テキストデータベースのためのキーワード抽出法

3S−1

別所 礼子 広瀬雅子 小川 泰嗣 西村 美苗
(株) リコー 研究開発本部 中央研究所

1 はじめに

オフィスには大量のテキストが氾濫し、その管理が問題となっている。しかし、DBMS は定型データの管理に広く用いられているが、テキスト・イメージなど非定型データのための機能は不十分である。そこで、われわれは非定型データの一つであるテキストを対象とするテキストデータベース管理システムを研究開発中である。その際、索引(キーワード)を用いる方式を採用した。

キーワードに基づくテキスト管理では、登録時の索引作成(キーワード抽出)を自動化することが必須である。しかし、従来のキーワード抽出法は単語の品詞からキーワードか否かを判断していたため、精度が低かった。不要語辞書を導入する方法もあるが、それでも十分な精度が得られず、これら辞書作成・維持も困難である。これに対し、われわれは「キーワード素性」を導入することで、高精度なキーワード抽出を実現した。

2 システムの概要

まず、われわれが開発しているテキストデータベース管理システムの概要を説明する。テキストの管理方法には、登録時に索引(キーワード)を作成しておくものと索引は用いない全文検索法がある。後者は専用ハードが必要なため、われわれは前者(特にキーワードを統制しない方式)を採用した。

テキストの登録・検索処理は図1のようになる。管理用データとして、プリサーチのための文字成分表とランキングのための抽出キーワードがある。文字成分表はテキストごとに各文字が存在するか否かを示す表である[2]。抽出キーワードはテキスト中の重要語を選択したものである。テキスト登録時には、登録テキストに応じて、文字成分表の更新とキーワード抽出が行なわれる。テキスト検索時には、ユーザの入力した検索語に応じて、文字成分表を用いてプリサーチし、得られたテキストに得点付けを行ない、得点の高い順にソートして結果とする[3]。

3 キーワード抽出法

3.1 概要

キーワードとはテキストの重要な単語(列)であり、テキストの内容を的確に表現していることと他のテキストに対する識別性が高いことが求められる。

キーワード抽出はつぎのように行なわれる。

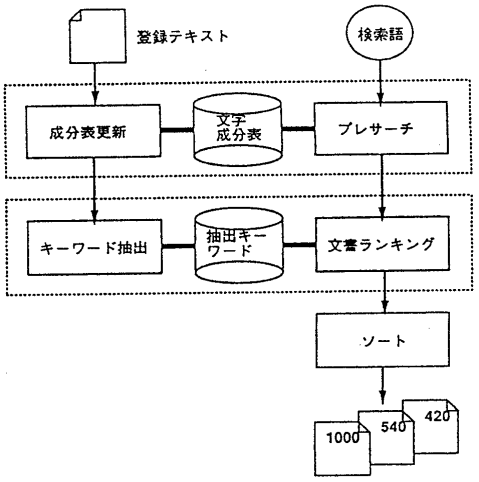

図1: テキストの登録/検索処理概要

1. 形態素解析:
   入力テキストを形態素解析し、単語に分割する[1][4]。
2. キーワード候補抽出:
   キーワードとなり得る単語を品詞により抽出する。
3. キーワード生成:
   キーワード候補の単語列から重要な部分のみを選択し、キーワードと判断する。

3.2 キーワード候補抽出

つぎの品詞の単語をキーワード候補として抽出する。

- 一般名詞・固有名詞
- その他名詞(サ変名詞、和語名詞、修飾名詞、相対名詞、複合名詞、転成名詞)
- 接頭辞・接尾辞
- 助詞(格助詞「の」)
- 数詞・助数詞
- 未登録語

3.3 キーワード選択

抽出ルールによってキーワード候補とされた単語列から最終的にキーワードとする部分を選択する。

3.3.1 キーワード素性

キーワードの判定には品詞だけでは不十分なため、品詞を補うものとして導入した。現在、つぎの6種類ある。

A Keyword Assignment Method for Text Database Systems
Ayako Bessho, Masako Hirose, Yasushi Ogawa and Mina Nishimura (Research & Development Center, RICOH Co., Ltd.)

