検索結果表示向け文章要約における情報利得比に基づく語の重要度計算

菊池 美和　　吉田 和史　　森 辰則
横浜国立大学 工学部 電子情報工学科
E-mail: {miwa,kazu,mori}@forest.dnj.ynu.ac.jp

1 はじめに

近年，情報検索システムが広く利用されているが，検索結果には要約文書が付与されていることが多い．

しかしながら，それは，原文書の最初の数バイトを出力したり，検索要求文に含まれる語の周囲のみを提示するといった単純な方法を採用しているので，十分な品質の要約を提供できていないことが多い．

そこで本稿では，情報検索の結果として得られた文書集合中の各々の文書を要約する一手法について提案する．検索された文書間の類似性構造を説明するのに寄与する単語に重きをおく重みづけを行なう．特に，その重みづけに決定木学習アルゴリズム C4.5 で導入された情報利得比を用いることを提案する．そして，この重みづけと他の重みづけを組み合わせることにより，重要文を抽出し，検索結果の文書を要約する手法について述べる．

2 検索結果文書を利用した語の重みづけ

検索結果文書の要約は次の点で単一文書要約と異なる．

- 検索要求文が与えられている．
- 複数の文書が同時に与えられ，一度の検索結果という点で文書間に類似性がある．

上記の手掛かりを語の重みづけに用いることを考えると，まず考えられる手法は，Tombros ら [TS98] が提案するように，検索要求中の語を重要語として考え，他の語よりも重みを高する方法である．しかし，この方法は，各種フィードバックや検索要求の拡張など検索エンジンにおける工夫が反映されないという問題がある．また，検索要求と関連性の低い文書においては，検索要求中の語がほとんど出現しないので効果を発揮しないと考えられる．

そこで，我々は，検索要求文を使わずに，検索文書集合のみを用いて重みづけすることを考える．すなわち，検索結果文書集合には検索要求に関する情報が暗に含まれていると考えられるので，その情報を引き出せばよい．

本稿では，次の二つの指針からなる手法を提案する．

1. 検索結果の文書集合に対し階層的クラスタリングを行い，文書間の類似性の構造を抽出する．
2. その類似性構造に基づき語の重みづけを行う．

1．においては，検索要求に関連する文章とそうでない文章がクラスタ構造の中で分離されることが期待され，なおかつ，それらの文書集合においても類似性に基づく細分類がなされると考えられる．ただし，ここで注意すべきことは，検索されなかった文書の存在を類似性構造の中に組み込む必要があることである．なぜならば，クラスタ構造において，一番上位のクラスタは与えられた構造として扱う以外になく，類似性の解析の対象とならないからである．よって，検索結果の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ものとする．そのクラスタには，検索結果の文書の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．

このようにして求められた類似性構造は文章を一つの単位とするマクロな情報であるので，要約のためには，これを文や句，単語を単位とする，よりミクロな情報に還元する必要がある．これが 2．である．

本稿では，1．については，語を次元とする文書ベクトルの類似度による階層的クラスタリングを用いる．2．については，各語に注目し，部分クラスタへの分割における寄与の度合を出現確率に基づく情報利得比により表現する．

このようにして求められた情報利得比を，既存の方法で用いられている重みである，語の文書内頻度 TF，文書頻度の逆数 IDF と組み合わせることにより，総合的な語の重要度を与える．

2.1 最大距離法による階層的クラスタリング

検索された文書集合の類似性の解析には，文書間の距離の定義とその類似性に基づく文書集合の構造化が必要となる．本稿では文書間の類似性として，$tf \cdot idf$ 法ならびにベクトル空間法に基づく方法を採用する．また，文書集合の類似性に関する解析には，階層的クラスタリングを用いる．階層的クラスタリングアルゴリズムには，文書間距離の絶対値をクラスタ中心の選択に反映させることができる最大距離法 [長尾 83] を採用した．このアルゴリズムは非階層的なクラスタを生成するが，これを各部分クラスタに対して再帰的に適用することにより，階層的なクラスタ構造を生成する．

クラスタリングでの文書間距離には，次に述べるユークリッド距離を用いる．まず，各文書 D_i はベクトル空間モデルに基づき，n 次元空間上の点 $(weight_{i1}, weight_{i2}, \ldots, weight_{in})$ で表す．$weight_{ik}$ は文書 D_i において語 w_k に割り当てられた重みである．重み $weight_{ik}$ としては語 w_k の $tf \cdot idf$ 値とする．このとき，文書 D_i と文書 D_j の距離 d を次のように定義す

