数における横軸を一語としているが，我々の手法では，横軸を一文として考える．文の重要度の再計算は以下のアルゴリズムで行なう．

1. ハニング窓関数の横軸は一文を単位とし，再計算の対象となる文を窓の中心位置 l とする．

2. 指定した窓幅 W に含まれる全ての文 i (中心位置の前後それぞれ $\frac{W}{2}$ の範囲の文) について，文 i の重要度 $S\_weight(i)$ とハニング窓関数の値 $h_l(i)$ を次式によって足し込む事で重要度の再計算を行ない，文間の結束性を考慮した重要度 $C\_weight(l)$ を算出する．

$$C\_weight(l) = \sum_{i=l-\frac{W}{2}}^{l+\frac{W}{2}} h_l(i) \cdot S\_weight(i) \tag{4}$$

3. 文書の先頭から順に各位置の重要度の再計算を行なう．文書の先頭と末尾で，重みを足し込む文がない場合には，疑似的に先頭と末尾の文が連続的に存在するものものとして，その重みを用いる事にする．

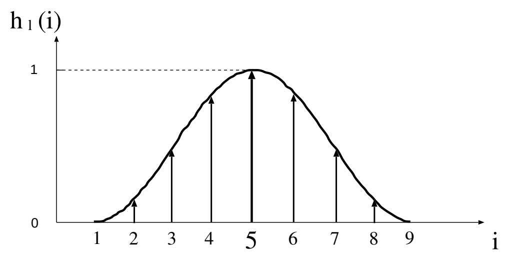

図 2: 窓幅を 8 とした時のハニング窓関数

ここで，複数の要約を対象に窓幅を変化させて予備実験を行なった結果，文間の結束性が感じられ，可読性 (要約の内容の理解のしやすさ) が最も高かった事から，窓幅を 8 文に決定した．

この時，式 (4) に $W = 8$ (窓幅 8) を代入し，一例として 5 番目の文について考えると，ハニング窓関数は図2のように表される．図 2において，$h_l(1) = h_l(9) = 0$ である事から，窓幅を 8 にすると，前後 3 文ずつ (2-8 番めの文) を考慮しながら各文の重要度の再重み付けを行なう事になる．

3 実験と評価

3.1 NTCIR TSC2 課題 B による順位付け評価

ハニング窓関数を用いないシステムにより NTCIR3 TSC2 課題 B[TSC 01] に参加し，評価を行なった． TSC2 ではタスクオーガナイザから与えられるトピック情報に従い，要約を生成し，その生成結果を提出した．

同 Formal Run は， 30 のトピックから構成される．各トピックは一つの情報検索結果に相当し以下の情報などから構成される．なお，対象となる文書は毎日新聞 98 年， 99 年の記事である．トピックの ID，要約の対象となる文書群，生成すべき要約文書の長さなどからなる．要約の長さには，長い要約 (Long) と短い要約 (Short) の 2 種類がある． Long は Short の倍の長さであるが，その長さは対象文書数などによって異なる．

順位づけ評価では，各トピック，各要約文書長に対して，評価対象システムの要約以外に， baseline として以下の 3 種類の要約を用意する．

1. 人間による自由作成要約 (表中では ‘人手’)
2. lead 法による要約
3. stein 法による要約 [SSW99]

次に，要約評価者に，トピック毎に原文書集合と各要約結果を読んでもらい，以下の内容の網羅性 (表中では ‘C’) と可読性 (表中では ‘R’) の観点から要約文書の順位づけを行なってもらう．評価対象のシステムの順位が小さい値ほどよいということになる．その内，対象文書数の少ない 15 トピック (7 文書以下) に関する結果を表 1に示す．文書数が少ない時には同表の示すように baseline と比較して評価が高かった．一方で文書数が多い残りの 15 トピックについては性能が勝っているとは言えなかった．

表 1: baseline との優劣 (7 文書以下の 15 トピック)

	C Short			R Short		
	W	L	T	W	L	T
v.s. 人手	6	9	0	6	9	0
v.s. Lead 法	**11**	**4**	**0**	6	9	0
v.s. Stein 法	**13**	**1**	**1**	7	8	0
	C Long			R Long		
	W	L	T	W	L	T
v.s. 人手	7	8	0	5	10	0
v.s. Lead 法	**10**	**5**	**0**	7	8	0
v.s. Stein 法	**10**	**4**	**1**	3	12	0

W:win L:lose T:tie

3.2 ハニング窓関数を組み込んだシステムに対する順位づけ評価

ハニング窓関数を導入した手法に対する評価を行なうにあたり，極力 NTCIR3 TSC2 順位づけ評価に環境を近付ける事を目指した． NTCIR3 TSC2 において baseline として用いた Lead 法と Stein 法の出力結果とソースコードは現在のところ公開準備中であり入手できなかったが，人手による自由作成要約は公開されている為，以下の 3 種類の要約についてトピック毎に順位づけ評価を行なった．要約評価は大学院生 4 人と大学生 2 人で行なった．なお，可読性に関する尺度として，「読みものとして，要約が述べる内容の理解のしやすさ」と定義し，被験者に周知した．

