自 然 言 語 処 理 152－10
（2002．11．12）

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　　森辰則
横浜国立大学大学院　環境情報学府 / 環境情報研究院
E-mail: {takuro,mori}@forest.eis.ynu.ac.jp

概要

本稿では，文書分類をされた後の複数文書を対象とし，内容の網羅性と可読性を併せ持つ原文の代わりとなる要約を提示するシステムを提案する．内容の網羅性の向上に対しては，情報利得比を用いた語の重要度とMMRの統合を，可読性の向上に対しては重要文の話題ごとの分類と，話題ごとのクラスタ説明の為のキーワード付与，を提案する．
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，対象とする文書集合の数が小さい時 (7文書以下の時) に効果的である事が示された．

Information Gain Ratio meets Maximal Marginal Relevance — A method of Summarization for Multiple Documents —

Takuro SASAKI and Tatsunori MORI
Graduate School of Environment and Information Sciences, Yokohama National University
E-mail: {takuro,mori}@forest.eis.ynu.ac.jp

Abstract

In this paper, we propose a method to make a summary from multiple documents with taking account of comprehensibility and readability. As for comprehensibility, we propose a method based on IGR. As for readability, we propose a method to generate a summary based on clustering important sentences according to subtopics and making a keyword list as a very brief summary for each cluster.
By the evaluation in NTCIR3 TSC2, we show that the proposed method works well to generate comprehensive summaries when the length of summary is short and the target is a small (7 or less) number of documents.

1　はじめに

デジタル文書が氾濫する現在，ユーザが必要とする情報を効率良く入手する事は困難な問題の一つである．その一助とするために，情報検索や文書分類など文書の組織化に関する研究が行なわれている．また，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．特に，近年，より効率の良い要約を目指して，複数文書を対象とした自動要約技術に注目が集まっている．

複数文書を要約するにあたり，大規模な文書集合を直接要約するのではなく，ある程度内容のまとまった文書集合を扱う事が現実的であると言えよう．例えば，情報検索システムのバックエンドに配置するナビゲーションシステムの一部として要約システムを用いることを考えよう．検索の初期の段階では，ナビゲーションシステムは得られた検索結果文書を分類し，利用者との対話により，順次，利用者が必要とする文書集合に絞り込んでいく．そして，利用者が必要とする複数文書が得られたとすると，そこで複数文書の概要を示す要約を提示する．この時にその要約が，原文の代わりとなる精度の良い要約となっていれば，それは個々の文書またはそれぞれの要約を読む事に比べて，利用者にとって遥かに効率の良いものとなる．

本稿では，利用者が持つある特定のトピックにより検索が行なわれ，ある程度取捨選択が行なわれた後の複数文書を対象とし，原文の代わりとなる要約を提示することを目的とする．原文の代わりとなる要約とするためには，次の二点が重要であると考える．

要約文書の要件 1 (内容の網羅性の高さ) 決められた要約率の中で，原文の持つ主要な内容を網羅した出力であること

要約文書の要件 2 (可読性の高さ) 生成された要約が一つの読み物として読みやすいこと

先行研究としては，要約文書の要件2を考慮しつつ，要約文書の要件1を中心に据えて議論を行なっているものが多い．例えば，Radevら [RJB00, RF00] は，与えられた文書群をクラスタリングし，各クラスタに対して独立に要約生成を行なう手法を提案している．また，Steinら [SSW99] の手法では，まず個々の文書を要約し，これらをクラスタリングする．そして，各クラスタの中から代表となる要約を一つずつ選択することにより，全文書の要約を得ている．Goldsteinらは [GMCK00] は新聞記事を対象とした情報検索結果文書集合に対して，その中から検索質問に関連するパッセージを抽出，収集し，それらを並べて要約を作成するMMR-MD(Maximal Marginal Relevance -Multi-Document) という手法