る．

$$
\begin{aligned}
d(D_i, D_j) &= \sqrt{\sum_k^n (weight_{ik} - weight_{jk})^2} && (1)\\
weight_{ik} &= tf(D_i, w_k) idf(w_k) && (2)\\
tf(D_i, w_k) &= \frac{freq(D_i, w_k)}{|D_i|} && (3)\\
idf(w_k) &= \log_2 \frac{N}{df(w_k)} && (4)
\end{aligned}
$$

ただし，本稿では語として名詞のみに注目し，以下のように定義する．

$freq(D_i, w_k)$： 文書 D_i での語 w_k の出現頻度
$|D_i|$： 文書 D_i 中の名詞の数
$df(w_k)$： 検索対象の全文書集合における語 w_k を含む文書数
N： 検索対象の全文書の数

文書からの名詞抽出には，形態素解析器 JUMAN version 3.61 を用いた．また，$df(w_k)$, N は検索対象である毎日新聞 1994 年，1995 年，1997 年，1998 年のすべての記事から求めた．

2.2 情報利得比に基づく語の重要度

クラスタの木における各接点 (内点) は，あるクラスタとそれを分割して得られた互いに素な部分クラスタの関係，すなわち，クラスタの分割の仕方を示している．この分割の仕方はクラスタ内の文書の類似度に従って決定されるので，これを文書内の語の重みに反映させることができれば，複数文書間の類似性というマクロな情報を，文書内の語の重みというミクロな情報に還元できると考えられる．その方法として，我々は次の 2 つの段階からなる手法を提案する．

1. 各クラスタについて，その部分クラスタの構造から，各語の重みを決定する．
2. 一つの文書は，クラスタの木の根接点から対応する葉接点に至るクラスタ分割の系列によって指し示される．よって，各文書における語の重みは，各分割で得られた語の重みを統合して得る．

このうち，特に重要なのは 1．である．その基本的な考え方は，クラスタの分割構造を決定するのに寄与する語に高い重みを与えるというものである．本稿では，この寄与の度合を，語の出現分布とクラスタ構造が一致する度合として捉え，その度合を表す尺度として情報利得比を用いる．

2.2.1 情報利得比

情報利得比は，決定木学習システム C4.5 において属性選択を行なうために導入された．我々は，表 1に示す対応の下，クラスタの構造を決定木の構造と見なすことにより，情報利得比を用いる．C4.5 においては属性の評価値として情報利得比を用いていたが，我々の方法においては，属性ではなくクラスに対応する単語に対する評価値として情報利得比を用いる．

表 1: 我々の方法と C4.5 における方法の対応

我々の方法	C4.5
クラスタの分割構造	属性によるテスト
単語の出現確率	クラスの出現確率

C_i をクラスタ 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 $gain\_r(w, C)$ は次の様に求められる．

$$
\begin{aligned}
gain\_r(w, C) &= \frac{gain(w, C)}{split\_info(C)} && (5)\\
gain(w, C) &= info(w, C) - info_{div}(w, C)\\
info(w, C) &= -p(w|C)\log_2 p(w|C)\\
&\quad -(1 - p(w|C))\log_2(1 - p(w|C))\\
p(w|C) &= freq(C, w)/|C|\\
info_{div}(w, C) &= \sum_i \frac{|C_i|}{|C|} info(w, C_i)\\
split\_info(C) &= -\sum_i \frac{|C_i|}{|C|} \log \frac{|C_i|}{|C|}
\end{aligned}
$$

2.2.2 情報利得比に基づく語の重要度

式 (5) に示される情報利得比は，各クラスタの分割毎に得られる．本稿では，すべての検索結果文書を同時に要約し，一覧形式でユーザに提示するという最も基本的なインタフェースを想定し，式 (6) に示す情報利得比の和を採用する．この方法では，すべてのクラスタ分割における情報利得比を等しく考慮する．

$$
\begin{aligned}
igr(w, D) &= \sum_{C \in Cs(D)} gain\_r(w, C) && (6)\\
Cs(D) &= \text{文書}D\text{の属するすべてのクラスタの集合}
\end{aligned}
$$

