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、検索結果の表示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。検索された文書群の類似性構造を説明するために寄与する単語に重きをおく重みづけを行う。特に、その重みづけに決定するアルゴリズム 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。
1. 検索要求文が与えられている。
2. 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、検索要求文中の語を重要度として考え、他の語よりも重みを高める方法である。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、検索要求と関連度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、検索文書集合成分のみを用いて重みづけることを考える。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、その情報を引き出すべきである。
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して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化がなされると考えられる。ただし、ここでの注意すべきことは、検索されていない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、一段階上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り得ないからである。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一つの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。
このようなにして求められた類似性構造は一個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、これを文や単語を単位とする、よりミクロな情報に還元する必要がある。これが2で、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性による階層的クラスタリングを用いる。2については、群に注目し、部分クラスタへの分割における寄与の度合を表現し、検索結果の文書集合から得られた情報利得比を表現する。
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文章でである。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-idf 等により文書間距離を用法に基づく距離を採用する。
また、文書集合の類似性に基づくには、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クラスタを生成するが、これを各部分クラスタに対して階層的に適用することにより、階層的なクラスター構造を生成する。
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 次元空間の頂点 (weight_1, weight_2, ..., weight_n) で表す。weight_i は文書 $D_i$ において語 $a_i$ に割り当てられた重みである。重み weight_i は文書 $D_i$ の tf-idf 値とされる。このとき、文書 $D_j$ と文書 $D_i$ の距離 $d$ は次のように定義す
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.

\[
d(D_i, D_j) = \sqrt{\sum_{k} (weight_{ik} - weight_{jk})^2}
\]

(1)

\[
weight_{ik} = tf(D_i, w_k) \cdot idf(w_k)
\]

(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)

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
スタをそれそのの集合として得られた互いに無部分クラス
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成か
ら，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積算して得る。

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，この寄与の度合いを，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する構造を見

表1：我々の方法とC4.5における方法の対応

<table>
<thead>
<tr>
<th>我々の方法</th>
<th>C4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>クラスタの分割構造</td>
<td>属性によるテスト</td>
</tr>
<tr>
<td>単語の出現確率</td>
<td>クラスタの出現確率</td>
</tr>
</tbody>
</table>

\[C_i\]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。

\[gain_r(w, C) = \frac{gain(w, C)}{split_{info}(C)}\]

(5)

\[gain(w, C) = info(w, C) - info_{dis}(w, C)\]

\[info_{dis}(w, C) = -\sum_i \frac{|C_i|}{|C|} \log \frac{|C_i|}{|C|}\]

\[split_{info}(C) = -\sum_i \frac{|C_i|}{|C|} \log \frac{|C_i|}{|C|}\]

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。

\[igr(w, D) = \sum_{C \in C(D)} gain_r(w, C)\]

(6)

\[Ce(D) = 文書Dの属するすべてのクラスタの集合\]

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，組み合わせ方として積を用いる。

\[weight(w, D) = igr(w, D) \cdot tf(w, D) \cdot idf(w)\]

(7)

3 評価実験

本稿ではNTCIR2 Text Summarization Challenge(TSC)での情報検索タスクに基づく実験評価を行い，本文中の評価を行なう。
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. 各文書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。

\[ s_{\text{imp}}(s, D) = \frac{\sum_{w \in s} \text{weight}(w, D)}{|s|} \]  

(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て、

終了とする。

さらに、上記の手順に以下の条件を加える。

- 要約を一視点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が、見やすい。そのため、要約文書の長さは要約度ではなく、適切な長さにより決定する。具体的には、
- 150 形態素を基本値とする。

- 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既往から抽出されたものである。TSC 実行委員会は各文書に対し、

で、検索要求に対する関連性評価を別途行なう A(関合), B(関連), C(無関係) の三段階を付与した。当然、

当時は TSC 参加者には非公開である。 TSC の参加者は各文のシステムで用いて、これらの文書を要約し、

事務局が提出した。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。被験者らは関連性の有無という三段階で評価してもらった。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