を提案している．基本となる枠組はパッセージ集合を検索質問に対する関連性に基づき順位づけすることであるが，各パッセージに対して，より上位に位置しているパッセージとの類似性に基づきペナルティを与え，再順位づけを行ない，冗長性の制御を行なう．この手法でもクラスタリングにより，類似したパッセージをグループ化し，パッセージの順位づけにおいて利用している．例えば，大きなクラスタに所属するパッセージを優遇し，また，同じクラスタにより上位に順位づけされたパッセージが存在する時にはペナルティを与える．

ここに述べたいずれの手法も，類似した文書グループの発見において，(非階層的な) クラスタリングを用いているが，その一方で，クラスタ間の関係や，各クラスタ内部のより詳細な部分クラスタ構造は利用していない．すなわち，各々のクラスタは独立に扱われ，クラスタ内の共通事項に関心があるのみである．これらの方式では，

- クラスタ間の関係により決まる重要箇所が考慮されない．
- 設定するクラスタの大きさにより重要箇所 (共通部分) が異なってくる．

という点で問題があるので，すべての文書に跨る共通事項と個別の文書グループに固有の事項の両者を過不足なく合わせ持つ，良い要約文書を生成するためには，さらなる工夫が必要である．

我々は，対象文書集合のより詳細な類似性構造を階層的なクラスタリングにより得ることで，各クラスタでの共通事項のみならず，クラスタ間の差異も考慮できると考えた．そして，この情報を要約の手がかりとすることができれば，より肌理の細かい重要箇所の抽出ができるのではないかと考える．

本稿では，この点を主要な問題意識としつつ，内容の網羅性ならびに可読性の二点について，次のように近接する．まず，内容の網羅性については，文書間の類似性構造を語の重みに写像する手法と文の冗長性制御に MMR(Maximal Marginal Relevance) を利用する手法を統合する枠組を提案する．森 [森 02] は，検索結果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クラスタ構造として抽出し，これを語の重みとする手法を提案している．この手法ではクラスタ間に存在する構造，すなわち，差異や共通点を同時に考慮して語の重みに反映できる．この手法は検索結果文書の各々を個別に要約する際に有効に機能することが示されている．各文書を個別に要約する場合にはこの重みづけにより，重要文抽出をするだけで十分であった．それは，各々の文書がもともと冗長性の無い，完結した読み物であるからである．しかし，複数文書から単一の要約文書を生成する場合には，文書間での内容の重複があり得る．そのため，最も基本的な重要文抽出による要約においてさえも，文の重要度と冗長度の両者を同時に考慮する必要がある．そこで，我々は，元来，検索質問文とパッセージ群の関連度とパッセージ間の冗長性を同時に扱う要約手法として提案されている MMR を，単文の重要度と冗長性を同時に考慮する手法として改訂した．

二点目の可読性については，抽出された重要文を話題ごとの分類するとともに，分類された各グループに対して説明記述の為に適切なキーワードを付与することを提案する．これは，要約文を完全な単一文書として再構成するというよりも，利用者が必要な情報を的確に得る為の情報として編成することを目的とするものである．

2 複数文書要約における課題

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．この状況においては，複数文書要約を行なうために，以下の事柄が必要とされる [奥村 99, 奥村 02].

観点 1 (複数文書要約)

1. 複数文書での重要箇所の抽出
2. 文書間の共通点を検出し冗長な箇所を削除する事
3. 文書間の相違点を検出しまとめる事

さらに，要約対象となる文書群があるトピックによる情報検索結果であるとするならば，要約において以下の点も考慮しなければならない．

観点 2 (検索要求) その検索において利用者が何を欲していたかという検索要求．

また，要約対象となる複数文書の性質は，検索要求や文書データベースに応じて変化しうる．NTCIR3 TSC2 のように新聞記事を対象とした場合には，少なくとも，以下のものが考えられる．

観点 3 (対象文書群)

1. ある事件について書かれた記事とその続報記事のように単一トピックについて，時系列に従って述べているもの
2. あるトピックについて，いくつかの観点から書かれた記事のようにサブトピックから構成されているもの

