3Y－8

視点を考慮した文書要約手法の提案

塩見 隆一　徳田 克己　青山 昇一　柿ヶ原 康二

松下電器産業（株）マルチメディア開発センター

1　はじめに

一般に、文書検索システムでは、文書データに対応する抄録データを用意している。検索者はキーワード検索などで文書データを絞り込んだあと、抄録データを参照し、検索された文書データの要・不要を判断する。しかしながら、用意された抄録は、検索者の意図とは無関係に作成されているため、検索者が文書データの要・不要を判断するのに十分なものとなっていない場合がある。

我々は、統計的な手法を用いた要約手法[1]を改良し、検索に用いられたキーワードを用いて、検索視点を反映した要約文書を作成する手法を提案する。

2　要約文書作成手法

2.1　従来アルゴリズム

要約文書の従来の作成方法[2]として、以下の手順を採用した。

1. **文・単語分割**

   句読点、記号を手がかりに文書を文に分割し、更に形態素解析を行ない単語に分割する。形態素解析は当社で開発したもの[3]を使用した。

2. **単語重要度の決定**

   単語 i の単語重要度 w_i は、次式で与えられる基本的な tf*idf を用いた。

$$w_i = tf_i * log\frac{N}{n_i}$$

   ここで、tf_i は、単語 i の文書内頻度、N は全文書数、n_i は、単語 i の文書頻度である。

3. **文重要度の決定**

   文 j の文重要度 s_j は単語重要度の和を文中の単語数で割る次式を用いた。

$$s_j = \frac{\sum_{i=1}^{M} w_i}{M}$$

   ここで、M は、文中の単語数である。なお、対象とする単語として助詞と助動詞は除外した。

4. **文の抜粋**

   必要な文数だけ文重要度の高い文を抜粋し、記事中での文の出現順に配置して要約文書を作成する。ただし、要約対象文書を新聞記事としたため、タイトル文と記事の第１文を特別扱いとし、必ず抜粋することとした。

2.2　改良アルゴリズム

検索視点を反映するため、従来アルゴリズムの単語重要度 w_i を次式のように変更した。

$$w_i = tf_i * log\frac{N}{n_i} + C(i)$$

ここで、$C(i)$ は単語 i がキーワード検索で用いられたキーワードと一致すれば定数 C、一致しなければ０である。この改良により、キーワード検索で用いられた単語を含む文の文重要度が高くなり要約文書に含まれる可能性が高くなる。

3　検索実験

定数 C と要約文書の文数を変え、被験者実験を行ない、改良アルゴリズムの効果を調べた。

3.1　データ

実験を行なう検索対象データとして、「情報検索システム評価用ベンチマークＶｅｒ１．０（ＢＭＩＲ－Ｊ１）[1]」を用いた。このデータは、日本経済新聞から抽出した新聞記事６００件、検索要求文６０件、検索要求文に対応する正解記事集合で構成されている。

また、ＢＭＩＲ－Ｊ１は検索要求文に対する正解記事をＡランクとＢランクで定義している。Ａランクは記事の主題が検索要求文の内容と一致するもの、Ｂランクは記事の主題と一致はしないが、検索要求文の内容を含むものである。

Ａランクの正解記事では、記事中の重要語と検索者が入力するキーワードが一致する可能性が高い。その結果、従来アルゴリズム、改良アルゴリズムのどちらを用いても類似の要約文書ができる可能性が高いと考えられる。本手法が効果を発揮するのは記事中の重要語と検索者が入力するキーワードが一致する可能性が低いＢランクの記事に対してと考えられる。そこで、実験には、構文解析や意味解析などを必要としない検索要求文から、Ｂランクの正解記事が多い６つの検索要求文を使用した。

An Abstruction Method Using Viewpoints
Takakazu Shiomi, Katsumi Tokuda,
Shoichi Aoyama, Koji Kakigahara,
Matsushita Electric Industrial Co., Ltd.
1006 Kadoma, Kadoma, Osaka 571-8501, Japan

[1]株式会社 日本経済新聞の協力によって、社団法人 情報処理学会データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、１９９３年９月１日から１２月３１日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）を利用

3.2 実験パラメータ

要約文書を作成する際の定数 Cの値3種類(0,15,100)と文数2種類(文書内文数/3, 5)を組み合わせた6種類の実験ケースを用意した。以下は、実験ケースをまとめたものである。

実験ケース	P1	P2	P3	P4	P5	P6
定数 C	0	15	100	0	15	100
文数	文書内文数/3			5		

定数 $C=0$ の実験ケースは、従来アルゴリズムでの要約を意味する。定数 $C=100$ の実験ケースは、検索キーワードを含む文が、ほぼ必ず抜粋されるような要約文が作成される。定数 $C=15$ の実験ケースは、その中間である。

