6M－7

文字認識と形態素解析を用いた類似文書検索の試み

間瀬久雄　小山幸子　木山忠博　辻　洋　絹川博之
（株）日立製作所　システム開発研究所　関西システムラボラトリ

１．はじめに

新聞や雑誌、特許、論文、報告書、議事録などの大量文書から必要な文書だけを検索するための技術が要求されている。我々は、印刷文書紙面を入力として、この文書に類似した文書を全自動で検索する方式を検討している。すなわち、印刷文書を文字認識し、認識結果を形態素解析し、解析結果を基にキーワードを抽出し、抽出したキーワードを基に検索式を生成し、検索式に基づいて文書ＤＢを全文検索し、検索結果文書に対して類似度を判定し、類似文書を抽出する。本方式は、特許の公知例調査などにとくに有効である。

本稿では、印刷文書の文字認識精度がキーワード抽出精度に及ぼす影響について考察する。

２．処理の流れ

図１に本方式の処理の流れを示す。

（１）文字認識

ＬＢＰ出力のＡ４横書き印刷文書を入力とする。印刷紙面やスキャナのノイズによる精度劣化を含めた文字認識精度が９５％前後のものを採用した。

（２）形態素解析 [1]

５万語の語彙を持つ辞書を参照して最長一致による単語分割を行い、品詞情報を取得する。

（３）キーワード抽出

品詞が名詞である単語と、辞書に未登録の単語（未知語）を対象として、出現頻度を算出し、頻度の高い単語をキーワードとする。ただし、ほとんどの文書に共通に出現する単語（「場合」「とき」「こと」「もの」等）は除去する。

（４）検索式生成

キーワードを論理演算子（ＡＮＤ／ＯＲ）で結合して検索式を生成する。

（５）全文検索

全文検索システム Bibliotheca/TS [2] により全文検索する。検索式のキーワードを含む文書を漏れなく抽出する。同義語や異表記も吸収できる。

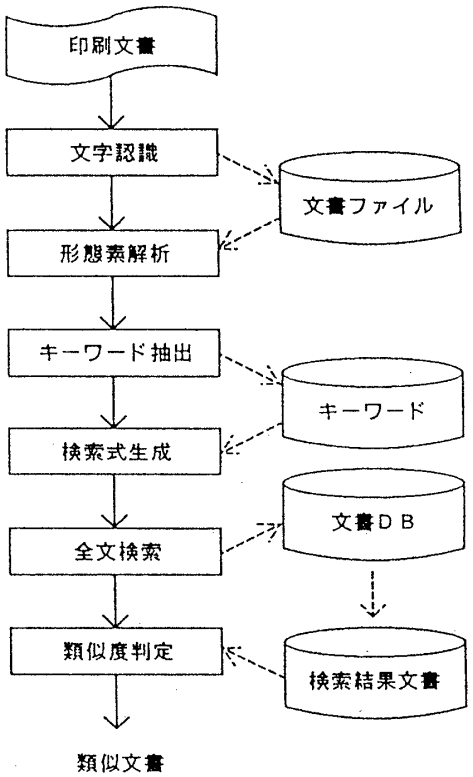

図１　処理の流れ

Text Retrieval using OCR and Morph Analysis
Hisao MASE , Sachiko KOYAMA , Tadahiro KIYAMA , Hiroshi TSUJI , Hiroshi KINUKAWA
Systems Development Laboratory , Hitachi Ltd.

表１　文字認識結果

正　解		４９３８字（９３．６％）
不正解	かな	２３２字（　４．４％）
	記号	５３字（　１．０％）
	漢字	２１字（　０．４％）
	数字	１６字（　０．３％）
	カナ	１４字（　０．３％）
	英字	３字（　０．１％）
	合計	３３９字（　６．４％）

表２　字種別認識結果

字種	正　解　文　字　数
かな	１７２０字／１９５２字（８８％）
記号	３９７字／　４５０字（８８％）
漢字	２４７７字／２４９８字（９９％）
数字	２０２字／　２１８字（９３％）
カナ	１２６字／　１４０字（９０％）
英字	１６字／　１９字（８４％）
合計	４９３８字／５２７７字（９４％）

表２　キーワード抽出結果

評価項目(%)	頻度					
	７以上	６以上	５以上	４以上	３以上	２以上
再現率	9/ 9 (100)	12/12 (100)	27/27 (100)	43/47 (91)	79/80 (99)	164/174(94)
適合率	9/11 (82)	12/14 (86)	27/30 (90)	43/50 (86)	79/88 (90)	164/189(87)
正当率	9/11 (82)	12/14 (86)	27/30 (90)	43/54 (80)	79/89 (89)	164/199(82)

※　再現率(%)=(両方の文書から抽出されたキーワードの種類数)/(認識率100%の文書から抽出されたキーワードの種類数)
※　適合率(%)=(両方の文書から抽出されたキーワードの種類数)/(文字認識された文書から抽出されたキーワードの種類数)
※　正当率(%)=(両方の文書から抽出されたキーワードの種類数)/(どちらか一方の文書から抽出されたキーワードの種類数)

（６）類似度判定

抽出された各文書に、何種類のキーワードが何回出現したかによって類似度を求め、類似度の高い文書から順に表示する。

３．本方式における仮説

本方式では、次の仮説を立てている。

［仮説１］文書に頻繁に現れる語句は、その文書の「キーワード」となりうる。

［仮説２］文書１の「キーワード」を多く含む文書ｎは、文書１の「類似文書」である。

［仮説３］「キーワード」を仮説１のように定義し、「類似文書」を仮説２のように定義すると、文字認識精度がキーワード抽出精度および類似文書検索精度に及ぼす影響は小さい。

４．文字認識精度とキーワード抽出精度の関係

文字認識精度がキーワード抽出に影響を及ぼすかを新聞記事１０文書を対象に実験評価した。

表１、表２に文字認識結果を示す。精度は９３．６％であった。また、表３にキーワード抽出結果を示す。ここでは文字認識精度を１００％とした場合に抽出されるキーワードを正解とした。

再現率については、誤認識結果が上位のキーワード（頻度５以上）には影響がなかった。適合率については、「た」を「左」と誤認識した結果、キーワード「左」がノイズとなった文書が多かった（１０文書中８文書）。適合率の改善については、全文検索方法（検索式の生成方法）の検討などがあり、今後の課題である。

５．おわりに

文字認識精度がキーワード抽出精度に及ぼす影響についての評価実験結果について述べた。文字の誤認識は、キーワード抽出の再現率にはほとんど影響を及ぼさなかった。

今後は、検索式の生成方法について検討するとともに、類似文書の検索精度について実験評価していく予定である。

参考文献

１）西森ほか：汎用日本語形態素解析ツールの開発、情処学会第４４回全国大会

２）浅川ほか：フルテキストサーチシステムBibliotheca/TSの開発、情処学会第４５回全国大会