統計的な抄録法を使った情報検索

隅田 英一郎　飯田 仁
ATR 音声翻訳通信研究所

1 はじめに

膨大な電子化された情報の中から必要な情報を適切に検索する技術の研究が盛んに行なわれている。

一般に、情報検索の対象である文書は、その長さが一定でなく、主題や関連話題への言及の度合などが多様である。この多様性が情報検索システムの精度や効率を低下させる。

文書から自動的に主題に関連した重要なセグメント（文、段落など）を一定個抜粋した抄録を検索対象とするという枠組を用いれば、精度と効率を同時に改善できると考えられる。

著者等は統計的な文抄録アルゴリズムの一つを用いて、この枠組の有効性を検討した。

手続の概略は次の通りである。（1）単語の重要度を決定する。（2）単語の重要度を用いて文の重要度を決定する。（3）文書から一定個の重要な文を抜粋し、これを抄録とする。（4）この抄録を対象として検索する。

新聞の記事を基に作成された日本語の情報検索システム評価用ベンチマークを用いて行なった実験で、ベクトル空間型情報検索システムの精度を改善することを確認した。2節で情報検索と自動抄録の現状を概観し、3節で自動抄録を使った情報検索システムの一実現法について説明し、4節で実験結果を述べる。

2 情報検索と自動抄録

2.1 ベクトル空間モデル型検索

1992年より、検索システムの精度比較を目的とした国際的な活動としてTRECが行なわれている。そこで、上位の成績を収めるシステムの一つにベクトル空間型システム（VSM）がある[8]。

VSM[3]では、文書と質問を単語の重要度を要素としたベクトルで表現する。重要度は単語頻度（term frequency, tf）と文書頻度の逆数（inverse document frequency, idf）の積を用いて算出されることが多い。文書のベクトルと質問のベクトルの類似度（内積など）を計算し、この順に文書をランキングする。

一般に、情報検索の対象である文書は、その長さ[1]が一定でなく、主題や関連話題への言及の度合などが多様である。この多様性が情報検索システムの精度や効率を低下させる。

本実験ではこれらの問題点を解消するために、VSMに自動抄録を組み込む。さらに、検索結果として文書の抄録を提示すれば、ユーザはシステムのランキングの根拠が理解できるし、当該文書全体を読む前に、ユーザは自らの要求に合致しているか否か判定できるという利点もある。

2.2 自動抄録法

2.2.1 統計的手法

統計的手法は1958年のLuhn[2]の研究に始まる。その骨子は、非常に高頻度の語（いわゆるstoplistの語）を無視して、文の重要度を高頻度語の関数として定義し、重要度の高い上位の文を抜き出すというものである。処理は単純で速く、文書の種類、ドメイン、言語に依存した知識やヒューリスティクスが不要であるという長所がある。最近、Zechner[6]によって提案された手法（3.1節）は、小規模な実験ではあるが、再現率・適合率の観点で、人間の被験者が作成した抄録と同様のレベルを達成できている。

2.2.2 表層に関する発見的な規則による手法

文書の修辞構造や表層的な手がかりに着目して、発見的な規則によって、適切な文を選ぶ手法も幾つか試みられている。キーワードの頻度、表層パターン（例 日本語の文頭、文末など特定の表現）、位置、長さなどの特徴を抽出し、これらに基づく重要度と既に選ばれた文との関連性など考慮した発見的な規則によって、順次選んでいく。重要度の計算式のパラメータを重回帰分析によって自動的に学習する方法[5]も提案されているが、特徴抽出部分の、文書の種類、ドメイン、言語に対する依存性は高い。