要約文書の長さは、元の文書の長さの1／3のものと、一定文数「5」のものについて比較することにした。なお、要約文書の文数にはタイトル文と記事の第1文を含んでいる。

3.3 被験者及び実験手順

被験者は、業務でキーワード検索を利用している6名とした。各被験者が使用する実験ケースは以下の通りとした。

検索要求文	1	2	3	4	5	6
被験者A	P1	P2	P3	P4	P5	P6
被験者B	P2	P3	P4	P5	P6	P1
被験者C	P3	P4	P5	P6	P1	P2
被験者D	P4	P5	P6	P1	P2	P3
被験者E	P5	P6	P1	P2	P3	P4
被験者F	P6	P1	P2	P3	P4	P5

被験者は検索要求文を読み、キーワード検索で記事を20件以下に絞り込む。ただし、検索式はANDのみ使用可能で、ORやNOTは使用できない。次に検索した各記事に対して、記事の要約文書を読み、以下の3段階で評価する。

- 検索要求文を満たす正解記事
- 検索要求文を満たさない不正解記事
- わからない

3.4 実験結果と考察

図1は、被験者が検索したBランクの正解記事に対する被験者の評価をまとめたグラフである。

実験ケースP1,P2,P3では、あまり大きな差がみられない。これは、元の新聞記事が長く、要約文書も長くなり、従来アルゴリズムでも判断材料となる文が要約文書に含まれているためと考えられる。また、要約文書が冗長なため、判断の難しいBランクの記事では「検索要求文を満たさない不正解記事」との誤った判断が50％程度になったと思われる。

一方、実験ケースP4,P5,P6では改良アルゴリズムの効果が伺える。要約文書がタイトル文と記事の第1文を含む5文で構成されていることから、従来アルゴリズムでは「検索要求文を満たす」と判断できる文が抜粋される可能性が低い。キーワード検索に使用された単語を含む文が抜粋された要約文書が、改良アルゴリズムによって作成され、被験者の正確な判断が可能となったと考えられる。

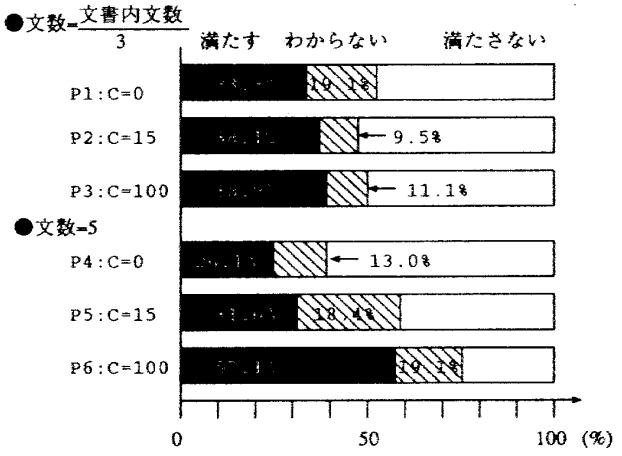

図1: 検索されたBランク記事に対する被験者の評価

なお、検索されたAランクの記事に対して、被験者が「検索要求文を満たさない正解記事」と誤って評価したのは全体でわずか1つであった。また、検索されたAランクでもBランクでもない記事のすべてに対して、被験者は「検索要求文を満たさない不正解記事」と正しく評価している。

4 まとめ

被験者実験を通じて、キーワードを利用した要約文作成方法の有効性を示すことができた。今後は以下の点について詳細に検討する必要がある。

- 定数 Cは必要か？
  検索キーワードを含む文から文重要度の高い文を抜粋する方法も考えられる。ただし、この方法で作成した要約文書はまとまりがなくなり、元の文書の全体像が掴めないということも考えられる。
- 定数 Cの制御が難しい。
  要約対象文書によって、抜粋される文の変化量と定数 Cの値の変化の関係が一定でない。定数 Cを文書内の平均単語重要度の定数倍にするなどの工夫が必要である。
- 要約文書の品質
  より高い検索視点の反映を実現するため、検索キーワードを基にシソーラスや単語共起を用いた意味間距離に応じて複数の単語の単語重要度を操作するなどの工夫が必要である。

参考文献

[1] Klaus Zechner :"Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences," COLONG '96, Vol.2, pp. 986-989, 1996.

[2] 隅田英一郎, 飯田仁 :" 統計的な抄録法を使った情報検索," 言語処理学会第3回年次大会発表論文集, pp.353-356,1997

[3] 杉村領一 :" 日英翻訳支援システム," National TECHNICAL REPORT, Vol.39, No.1, pp.80-88,1993