最も基本的な要約手法である重要文抽出法を想定すると，観点 1 が重要文の選択に関与し，一方，観点 3 は選択された重要文をどのように並べ，文書として成すかに関与する．観点 2 は，両者に関与すると考えられる．

特に本論文では以下のように対応することを考え，次節以降で検討する．

1. 観点 1 の 1，3，ならびに，観点 2 については，文書間の類似性構造を語の重みに写像する森の技法 [森 02] で同時に実現する
2. 観点 1 の 2 については，単文に対する MMR により実現する．
3. 観点 3 については，原文書を単一リンク法によるクラスタリングにより分類し，各グループに対するキーワードを付与することにより実現する．

3 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

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，次の二点を統合した重要文抽出手法を提案する．

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

この手法は次の特徴を持つ．

- 文書間の類似性構造は語の重みに反映されているので，文書クラスタに依存せずに文抽出が行なえる．よって，各クラスタから平均的に抽出する必要はなく，重要度のみを考慮することができる．
- 文書クラスタリングの過程に自然な形で検索質問によるバイアスが統合されているので，
  - Stein らの方法と違って，検索トピックを考慮でき，
  - その一方で MMR-MD のように検索質問が必要でない．

本節では，特に，MMR を有効に機能させるために，文の重要度や文間類似度の正規化，ならびに，MMR における冗長度を制御する変数 λ の値の自動決定法を検討する．

3.1 情報利得比に基づく語の重要度

森ら [森 02, 菊池 01] 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．

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．

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

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．

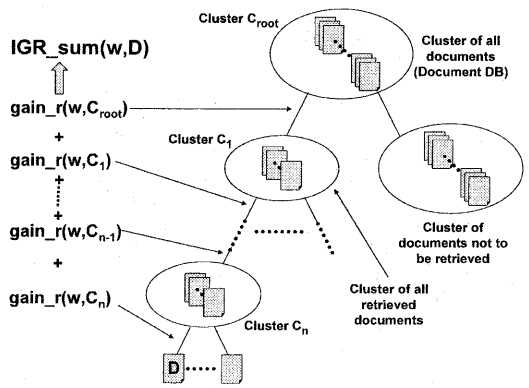

図 1: 情報利得比に基づく語の重み IGR_sum

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．

3.2 MMR

節1で述べたように，Goldstein らは MMR-MD と呼ばれる情報検索結果を対象とした複数文書要約手法を提案している [GMCK00]．特に抽出されたパッセージ間の冗長性に対処するために，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，パッセージの再順位づけが行なわれる．すなわち，MMR-MD は，複数文書要約における共通箇所の検出（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-MD における冗長性の削減は，Carbonell らが提案した MMR(Maximal Marginal Relevance) の概念に基づいている [CG98]．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．

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

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．

3.3 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，以下のように定義する．

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

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．

3.4 文の重要度と文間の類似度の正規化

MMI-MS による冗長性削除が適切に働くためには，$Imp(S_i)$ と Sim_s の値の範囲がほぼ等しくなければならない．しかし，Sim_s は 0 から 1 までの数であるのに対して，Imp はそのような制約はない．そこで，我々は以下の方法により Imp の値を正規化している．

1. 文書内での文の重要度の正規化

   各文書 D の中で文の重要度を平均値 0.5 の偏差値に変換する．

   $$Imp_s^d(S_i) = 0.5 + \frac{Imp_s(S_i) - Imp_s^{ave}(D)}{\sigma(D)} \quad (3)$$

   ただし，$Imp_s^{ave}(D)$ ならびに $\sigma(D)$ は，それぞれ，文書 D 中の文の重要度の平均値ならびに標準偏差である．この正規化は，与えられた文書群の全てが同程度に重要であると考え，文書内の重要度の平均値を揃えるためのものである．

