3V－8

検索キーワードを利用したTF・IDF法による要約文抽出法

蓮井洋志†　魚住超†　小野功一†
†室蘭工業大学情報工学科

1　はじめに

大規模な文書データベースから欲しい情報を得るために、その中の文書をすべて読むのはほとんど不可能に近い。そこで必要な文書を得る方法として、文書の検索や分類、要約などの研究が盛んである。この中で、文書の検索はキーワードによって欲しい文書を特定する。まず、システムがキーワードを索引として抽出しておく。その後、ユーザが文書の特徴を表したキーワードで問い合わせると、索引にそのキーワードが存在する文書を取り出す。

文書から索引を自動的に抽出する方法の一つにTF・IDF法がある。TF・IDF法とは、キーワードがデータベース内の文書に出現する頻度から、そのキーワードの重要度を計算する方法のことである。この方法では、重要度が大きいキーワードを索引に登録する。

また、文書データベースシステムは、検索した結果とともに要約文を表示する。ユーザはそれを検索結果の中から欲しい文書を取り出す時の判断の材料とする。

本稿では、我々は検索要求で指定したキーワードを視点とした要約文抽出法を提案する。文の重要度の計算にTF・IDF式を利用し、重要度の大きい方から好きな数だけ文を抽出する。検索キーワードを含んだ文の重要度をTF・IDF値より大きくすることで、検索キーワードに関する情報を抽出する。この方法を使えば、索引抽出に必要であるTF・IDF値を要約文の抽出に流用できる。

本稿では、提案したシステムを実現し、議論を行った。2章でTF・IDF法について説明し、3章で実現したシステムの構成について記述する。4章でこのシステムについて考察する。

2　TF・IDF式を利用した要約手法

2.1　TF・IDF式と検索キーワード

TF・IDF式とは、キーワードが文書データベース内の文書に出現する回数から、文書中のキーワードの重要度を定める計算式のことである。

表1: CF値

格	CF値	格	CF値
と	0.18	を	0.15
に	0.16	から	0.13
が	0.19	まで	0.09
は	0.21	も	0.10
で	0.12	では	0.16
には	0.22	とは	0.26
からは	0.18	までは	0.09
その他	0.00		

TF・IDF式は、キーワードが文書中に出現する回数TFにIDFをかけた式である。この式の値がキーワードの重要度$Word$を表す。IDF値はデータベース内のすべての文書数をキーワードの出現する文書数で割って、それの対数をとった値である。文書データベース中の多くの文書に現れるキーワードはIDF値が小さくなり、対象とする文書だけにしか出現しないキーワードはIDF値が大きくなる。その文書にしか出現しないキーワードを何度も使うとTF・IDF値が大きくなる。以下にTF・IDF式を表す。

$$TF = \begin{cases} Max\ Term\ Frequency \\ \quad (For\ query\ keyword) \\ Term\ Frequency \\ \quad (Otherwise) \end{cases}$$

$$IDF = \begin{cases} log(All\ Documents\ Number) \\ \quad (For\ query\ keyword) \\ log(\frac{All\ Documents\ Number}{Document\ Frequency}) \\ \quad (Otherwise) \end{cases}$$

$$Word = TF \times IDF$$

要約文は、文書データベースで検索した結果、得た文書がユーザの興味とあっているかどうかを判断するために利用する。ユーザは検索キーワードに興味があるので指定した。つまり、ユーザはその語に関連した内容が知りたいことが推測できる。検索キーワードを含んだ情報を抽出するために、検索キーワードの重要度は、TF値としてその文書内の最大出現頻度、IDF値はデータベース内の総文書数の対数として計算した。この結果、検索キーワードの重要度は他のキーワードの重要度より大きくなる。

Document Summarization by the TF · IDF method with Query Keywords
Hiroshi Hasui, Takashi Uozumi, Koichi Ono at Department of Computer Science and Systems Engineering in Muroran Institute of Technology