1. 人間による自由作成要約
2. ハニング窓関数を導入した手法による要約
3. NTCIR3 TSC2 課題 B 時の手法による要約 (表中では ‘Formal Run’)

その内，Formal Run の時に評価の悪かった，対象文書数の多い 15 トピック (8 文書以上) に対する結果を表 2に示す．

表 2: baseline との優劣 (8 文書以上の 15 トピック)

	C Short			R Short		
	W	L	T	W	L	T
v.s. 人手	0	15	0	0	15	0
v.s. Formal Run	3	6	6	**5**	**1**	**9**
	C Long			R Long		
	W	L	T	W	L	T
v.s. 人手	0	15	0	1	14	0
v.s. Formal Run	**10**	**3**	**2**	**9**	**3**	**3**

W:win L:lose T:tie

4 考察

4.1 情報利得比と MMI の統合による手法

表 1に示した NTCIR3 TSC2 課題 B の実験による評価結果より，我々のシステムは baseline である lead 法，stein 法と比較して，内容の網羅性においては優れていると考えられる．特に，対象文書数が少なく，要約率が小さい時においてその傾向が顕著に表れている．この結果は，IGR による語の重みづけと MMR の統合に基づく重要文抽出が効果的である事を示す．

これに対して，上に示したものと反対の状況では評価が格段に下がる．特に対象文書数が多い場合には，文間の結束性が低下する事が一因であると考えられる．

4.2 ハニング窓関数による文重要度平滑化

表 2より，ハニング窓関数を組み込み，文間の結束性を向上させる事で，要約率が大きい時の内容の網羅性と，全トピックに亙っての可読性に対して，要約の性質が改善されたと言える．また，可読性 (R Short，R Long) に関しては対象文書数によらず，改善されるが，内容の網羅性に関しては，対象文書数が多く，要約率が大きい場合 (C Long) に大きく改善される傾向が見られる．また，対象文書数が少なく，要約率が小さい場合 (C Short) に関しては，多少評価が下がった．

表 1，2の結果を比較すると，対象文書数が少なく，要約率が小さい時にはハニング窓関数を組み込まない時に評価が良く，逆に対象文書数が多く，要約率が大きい時にはハニング窓関数を組み込んだシステムが良好な評価結果であった．

本稿ではハニング窓関数の窓幅を 8 に固定したが，対象文書数と要約率に応じて表 3に示すように窓幅を調節する事で，より精度の良いシステムが実現できると考えられる．

表 3: ハニング窓関数の窓幅の調節

	対象文書数		要約率	
	多い	少ない	大きい	小さい
窓幅	広く	狭く	広く	狭く

5 まとめと今後の課題

本稿では，文書分類をされた後の複数文書を対象とし，情報利得比を用いた語の重要度と MMR の統合による重要文抽出を行なう事で内容の網羅性を考慮した要約を提示するシステムを提案した．NTCIR3 TSC2 による評価においては，我々のシステムは，内容の網羅性を考慮した複数文書要約を作成するにあたり，特に要約率が小さい時と，対象文書数が少ない時 (7 文書以下の時) に効果的である事が示された．また，文間の結束性を高めるためにハニング窓関数による文重要度平滑化手法を導入し，人手による評価を行なったところ，組み込む前のシステムと比較して，内容の網羅性に関して，対象文書数が多い時 (8 文書以上の時) に効果的である事が示された．さらに，可読性に関しては，対象文書数によらず，改善される事が示された．

今後の課題として，ハニング窓関数の窓幅を対象文書数と要約率に対応するように自動的に調節し，より精度の良い重要文書部分の抽出を行なう事が考えられる．

参考文献

[CG98] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

[Qui93] J. Ross Quinlan. *C4.5: progarms for machine learning*. Morgan Kaufmann Publishers, May 1993.

[SSW99] Gees C. Stein, Tomek Strazalkowski, and G. Bowden Wise. Summarizing Multiple Documents using Text Extraction and Interactive Clustering. In *Proceedings of the sixth Pacific Association for Computational Linguistics (PACLING 99)*, pp. 200–208, 1999.

[TSC 01] TSC 実行委員会. NTCIR 3 テキスト自動要約タスク TSC-2. http://lr-www.pi.titech.ac.jp/tsc/tsc2.html, 2001.

[黒橋 96] 黒橋禎夫, 白木伸征, 長尾真. 出現密度分布を用いた語の重要説明箇所の特定. 自然言語処理研究会報告 96-NL-115, 情報処理学会, 1996.

[森 02] 森辰則. 検索結果表示向け文書要約における情報利得比に基づく語の重要度計算. 自然言語処理, Vol. 9, No. 4, pp. 3–32, 7 月 2002.

[栗山 02] 栗山義明, 絹川博之. ターム群の出現密度分布を用いた重要文抽出方式 – ターム種別重みの評価実験 –. 第一回情報科学技術フォーラム講演論文集 (FIT2002), 2002.