1. 複合語語基：複合語の末尾となりやすい名詞類
2. 固有名詞構成語:固有名詞の後続語になりやすい名詞類
3. 接頭修飾：後続の語を修飾する働きの強い接頭辞
4. 情報処理分野助数：情報処理分野に固有な助数
5. 地名識別語：地名の中でも識別性の低い固有名詞
6. 元号識別語：元号の一般名詞

3.3.2 選択ルール

キーワード素性を導入しても、単語ごとの判断では正確な判定ができない。そこで、単語の連鎖関係も用いた選択ルールにより精度向上をはかる。主なルールはつぎの通り。

1. キーワード素性なしの一般名詞・固有名詞、未登録語は単独でもキーワードとする。
2. キーワード素性付きの一般名詞・固有名詞、その他名詞は後続語があればキーワードとする。
3. 数詞の連続＋素性なしの助数はキーワードとしない。
4. 数詞の連続＋素性付きの助数はキーワードとする。
5. 数詞の連続はキーワードとする。

4 例

対象テキストを「リコーの中央研究所は超音波センサーを使った形状識別装置を９月に開発した。」とする。抽出処理は表１に示す通りで（○はキーワード候補、◎はキーワードとして選択された単語)、抽出キーワードは「リコー中央研究所」「超音波センサー」「形状識別装置」となる。

表 1: キーワード抽出の例

単語	品詞	キーワード素性	
リコー	固有名詞		◎
の	格助詞		○
中央	名詞相対		◎
研究	名詞サ変	複合語語基	◎
所	接尾辞		◎
は	副助詞		
超	接頭辞		◎
音波	一般名詞		◎
センサー	一般名詞		◎
を	格助詞		
使っ	動詞５		
た	助動詞		
形状	一般名詞		◎
識別	サ変名詞		◎
装置	一般名詞	複合語語基	◎
を	副助詞		
９	数詞		○
月	接尾辞		○
に	格助詞		
開発	名詞サ変	複合語語基	○
し	助動詞		
た	助動詞		

表 2: キーワード抽出精度の評価結果

	部分一致		完全一致	
	再現率	適合率	再現率	適合率
情報処理分野	95.3	58.2	64.3	26.5
一般	94.2	54.3	60.8	26.0
平均	94.5	55.1	61.4	26.1

表 3: キーワード抽出速度

	速度 (chr/sec)	割合 (%)
形態素解析	172.7	81.6
候補抽出＋選択	766.1	18.4
全体	140.9	100

5 評価

5.1 抽出精度の評価

評価対象は新聞記事 200 件。評価基準には、再現率（抽出洩れの少なさ）と適合率（ノイズの少なさ）を用いた。抽出結果の判定は人手で行ない、部分一致（人が判断したキーワードとシステムが抽出したキーワードが一部分でも一致すれば正解と判断する）と完全一致の両方について再現率・適合率を計算した。評価結果を表 2 に示す。

5.2 抽出速度の評価

評価対象は前述の新聞記事 200 件。使用マシンは SUN SPARCstation2（メモリ 32MB、内蔵 SCSI ディスク）である。登録テキストの長さとキーワード抽出時間の関係を調べた結果、比例関係のあることがわかった。この関係を使って抽出速度を計算したものが表 3 である。

6 おわりに

われわれはキーワードに基づくテキストデータベース管理システムを研究開発中であり、そのためのキーワード抽出法を提案した。従来のキーワード抽出法には、精度が低い、不要語辞書の作成・維持が困難などの問題があった。そこで、キーワード素性および単語の連鎖関係に基づく選択ルールの採用により高精度なキーワード抽出を実現し、その有効性を示した。

参考文献

[1] 伊藤篤他. 日本語形態素解析における素性を用いた解析方式. 第 *43* 回情報処理学会全国大会予稿集, 1991.

[2] 岩崎雅二郎他. テキストデータベースのための文字成分表を用いたブリサーチ. 第 *45* 回情報処理学会全国大会予稿集, 1992.

[3] 小川泰嗣他. テキストデータベースのための文書ランキング法. 第 *45* 回情報処理学会全国大会予稿集, 1992.

[4] 望主雅子他. 日本語形態素解析における素性の導入. 第 *43* 回情報処理学会全国大会予稿集, 1991.