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。

実験成績結果を示す。評価尺度には、被験者が 1 検索要求に関する文書 (50 文書) を要した時間 (TIME),

文書の程度を示す指標 (要約率 (Recall), 適合率 (Precision), F 値 (F-Measure))

と要約文書の長さ (1 文書あたりの平均文字数, LENGTH) を用いた。

4 考察

情報検索文書の要約においては、利

用者が行う必要性を判断のための時間内は短い、適合

性の正確さが必要であることが必要である。

そこで、まず、タスク有効性について単純に考察し、

次に本研究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文における、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。全ての参加システムの平均タイムは 1 トピック当り 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であることがない。

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、現実率、適合率、F値を示し、全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、

しているものので、それ以外は確保している。

検索要求によるパワースペクトル化を行い、TF 法と比較してみると、現実率において 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

を示している。これは、検索文書の要約において検索要

求を使用し、検索文書だけでなく同等以上の質

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、現実率が 1 位と高いもの

の、適合率は 7 位、F値は 9 位と相対に低くなっ

た。Answer Level B の評価においては、Anwe

Answer Level A で高い数値が多かったので、次に、

Answer Level A において、現実率が高くなり、現実率についていけば、Answer

Answer Level A において、高い適合度をえた回答はTracking

にとまった。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、

その他の上位差が低い。

我々の場合は、現実率が 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ある。つまり、適

正性の判定が実現さ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

よりも低いので、C 判定文書の要約に関しては適

合する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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.907</td>
<td>0.833</td>
<td>0.899</td>
<td>0.793</td>
<td>0.818</td>
<td>0.885</td>
<td>0.831</td>
<td>0.824</td>
</tr>
<tr>
<td>0.751</td>
<td>0.728</td>
<td>0.717</td>
<td>0.685</td>
<td>0.674</td>
<td>0.718</td>
<td>0.739</td>
<td>0.738</td>
</tr>
<tr>
<td>0.808</td>
<td>0.761</td>
<td>0.785</td>
<td>0.715</td>
<td>0.718</td>
<td>0.763</td>
<td>0.766</td>
<td>0.749</td>
</tr>
<tr>
<td>0.754</td>
<td>0.741</td>
<td>0.793</td>
<td>0.715</td>
<td>0.737</td>
<td>0.745</td>
<td>0.719</td>
<td>0.719</td>
</tr>
<tr>
<td>0.897</td>
<td>0.921</td>
<td>0.904</td>
<td>0.898</td>
<td>0.875</td>
<td>0.892</td>
<td>0.908</td>
<td>0.913</td>
</tr>
<tr>
<td>0.797</td>
<td>0.808</td>
<td>0.828</td>
<td>0.776</td>
<td>0.773</td>
<td>0.785</td>
<td>0.779</td>
<td>0.775</td>
</tr>
<tr>
<td>234.4</td>
<td>297.8</td>
<td>595.7</td>
<td>80.5</td>
<td>136.4</td>
<td>298.4</td>
<td>292.9</td>
<td>266.3</td>
</tr>
<tr>
<td>819.4</td>
<td>253.6</td>
<td>174.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sys 1 から Sys 9: TSC参加の候補システム
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, 検索結果中の単語に 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, タイトルは欠如しない。

以上より結論することは, 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, 合理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, 各クラスターごとにクラスターに多く含まれる語と, そのクラスターを代表する文書のタイトルを, そのクラスターの要約として利用者に提示する。利用者には, その情報と文書間の類似度に基づいて文書を検索することを応用して, そのクラスターの文書を用いて適合性フィードバックを広げる。

Fukuhara らの手法でも, 検索結果文書をクラスタリングし, 文書間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ーごとにトピックを発表する手法を挙げており, それぞれトピックを抽出し, 焦点-主題関連を考慮して並べ替え, 各クラスの要約を生成している。

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のみに利用している。直接的評価に反映されていない。重要な点として, 評価条件としてされるクラスター内の類似的文書を用いているだけである。我々の手法においては, クラスター間の情報の取り入れて重みづけをしていますので, 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の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。クラスター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, 対話形態のインターネットへの利用が挙げられる。

参考文献