2. 文書集合全体における，文の重要度と文間類似度の正規化

   $Imp_s(S_i)$ と $Sim_s(S_i,S_j)$ の値の範囲が同程度になるようにするために，まず，それぞれの値を最大値で除して，0 から 1 の値に正規化する．さらに，それぞれの値 $Imp_s(S_i)$ と $Sim_s(S_i,S_j)$ を平均値 0.5 の偏差値に変換する．これは，MMI-MS の効果が確実に現れるようにするための正規化である．

3.5 MMI-MS における変数 λ の自動決定

MMR-MS における変数 λ は 0 から 1 の値をとり，0 に近づく程，文の重要度が高くとも冗長であれば削除されるようになる．この変数 λ の適切な値は与えられた文書群によって異なると考えられるので，ある程度の自動推定が必要となる．例えば，内容が多岐に亙る文書集合のように重複箇所が少なければ λ を 1 に近い値にし，元の文の重要度による順位づけに近づけるべきである．逆に，ある事件に関する続報記事集合のように重複箇所が多い場合にはより小さい値を設定し，冗長部分削除が促進されるようにすると良いであろう．このように，文書集合の性質によって λ の値を適切に変化させる事が出来れば，より充実度の高い要約を作成する事ができると考えられる．

本稿では，文書集合の性質，すなわち，冗長度を判断する指標として，*cosine* 類似度の平均値に着目した．*cosine* 類似度の平均値が大きい時には，類似度の高い文が多く冗長性を削除する方向に MMI-MS を働かせるべきであるので，λ の値を小さくすればよい．このとき，次の 2 点に注意し，式 (4) を採用した．

1. λ の最大値は 1 である．
2. 文間の類似度の尺度として文ベクトルの cosine 値を利用しているので，内容としての類似性を直接計っているわけではない．よって，cosine 類似度の高さを，そのまま，冗長度の高さとして判断し，文の重要度を無視して順位づけをする状況は好ましくない．このため，文の重要度をある程度重視する設定が望ましい．

$$\lambda = 0.5 + 0.5(1 - \text{cosine 類似度の平均値}) \quad (4)$$

4 要約文書の生成

前節までに，複数文書から重要文を選択する方法について述べた．しかし，要約した文書は人間が読むものである以上，重要な内容を取りこぼさない為の内容の網羅性の高さだけではなく，読み物としての可読性の高さ，すなわち，読みやすさも考慮するべきである．

生成された要約文章の読みやすさは，文章としての自然さの他にも，各文に対して適切な文脈が与えられていて，誤った理解を与えないことも重要である．例えば，重要文を脈絡無く並べただけでは文間の結束性がなく，理解不能な要約が生成される．

抽出された重要文の集合を出発点とした場合，結束性のある文章を生成するためには，最終的には以下の一連の作業が必要となろう．

1. 重要文抽出の過程において結束性を考慮して抽出する．例えば，参照や語彙的連鎖などにより結び付いている一連の文を塊として抽出する．あるいは，照応表現の先行詞を同定し，書き換えることにより，自己完結した表現に書き換える．以下では抽出された文章小部分を単にパッセージと呼ぶ．

2. 抽出されたパッセージ群を，パッセージ間の時間順序や話題の関係を考慮しつつ並べる．

3. 必要に応じてパッセージ間の関係を明示する表現を生成し挿入する．

4. 言い換えなどにより，表現の統一や簡略化などを行ない，最終調整を行なう．

Mani ら [MGB99] は共参照情報 (coreference informaiton) 付きの構文木の並びとして表現された抜粋を書き換え，報知性 (informativeness) や可読性 (Readability) を高める手法を提案している．難波ら [難波 99] は重要文を書き換えることにより読みやすい要約を生成する手法を提案している．大塚ら [大塚 01] は，照応表現を先行詞で書き換えることにより，抽出文のつながりの悪さを改善している．