以上で定義された情報利得比に基づく重みにより，文書 D 中の語 w の重要度 $weight(w, D)$ を定義する．前述の通り，語の重要度には tf，idf，igr の各値の組み合わせを考えるが，各値が独立に重要度に寄与すると考え，組み合わせ方法として積を用いる．

$$
weight(w, D) = igr(w, D) \cdot tf(w, D) \cdot idf(w) \qquad (7)
$$

3 評価実験

本節では NTCIR2 Text Summarization Challenge(TSC) での情報検索タスクに基づく実験評価を行ない，本手法の評価を行なう．

3.1 重要文抽出に基づく要約文書生成

本実験では，我々の手法の有効性を示すために，次に示す，語の重要度だけによる最も基本的な要約手法を用いた．

1. 各文中の名詞の重要度の平均値を求め，それを文の重要度とする．

$$s\_imp(s, D) = \frac{\sum_{w \in s} weight(w, D)}{|s|} \quad (8)$$

2. 重要度の高いものから順番に文を選択することを繰返し，ある決められた文書の長さに達したら，選択した文を文書中での出現順に並べなおして，終了とする．

さらに，上記の手順に以下の条件を加えた．

- 要約を一覧形式で提示することを想定すると，要約文書の長さが統一されているほうが，見やすい．そのため，要約文書の長さは要約率ではなく，絶対的な長さにより決定する．具体的には，150 形態素をしきい値とする．
- 文が省略されている箇所には「...」を加え，原文書の段落終了箇所には改行を加える．

3.2 情報検索タスクにおける実験結果

評価は NTCIR2 TSC における「課題 B IR タスク用要約」での結果に基づいて行なう．

TSC 実行委員会より配布されたデータセットには，12 の主題があり，それぞれ，検索要求 1，検索文書 50 文書が含まれている．これらの文書は 1994 年，1995 年，1998 年の毎日新聞の記事の集合から検索されたものである．TSC 実行委員会は各文書に対して，検索要求に対する関連性評価を別途行ない A(適合),B(関連),C(無関係) の三段階を付与した．当然，当初は TSC 参加者には非公開である．TSC の参加者は各自のシステムを用いて，これらの文書を要約し，事務局に提出した．これら提出された要約文書と検索要求に対して，TSC 事務局が被験者 36 名 (学生) による関連性評価を行なった．被験者らには関連性の有無という二段階で提示してもらった．よって，両者の一致の判定においては，A 判定の文書だけを関連文書とする場合 (Answer Level A) と A 判定に加えて B 判定の文書も関連文書とする場合 (Answer Level B) が考えられる．

表 2に実験結果を示す．評価尺度には，被験者が 1 検索要求に関するタスク (50 文書) に要した時間 (TIME)，タスクをどの程度適切に行なえたかを示す指標 (再現率 (Recall)，適合率 (Precision)，F 値 (F-Measure))，要約文書の長さ (1 文書あたりの平均文字数,LENGTH) を用いた．

4 考察

情報検索の結果の文書に対する要約においては，利用者が行なう適合性判断のための時間の短さと，適合性判断の正確さが共に達成されることが必要である．そこで，まず，タスク遂行時間について簡単に考察し，次にタスクの精度について詳しく考察する．

4.1 タスクに要する時間

我々のシステムが生成した要約における，適合性判定に要した時間は，1 トピック (50 文書) あたり 8 分 33 秒であった．すべての参加システムの平均タスク時間は 1 トピック当たり 9 分 8 秒 であり，我々の要約の適合性判定に要する時間はこれよりも短い．したがって，次に述べる精度の比較においては，時間は考慮せずに，各評価値を直接比較する．これによって，我々のシステムに有利になることはない．

4.2 タスクの精度

4.2.1 Answer Level A

Answer Level A では，我々の手法は，再現率，適合率，F 値すべてにおいて，他のすべての参加システムよりも高い値を示している．ベースラインシステムとの比較においては，我々のシステムの適合率は適合率重視の Lead 手法よりも 1.5 ポイント低い値を示しているものの，それ以外は勝っている．

検索要求によるバイアスを採り入れた TF 法と比較してみると，再現率において 10.9 ポイント，適合率において 2.7 ポイント，F 値において 7.0 ポイント勝っている．これは，検索文書の要約において検索要求を使用しなくても，検索文書群だけで同等以上の質の要約が可能であることを示している．

4.2.2 Answer Level B

本節では Answer Level B について考察を行なう．他のシステムと比較して，再現率が第 2 位と高いものの，適合率は第 7 位，F 値は第 4 位と相対順位が低くなった．Answer Level B の評価においては，Answer Level A よりも正解の数が多くなるので，一般に，Answer Level A に比べて，再現率が下降し，適合率が高くなる．再現率についていえば，Answer Level A において，高い適合率となったシステムほど減少が激しくなる．一方，適合率については，B 判定のものが Answer Level A での誤判定となっているのであれば，その値の上昇が著しい．

