シグネチャ法を用いた日本語文書検索システム

4F－1

○(学) 玉置 志津 、(正) 藤原健史、(正) 西谷紘一
奈良先端科学技術大学院大学 情報科学研究科

1. はじめに

文書検索の代表的手法の一つキーワード検索では、あらかじめ各文書にキーワードを与えておく必要がある。キーワードの選択は学術論文など主題分野が限定された文書の場合には、専門用語が中心になるため比較的容易であるが、エッセイや新聞記事など我々の日常生活で目にするような文書の場合は難しい。一方、キーワードを用いない検索としてはフリーワード検索(指定した単語が文書中に出現するかどうかで検索を行う)があるが、指定した単語ごとにオンラインですべての文書をスキャンしなければならないため、検索速度の点で問題がある。

キーワードの選択が不要でしかもフリーワード検索より高速な検索方法としてシグネチャ法の利用を考えた。本研究では新聞紙上に掲載されたコラム記事を対象にシグネチャ法を用いた検索システムを作成・評価した。

2. シグネチャ法の日本語文書への応用

シグネチャとは0と1から成るビット列のことで、各文書データを一定の方式で処理してシグネチャを作成する。文書ごとのシグネチャをデータベース化したものを対象として、検索要求のシグネチャとのパターンマッチングにより適合文書を決定する方法をシグネチャ法という [1]。シグネチャ法は元来英語文書の検索に利用されてきた [2,3]。例えば、アルファベット各文字が特別な意味を持たないことからハッシュ関数を用いて単語シグネチャを計算し、単語の集合である英語文書の文書シグネチャを求めてデータベース作成し、検索を行った例が報告されている [4]。

Japanese Document Retrieval System Using the Signature Method.
Shizu Tamaoki, Takeshi Fujiwara, Hirokazu Nishitani.
Nara Institute of Science and Technology.
8916-5 Takayama, Ikoma, Nara 630-01, Japan.

シグネチャ法を日本語文書へ適用するとき、まず英語文書と日本語文書との構造の相違を把握しなければならない。すなわち、日本語文書はベタ書きの文章であるから、形態素解析を行い、単語の集合に変換する必要がある。また単語を構成する文字のうち漢字は表意文字であることから、日本語文書を構成している最小単位は単語ではなくて文字 (漢字) であるとも考えられ、ハッシュ関数などの利用は適さない。本研究では第一水準漢字の JIS コードを利用して漢字シグネチャを決定し、それをもとに単語シグネチャ、文書シグネチャを計算する方法を採用した。

3. 日本語文書検索システム

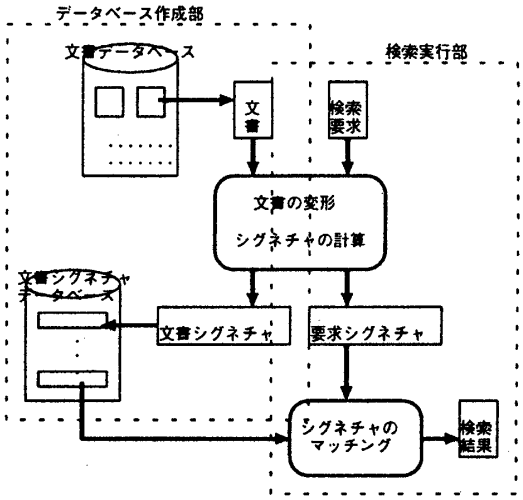

図１　日本語文書検索システム

本システムでの処理の流れを図1に示した。データベース作成部はあらかじめ実行しておき、検索要求がなされたとき検索実行部がオンライン処理される。主なデータ処理について説明する。

(1) 文書の変形

形態素解析を行い、名詞、動詞、形容詞のみを選択する。これらの単語が処理対象単語となる。選択された単語の中の漢字に着目して漢字二文字の単語と漢字一文字単語とに変換する。これらを各々

二文字単語、一文字単語と呼ぶ。このとき、ひらがなやカタカナ、アルファベットは無視される。

例 ・学校長 → 学校＋校長
情報科学 → 情報＋科学
・通り雨 → 通雨
分かれ道 → 分道
取り締まる → 取締
・ハレー彗星 → 彗星

(2) シグネチャの計算

(a) 単語シグネチャ

一文字単語シグネチャは漢字シグネチャそのままである。二文字単語シグネチャは一文字目と二文字目の漢字シグネチャから図2左下部分のように表す。これらを一列につなげたものを単語シグネチャとする。

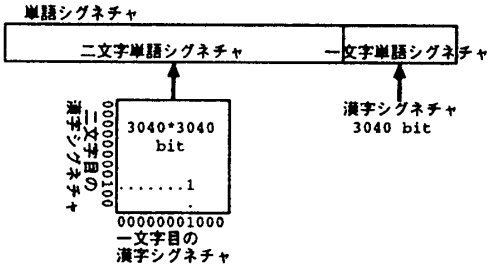

図 2 単語シグネチャ

(b) 単語シグネチャの圧縮

漢字シグネチャのサイズを第一水準漢字数から3040 ビットとしたとき、単語シグネチャのサイズは $3040 \times 3040 + 3040$ (約 10^7) ビットとなる。これは記憶容量の点から大きすぎるため、過去数年間の記事の中に出現しなかった漢字に対応するビットを省略するなどして、単語シグネチャのサイズを縮小した。

(c) 文書シグネチャの計算

文書中のすべての処理対象単語の単語シグネチャに対して、各ビットごとに論理和を取ったものを文書シグネチャとした。

(3) シグネチャのマッチング

二つの文書シグネチャの対応するビット同士を比較したとき、どちらも1であるビット数が多いほど二つの文書の内容が似ていると考えられる。検索時には検索要求に対するシグネチャ(要求シグネチャ) とよく似た文書シグネチャを持つ文書を検索結果とする。

4. システムの評価

検索システムの評価尺度として一般には呼出率と適合率が用いられる [5]。

$$\text{呼出率} = \frac{\text{検索された適合文書数}}{\text{すべての適合文書数}}$$

$$\text{適合率} = \frac{\text{検索された適合文書数}}{\text{検索された文書数}}$$

検索対象とした文書集合はあらかじめ人間の手によって大項目・中項目・小項目に分類されており [6]、検索結果の文書のうちこれと一致したものを適合文書であると定めた。ただし完全に一致した場合だけでなく大項目のみが一致した場合も適合に準じると解釈し、この場合の評価尺度を準呼出率、準適合率と呼ぶことにした。

5. おわりに

日本語文書の検索にシグネチャ法を応用した手法を提案し、それに基づいたシステムを構築して、検索効率などの点からの評価を行った。

6. 参考文献

[1]小川隆一他:フルテキスト・データベースの技術動向, 情報処理,Vol.33,No.4,pp.404-412(1992)

[2]C.Faloutsos and S.Christodoulakis:Description and performance analysis of signature file methods for office filing, ACM Transaction on Office Information Systems,Vol.5,No.3,pp.237-257(1987)

[3]E.J.Schuegraf:Signature searching:a review of theory and application, Canadian Journal of Information Science,Vol.12,No.2,pp.22-35(1987)

[4]川本芳久:シグネチャ法を用いたネットワークニュースの検索, 大阪大学大学院基礎工学研究科物理系専攻情報工学分野修士学位論文 (1993)

[5]伊藤哲郎:情報検索, 昭晃堂,pp.10(1986)

[6]朝日新聞社:朝日新聞−天声人語・社説 1985-1991 増補改訂版 (英訳付), 日外アソシエーツ (1992)