単文書要約においては，文内に元々ある結束構造を保持する努力をすれば，最低限の要約が可能であった．一方，複数文書要約においては，文書に亙る結束性を見出す必要がある．さらに，個々の文書で文体が異なることもあり得る．しかしながら，これらはいずれも難しい問題であるので，記事のもつ日付情報や元文書における文の位置などの情報を用いて読み辛くならない程度に対処することが多い．例えば，Goldstein ら [GMCK00]，Stein ら [SSW99]，Radev ら [RJB00, RF00] はいずれも，記事の記述された時間順序を保つように要約の構成要素を並べている．更に，Stein ら [SSW99] の手法では，隣接する要約の構成要素の類似度が高くなるように並べかえを行なっている．これは，疑似的にではあるが，構成要素間に結束性を持たせる試みと考えることができよう．

ここで，上述の方法はいずれも，一つの読み物としての要約を目指している点に注意されたい．事件記事のように連続した一つのトピックを扱う場合にはこれでもよいかもしれない．しかし，検索結果の文書を要約する場合には，同じトピックであっても，少しずつ異なる視点 (サブトピック) の文書が得られると考えるのが普通である．この場合，無理に一つの読み物とするよりも，サブトピックの違いをアウトラインにより明確に示し，そのアウトラインに従ってサブトピック毎の要約を個別に示すほうがよいのではないかと我々は考えた．

Barzilay ら [BEM01] は，同じ文書群から抽出した重要文を複数の人間に読み易くなるように並べ変えをしてもらう実験を行なったところ，次の知見を得たことを報告している．

- 作成された文の順序が一致することなはく，文をならべる順序において複数の可能性がある．
- しかしながら，完全に自由な順番で文が出現しているわけではなく，幾つかの文から構成されるブロックが見受けられる．
- 各ブロックを構成する文は関連するトピックのものである．

つまり，ブロックの順番とブロック内の文の順番は様々であるものの，個別トピックによりブロックが構成される点で一致している．

本稿ではこの点を考慮し，抽出した重要文をサブトピックごとに分類することにより，部分的な要約を生成するとともに，その要約に対して，どのようなサブトピックであるかを示すキーワードを付与する事で，複数文書の性質に対応した出力方式を目指した．

4.1 重要文の話題ごとの分類

あるトピックに関連する文書集合のサブトピックの構造については，少なくとも以下の 2 つが考えられる．

1. ある事件について書かれた記事とその続報記事のように単一トピックについて，時系列に従って述べているもの
2. あるトピックについて，いくつかの観点から書かれた記事のようにサブトピックから構成されているもの

出力方法としては，1 のような文書集合では，全ての文書を時系列で並べる事が適していると考えられる．それに対して，2 のような文書集合では，文書をサブトピックごとに分類し，必要に応じて時系列に並べかえる事が必要であると考えられる．いずれの場合も，話題の関連性に従って文書をクラスタに分類することができれば，対処が可能であると考えられる．

さて，クラスタリングアルゴリズムには様々なものが存在するが，どれが上記の目的に適しているであろうか?

時系列により整理すべき続報記事の集合では，扱う事件等は同じであるものの，記述される事柄は捜査の進展に従って次第に焦点が遷移していくこともある．この場合，クラスタ全体の半径が小さくなることには，さほど重きがなく，むしろ，直近の文書間距離がなるべく小さくなるようにクラスタを構成すべきである．そこで，本稿では，話題の繋がりが深い文書ごとを同一のクラスタに分類する為に，原文書間類似度を用いた単一リンク法を用いた [FBY92, BYRN99]．単一リンク法では，ある文書を分類する際に，最も近い文書の存在する既存クラスタに分類する．文書間距離としては，原文書間の距離，記事表題 (HEADLINE) 間の距離，要約文書間の距離が考えられる．しかし，HEADLINE よりも原文書の方が情報が多い点，要約文書は MMR により冗長箇所が削除されている点，の 2 点より，原文書間距離を用いた．

以上の点を考慮して，抽出された重要文を以下の手順で分類する．

1. 得られた重要文を文書単位でまとめ，文書ごとの要約を得る．以後これを単位として扱う．

2. 原文の文書間類似度 (*cosine* 類似度) を用い，クラスタリングを行う．クラスタリングアルゴリズムは単一リンク法を採用し，併合を行なう類似度の閾値を 0.5 とする．