3 統計的な自動抄録法を使った検索

本実験では、手法が単純で一般的であり、抄録の品質面でも良い結果が報告され、VSMとデータを共有できるZechnerの文抄録アルゴリズムを採用した。

3.1 抄録処理

（1）単語の重要度を決定する。（2）文中の単語の重要度の和[2]を文の重要度とする。（3）文書から一

[1] テストデータの新聞記事に関する表1から、新聞記事の多様性の一端が分かる。

[2] Zechnerの原論文では正規化したら性能が悪かったとあるが、本実験では、この和を単語数で割って正規化した方が良かった。

定個の重要な文を抜粋し[3]これを抄録とする。（4）この抄録を対象として検索する。

単語の重要度は、次式で与えられる基本的な tf * idf を用いる [6]。

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

ここで、w_i: 語 i の重要度、f_i: 語 i の頻度、n_i: 語 i の出現する文書数、N: 全文書数とする。

3.2 検索処理

抄録の前処理と出来た抄録を対象とした検索処理について述べる。

3.2.1 前処理

従来、日本語を対象とした VSM の研究は余りなかった。単語が処理単位となるが、高速・高精度で頑健な形態素解析の実現・入手が困難だったためと考えられる。しかし、最近、高精度の形態素解析システムが幾つか報告され、一般に利用可能となって来ている。本実験では、ALTJAWS を利用している。

英語の検索システムでは、通常 stemming による単語のマージ処理と stoplist による不要語除去を行なう [1]。本稿の日本語の VSM では、単語のマージ処理は、「活用語はその連体形[4]を代表の形とし、以降の処理では品詞などの形態素情報を無視する方法」で集約させた。また、不要語除去は「stoplist を用いるのではなく、品詞を参照して助詞や助動詞などの機能語を削除した。」stoplist は一般的に高頻度語を中心に構成されるが、助詞と助動詞は高頻度語であり、ほぼ同等の効果があると考えられる。

3.2.2 検索処理

単語の重要度として、tf 部分の拡張とベクトル長での正規化を行なった Salton 等の推奨する [4] 以下の tf * idf を使った。

$$w_i = \frac{(0.5 + 0.5\frac{f_i}{\max f}) * log\frac{N}{n_i}}{\sqrt{\sum_{j=1}^{j=W}((0.5 + 0.5\frac{f_j}{\max f})^2 * (log\frac{N}{n_j})^2)}}$$

ここで、w_i: 語 i の重要度、f_i: 語 i の頻度、n_i: 語 i の出現する文書数、N: 全文書数、W: 文書中の単語数とする。

[3] Zechner の原論文では「見出し」などを特別扱いしても性能が良くなかったとあるが、本実験では、見出しをその重要度にかかわらず抄録に含めた方が良かった。

[4] サ変は語幹を用いる。

4 実験

日本語の新聞記事を用いて、評価実験を行なった。以下では、使用したデータ、評価方法、実験結果について述べる。

4.1 データ

日本語の情報検索システム評価用のベンチマークとして、BMIR-J1[7] が作成され、供給されている。600 件の文書 (1993 年度 9 月から 12 月までの日経新聞朝刊の経済面の記事 41843 件から選択) と 60 個の自然言語文の質問が含まれている。

表 1: BMIR-J1 の記事（見出しを含む）の諸元

平均段落数	平均文数	平均語数
6.6	16	380
最大語数	最小語数	平均ベクトル長
2058	42	170

質問は、その質問を正しく処理するために必要と想定される機能によって分類されている。本実験では、構文解析や意味解析などの処理が要求されない分類 A の質問に限定した。

文書の正解判定は各質問毎に 2 段階 A,B で与えられている。A はその文書の主題が質問に合致していることを意味し、B はその文書の主題は質問に合致しないが文書の一部が質問に関連することを意味する。

実験では、正解判定 A に対する正解率を中心に評価した。また、後述するように、本論文の手法が正解判定 A の文書を中心に検索し、逆に正解判定 B の文書を検索しない傾向を確認した。

4.2 評価方法

評価は、情報検索の分野で一般的に使われている**再現率** (recall) と**適合率** (precision) を用いた。再現率および適合率は次式で定義される。

$$\text{再現率} = \frac{\text{検索文書中の関連文書数}}{\text{関連文書数}} \quad (1)$$

$$\text{適合率} = \frac{\text{検索文書中の関連文書数}}{\text{検索文書数}} \quad (2)$$

但し、検索文書はランク N 位までの検索文書とし、関連文書の正解判定は、図 1、図 2、図 3、図 4では A、図 5では B である。

再現率・適合率のグラフは次に示す手順で作成した。

[再現率・適合率グラフの作成手順]