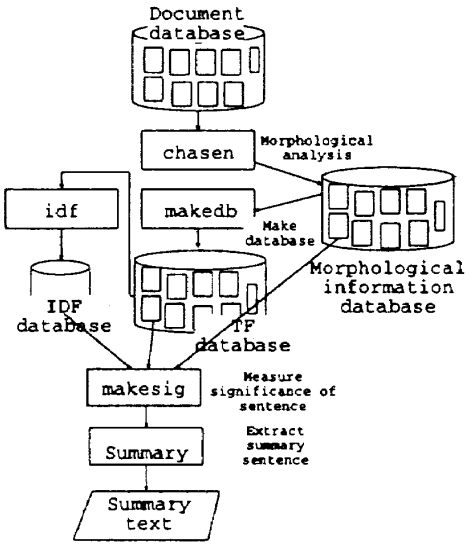

図 1: 要約文抽出システム ExtraSummary の構成

2.2 文の重要度と格情報

本研究における文の重要度は以下の式で表す。文の重要度は、その文を構成するキーワードの重要度 $Word_i$ とその単語が持つ格の偏重度 $Case_i$ を加え合わせたものである。

$$
\begin{aligned}
Case_i &= Word_i \times CF \times PN \\
PN &= Phrases\ Number \\
CF &= \frac{Significant\ Case\ Frequency}{All\ Case\ Frequency} \\
Sentence &= \sum_i (Word_i + Case_i)
\end{aligned}
$$

「は」格、「が」格の単語は話題を表す場合が多い。それに対して、「に」格、「で」格の単語は話題とは関係が薄い。「は」格や「が」格を構成する単語の重要度が大きければ、その文は文書の中でより重要な役割を果たすことが想像できる。そこで、格の種類によって定められる CF と格を構成するキーワードの重要度 $Word_i$、文を構成する格の数 PN の積を、その格を持つ文の重要度に加算する。これを格の偏重度 $Case_i$ という。

CF 値は、文書データベース内の全部の文書において、重要なキーワードをどの格で良く使ったかを表している。文書中の格の出現頻度に対する、重要なキーワードを主辞に持つ格の数の割合で表す。これは格の種類別に決定する。本研究で計測した CF 値を表 1に示す。

3 要約文抽出システムの構成

本研究で提案された手法を実現した要約文抽出システムの構成を図 1に示す。chasen[1] を用いてすべての文書を単語に分割し、その結果を形態素解析情報ファイルに保存する。それを makedb が単語の出現頻度の情報を別のファイルに保存する。登録対象の語は名詞、サ変名詞、固有名詞および未登録語であり、それ以外の品詞の単語は重要度を持たない。これらのファイルは文書ごとに存在する。idf が出現頻度ファイルの中からそれぞれの単語が出現した文書の数を特定し、idf ファイルに出力する。makesig が形態素解析情報ファイル、単語出現頻度情報ファイル、idf ファイルの中の情報を利用して文の重要度を計測し、最後に summary がその中から重要度の大きい文を要約文として抽出する。

ExtraSummary は、朝日新聞の「天声人語」と「社説」それぞれ 1000 件からなる文書データベースシステムにおいて、検索結果を表示する時に利用する。このデータベースシステムおよび ExtraSummary は、Linux 上で Perl および Shell Script で記述した。この文書データベースシステムは Internet 上でブラウザを介して利用できる。

4 考察

抽出型のシステムは、文書から要約文を取り出す。そのためには、主題や結論を明確に表す文が文書の中に存在する必要がある。しかし、小説は主題を物語に託して述べるために、主題を直接的に表現した文は存在しない場合が大半である。本研究の要約文抽出法は小説などの芸術一般の文書には向かない。

科学技術文書は書き方が定められている。短い論文でははじめに全体の要点をまとめる。長い論文でははじめに問題提起をして、最後に要点をまとめる。つまり、文書内に要約された文があり、かつその場所がわかっている。こういった文書は文の位置から判断して要約文を取り出す方策が適している。しかし、社説や天声人語などの一般の論説文では、普通内容を統括した場所はない。本研究の要約文抽出法は、要点を統括した部分のない文書を対象にしている。

文書データベースシステムにおいてユーザに要約文を提示することは、再検索する時に検索キーワードの発想を支援する効果がある。検索結果の要約文から欲しい情報を絞り込む時に追加したい概念のキーワードが要約文から選び出せる。

参考文献

[1] 松本 裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明: "日本語形態素解析システム 茶筅 version 1.5 使用説明書", Freeware, 1997.