3. 得られた各クラスタにおいて最も古い記事の日付をクラスタの日付とする．この日付情報を用いてクラスタ群を時間順に並べる．

4. 各々のクラスタの内部においては，要約文書を時間順に並べる．

5. 並べられた順序に従って要約文書を出力する．ただし，要約文書間には空行をいれ，要約文書間の区別を明示する．

4.2 要約クラスタ説明の為のキーワード付与

前節で提案した手法により分類された各クラスタに対して，単に重要文を並べるだけではなく，クラスタの説明記述を付与する事ができれば，利用者にとってクラスタの概要を把握するための手がかりとなり，可読性の向上に役立つと考えられる．

これは，複数文書の要約として単一文書を生成するという戦略ではなく，先に述べた Barzilay ら [BEM01] の実験に従い，サブトピック毎の要約を個別に示すという戦略をとるものである．

本稿では，クラスタ説明記述として，以下の手順により各クラスタから抽出したキーワードを用いる．

1. 文書ごとの要約から，各々，最重要文を抽出する．

2. 抽出した最重要文から，最重要語を抽出する．これにより，各文書毎に最重要語が一つずつ抽出される．

3. 原文書において，抽出した最重要語が複合語の一部を構成するのであれば，最長の複合語をその文書のキーワードとする．そうでなければ，最重要語をその文書のキーワードとする．

4. 各クラスタにおいて，所属する要約文書のキーワードを並べたものをクラスタの説明記述とする．

キーワードの抽出において，複合語を優先するのは，それが，より個別な情報を伝えるためである．

要約文の出力において，抽出されたキーワードの列を各クラスタの冒頭に付与した．

5 評価実験

NTCIR3 TSC2 課題 B[TSC 01] に参加し，評価をおこなった．TSC2 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，その生成結果をタスクオーガナイザに提出する．

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
- 検索要求となるキーワードのリスト
- 検索要求の簡潔な表現
- 要約の対象となる文書 (の ID) 群．3 文書から 17 文書と，トピックによって数が異なる．
- 生成すべき要約文書の長さ．長い要約 (Long) と短い要約 (Short) の 2 種類があり，(改行を含めない) 文字数で与えられている．Long は Short の倍の長さ (例えば，1000 と 500，500 と 250，250 と 125．) であるが，その長さは対象となる文書数などによって異なる．

TSC2 課題 B においてタスクオーガナイザから提供される評価結果は以下の通りである．

5.1 順位づけ評価

各トピック，各要約文書長に対して，以下の 4 種類の要約を用意する．

1. 人間による自由作成要約 (upperbound) (以下の表中では '人手')
2. 評価対象のシステムの要約
3. lead 法による要約 (baseline1)
4. stein 法による要約 (baseline2)

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 2 つの観点から要約文書の順位づけを行なってもらう．すなわち，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．

要約文書の要件 1 (内容の網羅性の高さ) 決められた要約率の中で，原文の持つ主要な内容を網羅した出力であること *(以下の表中では 'C')*

要約文書の要件 2 (可読性の高さ) 生成された要約が一つの読み物として読みやすいこと *(以下の表中では 'R')*

なお，TSC2 では，Stein 法の各ステップにおいて以下のものを用いている．

- 単一文書要約作成段階では，Lead 法を用いる．
- クラスタリングでは，クラスタ数がテキスト数の半分程度になるように設定している．

5.2 添削評価

この評価では，一つの要約について一人の被験者が添削を行うことにより，要約文書の評価を行なう．被験者には要約を読んでもらい，原文書の重要な内容を過不足なく記述しているように，また，一つの文書として読み易くなるように添削をしてもらう．添削は，挿入，削除，置換の 3 つの操作のみで行なわれる．なお，要約の半分以上を書き換えないと，条件を満足できないと思われる要約に関しては，添削作業を最後まで行わず，放棄をしてよいという指示を被験者に与えている．