(1) 値 N を任意に複数個決める。
(2) 質問ごとに、各 N における再現率と適合率を求める。
(3) N における再現率と適合率の全質問に対する平均を求め、プロットする。

次節の結果では、N を 5, 10, 15 に設定し、次の 5 種類の検索手法を比較する。

- **Z(echner).i**: 見出し及び上位 i 個の重要な文からなる抄録を使った検索。i=1,2,3。
- **L(ead).j**: 見出し及び先頭の j 個の文からなる抄録を使った検索。j=1,2,3。
- **T(itle)**: 見出しのみの検索。
- **W(ord).k**: 上位 k 個の重要な語の集合を使った検索。k=10,20,30,40。
- **F(ull)**: 文書全体を使った検索。

4.3 結果と考察

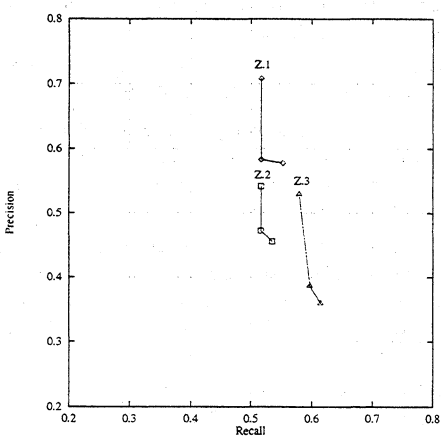

図 1: 再現率・適合率 (Zechner 法)

- 図 1、図 2、図 3から、文単位の手法 **Z**、**L** と単語単位の手法 **W** のそれぞれにおいて、抄録の長さと再現率・適合率の関係をみることが出来る。**単語単位の手法では、抄録を長くすると、適合率が下がるのに伴い再現率が上がる。文単位の手法でも、抄録を長くすると、適合率は下がるが再現率は相対的に高く安定している。** また、**概して、文単位の手法の方が単語単位の手法より適合率が高い。**

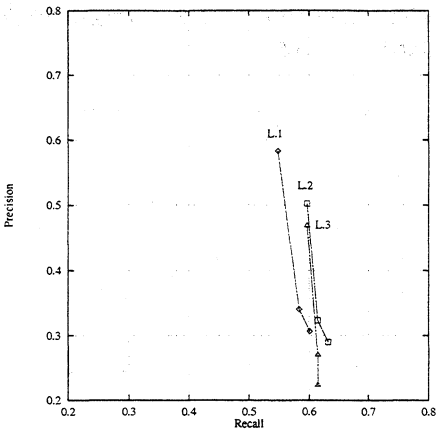

図 2: 再現率・適合率 (Lead 法)

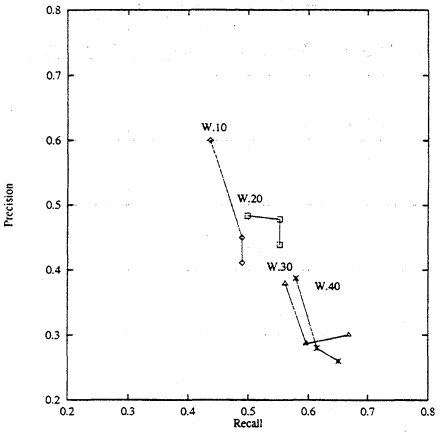

図 3: 再現率・適合率 (Word 法)

- 図 4から **Z**、**L**、**W の順に、適合率が高いことが分かる。**見出しのみの検索 **T** は再現率で、文書中の全ての語を使った検索 **F** は適合率で他の 3 手法に劣る。
- 図 5では**関連するが主題ではない文書が Z、L、W の順に検索されにくいことが分かり、質問が文書の主題か否かに関する感度はこの順に高い**といえる。
- また、**F** は主題、非主題いづれの場合にも高い再現率を示してる。
- 記事の平均文数が 16 で、**Z.2** の文数が高々 3 であるから、圧縮率は約 1/5 である。この率はそのままメモリ効率や速度に反映する。**同じ程度の圧縮**