我々の場合，再現率が 0.907 から 0.754 へと激しく低下しているが，順位が 2 位であるので相対的には他のシステムよりも高いことがわかる．つまり，正しく関連性の判定が行なわれた要約文書の数は他のシステムよりも多い．一方で，適合率の上昇は他のシステムより低いので，C 判定文書の要約に対しても適合であると判定を下した数が多かったことになる．

表 2: 総合評価一覧

	Our System	Sys 1	Sys 2	Sys 3	Sys 4	Sys 6	Sys 7	Sys 8	Sys 9	Fulltext	TF	Lead
Recall (Ans. A)	0.907	0.833	0.899	0.793	0.818	0.858	0.831	0.824	0.849	0.843	0.798	0.740
Precision (Ans. A)	0.751	0.728	0.717	0.685	0.674	0.718	0.739	0.738	0.741	0.711	0.724	0.766
F-Measure (Ans. A)	0.808	0.761	0.785	0.715	0.718	0.763	0.766	0.749	0.768	0.751	0.738	0.731
Recall (Ans. B)	0.754	0.741	0.793	0.715	0.737	0.745	0.719	0.719	0.752	0.736	0.700	0.625
Precision (Ans. B)	0.897	0.921	0.904	0.898	0.875	0.892	0.908	0.913	0.923	0.888	0.913	0.921
F-Measure (Ans. B)	0.797	0.808	0.828	0.776	0.773	0.785	0.779	0.775	0.805	0.773	0.776	0.712
TIME	8:33	9:41	12:48	6:25	6:44	9:01	10:16	9:16	9:31	13:46	8:44	7:32
LENGTH	234.4	297.8	585.7	89.5	136.4	288.4	292.9	266.1	262.5	819.4	253.6	174.5

Sys 1 から Sys 9: TSC 参加の他システム
Ans. A, Ans. B: Answer Level A, Answer Level B にそれぞれ対応
Fulltext: 原文書
TF: TF による重要文抽出手法．検索要求中の単語に 2 倍の重み．要約率 20%(文ベース)
Lead: 先頭から 20% の文を抽出する手法．タイトルは出力しない．

以上より結論されることは，我々の手法は，平均的な要約率において，やや再現率を重視した要約を生成する方法であると考えることができる．

5 関連研究

本研究では，検索結果文書の情報を利用したが，その点で，Eguchi ら [KHAY99]，Fukuhara ら [THT99] の手法が関連する．

Eguchi らは，適合性フィードバックに基づく検索システムを構築している．このシステムでは，検索結果を文書間の類似度に基づいてクラスタリングし，各クラスタごとにクラスタに多く含まれる語と，そのクラスタを代表する文書のタイトルを，そのクラスタの要約として利用者に提示する．利用者には，その情報を元に関連するクラスタを選択してもらい，そのクラスタ内の文書を用いて適合性フィードバックを行なう．

Fukuhara らの手法でも，検索結果文書をクラスタリングし，文書中の単語の出現頻度に基づく skewness と kurtosis という尺度を用いてクラスタごとにトピックを表す語を抽出する．そして，それらトピックを含む文を抽出し，焦点–主題連鎖を考慮して並べ替え，各クラスタの要約を出力している．

これらの手法では，クラスタリングを文書のグループ分けのみに利用していて，直接語の重みには反映させていない．語の重要度としては単純にクラスタ内の頻度情報を用いているだけである．我々の手法においては，クラスタ間の構造の情報も取り入れて重みづけをしているので，この点において類似性構造をより反映していると考えられる．

6 まとめ

本稿では，複数の検索文書の間に存在する類似性の構造を階層的クラスタリングにより抽出し，その構造を適切に説明するか否かに応じて語に重みをつける手法を提案した．タスクに基づく評価実験の結果，この方法に基づく重要文抽出型の要約手法は，検索文書の要約において，非常に有効であることが示された．

今後の課題としては，対話形式のインターフェースへの利用が挙げられる．

参考文献

[KHAY99] K.Eguchi, H.Ito, A.Kumamoto, and Y.Kanata. Adaptive Query Expansion Based on Clustering Search Results. 情報処理学会論文誌, Vol. 40, No. 5, pp. 2439–2449, 1999.

[THT99] T.Fukuhara, H.Takeda, and T.Nishida. Multiple-text Summarization for Collective Knowledge Formation. In *Workshop on Social Aspects of Knowledge and Memmory*. Man and Cyberbetics Conference, IEEE Systems, 1999.

[TS98] A. Tombros and M. Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 2–10, 1998.

[長尾 83] 長尾真. パターン情報処理, pp. 116–117. 電子通信学会大学シリーズ. コロナ社, 1983.