6 実験結果

6.1 順位づけ評価の結果

順位づけ評価の結果を表1に示す．ここで，各々の順位の平均値は必ずしもシステムの性能を正確に評価する基準とはなり得ない事に注意されたい．順位づけ評価では各トピック毎に，2つのbaseline, upperboundと各システムによる出力を1位から4位に順位づけをし，これをトピックに亙って平均値を求めている．しかしながら順位は順序尺度であり，距離尺度ではないので，平均値による比較は目安に過ぎない．一つのシステムを評価する際に，baseline, upperboundを利用する場合には，それらと個別に比較した時のそのシステムの優位性を評価する必要性がある．つまり，評価対象のシステムと，baseline, upperboundとの直接比較が必要である．すなわち，今回の順位づけ評価においては，1位から4位という順位自体は重要な意味を持たず，評価対象のシステムと，baseline, upperboundのそれぞれと比較した時の順位の優劣に意味がある，と我々は考える．従って，我々は平均順位による評価は参考程度に位置付け，提案システムをトピックごとにbaseline, upperboundのそれぞれとと比較した時の優劣を評価基準として考える．その結果を表2に示す．

また，評価結果を分析した所，我々のシステムは対象とする文書集合の数が少ない時により評価が高い事がわかった．そこで我々は7文書以下の文書集合から構成されるトピックに限定した評価も行なった．7文書以下のトピックは与えられた30トピックの内で半数である15トピックに相当する．この結果を表3に示す．

表1: 順位づけ評価の結果 (30トピック)

	C short	R short	C long	R long
平均順位 (a)	2.53	3.10	2.73	3.30
(a) の順位	2	8	5	8

平均順位 (a): baseline, upperboundとの比較による順位の平均値

(a) の順位: 9つの参加システム中における(a) の順位

表2: baseline, upperboundとの比較による優劣 (30トピック)

	C Short			R Short		
	W	L	T	W	L	T
v.s. 人手	7	22	1	8	20	2
v.s. Lead法	**15**	**12**	**3**	6	21	3
v.s. Stein法	**16**	**12**	**2**	7	22	1
	C Long			R Long		
	W	L	T	W	L	T
v.s. 人手	9	21	0	7	21	2
v.s. Lead法	**17**	**13**	**0**	10	20	0
v.s. Stein法	11	18	1	3	27	1

W:win L:lose T:tie

表3: baseline, upperboundとの比較による優劣 (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

6.2 人手による添削結果

添削結果として，1テキストあたりの平均添削文字数を表4に示す．表4より，比較対象とした全てのシステムでの平均値と比べて，全体的に添削文字数が多い事がわかる．

表4: 1テキストあたりの平均添削文字数

	short		long	
	提案手法	Ave.	提案手法	Ave.
削除				
C[%]	24.4	18.6	21.5	16.9
R[%]	0.4	1.6	0.9	1.7
挿入				
C[%]	16.2	25.9	18.1	19.5
R[%]	1.4	1.0	0.5	1.3
置換				
C(削除)[%]	4.0	2.0	2.9	1.5
C(挿入)[%]	9.1	2.9	4.6	2.0
R(削除)[%]	1.0	0.6	0.5	0.5
R(挿入)[%]	1.0	1.2	0.4	0.3

Ave.: 全システムの平均値

7 考察

7.1 内容の網羅性

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，以下の状況下においてその傾向が顕著に表れている．

1. 要約率が小さい時
2. 対象とする文書集合の数が少ない時

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に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．これは，MMRにおける冗長

性を制御する変数である λ の自動決定法が不完全である事が一因であると考えられる．今回，我々は式(4)で示したように，冗長性削除よりも重要文抽出を優先する為に λ に 0.5 のバイアスを掛けている．その為 λ が比較的大きな値を取り，MMI-MS の効果が押えられる傾向が見られた．この結果，対象とする文書集合の数が大きくなった時に，MMI-MS の効果が押えられている為に，全ての文書から重要な文を抽出する事を優先する傾向が見られた．