率[5]の W.30 と比較すると Z.2 の適合率での優位性は明らかである。

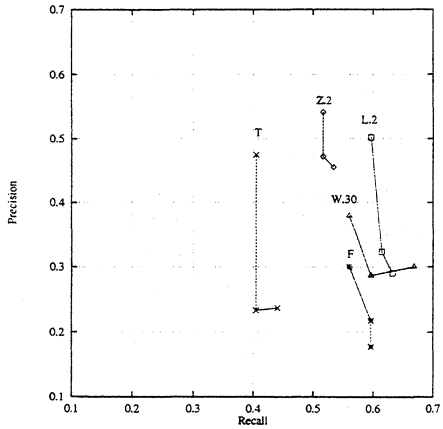

図 4: 主題（正解判定 A）の場合 (5 手法の比較)

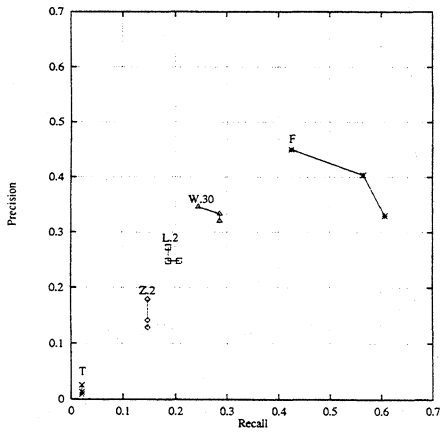

図 5: 非主題（正解判定 B）の場合 (5 手法の比較)

5 おわりに

文書から自動的に抜粋した抄録を検索対象とするという枠組を用いれば、情報検索の精度と効率を同時に改善できると考えられる。実験で統計的な文抄録アルゴリズムの一つを用いて、ベクトル空間型情報検索システムの精度を改善することを確認した。

本稿で採用した方法は文抄録なので、修飾語句などの削除 [9] もされることなく要約としては冗長になりがちであり、精度・効率ともに改善の余地があることになる。単語から組み上げていく方法または文から削っていく方法などを検討する必要がある。

[5]平均ベクトル長は約 10 である。

VSM の様々な改良が報告されているので、これらとの比較も重要な課題である。

謝辞 本研究では以下のデータ・プログラムを利用している。ここに記して関係各位に感謝する。

- 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

- NTT の形態素解析プログラム ALTJAWS をモニター利用。

参考文献

[1] William B. Frakes. **Introduction to Information Storage and Retrieval Systems**. In William B. Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, chapter 1, pp. 13–27. Prentice Hall, 1992.

[2] Luhn H. P. **Automatic Creation of Literature Abstracts**. *IBM Journal of Reaserach and Development*, Vol. 2, No. 2, 1958.

[3] Gerard Salton. **Introduction to Modern Information Retrieval**. McGraw-Hill, 1983.

[4] Gerard Salton and Christopher Buckley. **Term-Weighting Approaches in Automatic Text Retrieval**. *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, May 1988.

[5] Watanabe. **A Method for Abstracting Newspaper Articles by Using Surface Cues**. In *Proc. of the 16th International Conference on Computational Linguistics (COLING'96)*, Vol. 2, pp. 974–978, August 1996.

[6] Zechner. **Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences**. In *Proc. of the 16th International Conference on Computational Linguistics (COLING'96)*, Vol. 2, pp. 986–989, August 1996.

[7] 小川泰嗣, 他. **日本語情報検索のためのベンチマークの構築**. 情報処理学会データベースシステム研究会, 100-16, pp. 145–152, October 1994.

[8] 住田一男, 三池誠司. **知的情報検索の動向**. 人工知能学会誌, Vol. 11, No. 1, pp. 10–16, January 1996.

[9] 山本和英, 増山繁, 内藤昭三. **文章内構造を複合的に利用した論説文要約システム GREEN**. 自然言語処理, Vol. 2, No. 1, pp. 39–56, January 1995.