7.2 可読性

人手による添削評価より，我々のシステムと他システムを比較すると，添削された数が多い．この原因としては，我々のシステムでは重要文抽出のみによる要約生成を試みている事が考えられる．文間の結束性の考慮や言い換えのような別の手法を採り入れる事で改善が見込まれる．

しかしながら，可読性に関する評価を考える際に，我々の方針と TSC2 タスクオーガナイザの方針に差異が生じている点に注意されたい．TSC2 タスクオーガナイザによる可読性に対する位置付けは，「単一文書としての文書の読みやすさ」であった．一方で，我々は「サブトピックによるグループを積極的に探し，その単位で要約を行なう．さらに，読み進める為の指針としてキーワードを与える」という方針であり，あえて単一文書にしないようにした．この可読性の位置付けに対する差異は，我々のシステムの出力結果において，サブタイトルとしてのキーワードの部分の大多数が添削の際に削除されている事などからも伺う事が出来る．

8 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，内容の網羅性と可読性を併せ持つ原文の代わりとなる要約を提示するシステムを提案した．内容の網羅性の向上に対しては，情報利得比を用いた語の重要度と MMR の統合を，可読性の向上に対しては，重要文の話題ごとの分類と，話題ごとのクラスタ説明の為のキーワード付与，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，対象とする文書集合の数が小さい時 (7 文書以下の時) に効果的である事が示された．

今後の課題として，文書集合が大きい場合についての改善を検討したい．さらに現在のシステムに文間の結束性を保つ機構や言い換え手法などを採り入れる事による，可読性向上を考慮した要約生成も考えたい．

謝辞

国立情報学研究所主催の NTCIR ならびに TSC を企画・運営し，評価用データを作成していただいた皆様に感謝致します．なお，本研究を遂行するにあたって，CD-毎日新聞 98 年版，99 年版を利用させていただきました．使用許諾をしていただいた毎日新聞社，ならびに，同データの研究利用に対して御尽力いただいた皆様に感謝致します．本研究の一部は文部科学省科学研究費特定領域研究「IT の深化の基盤を拓く情報学研究」(課題番号 14019041) により支援を受けております．

参考文献

[BEM01] Regina Barzilay, Noemie Elhadad, and Kathleen R. MaKeown. Sentence ordering in multidocument summarization. In *Proceedings of the the first International Conference on Human Language Technology Research (HLT 2001)*, pp. 149–155, 2001.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* ACM press, Addison Wesley longman Ltd., 1999.

[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.

[FBY92] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval – Data Structure & Algorithms.* Prentice Hall PTR, 1992.

[GMCK00] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-Document Summarization by Sentence Extraction. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, pp. 40–48, 2000.

[MGB99] Inderjeet Mani, Barbara Gates, and Eric Bloedorn. Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL 99)*, pp. 558–565, 1999.

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

[RF00] Dragomir R. Radev and Weiguo Fan. Automatic summarization of search engine hit lists. In *Proceedings of ACL Workshop on Recent Advances in NLP and IR*, 2000.

[RJB00] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of ANLP/NAACL Workshop on Automatic Summarization*, 2000.

[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.

[大塚 01] 大塚敬義, 内海彰, 廣田薫. 要約文生成における照応処理. 言語処理学会第 7 回年次大会, pp. 425–428, 3 月 2001.

[奥村 99] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向. 自然言語処理, Vol. 6, No. 6, pp. 1–26, 1999.

[奥村 02] 奥村学, 難波英嗣. テキスト自動要約に関する最近の話題. 自然言語処理, Vol. 9, No. 4, 7 月 2002.

[菊池 01] 菊池美和, 吉田和史, 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. 言語処理学会第 7 回年次大会, pp. 189–192, 3 月 2001.

[難波 99] 難波英嗣, 奥村学. 書き換えによる抄録の読みやすさの向上. 自然言語処理研究会 NL-133-8, 情報処理学会, 1999.

[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, 7 月 2002.