類似用例文の効率的検索手法とその応用

5 L－1

溝渕　昭二　泓田　正雄　獅々掘　正幹　青江　順一
徳島大学工学部知能情報工学科

1．はじめに

機械翻訳や文書管理システムにおいて複数の検索要求に該当する用例文をいかに効率的に検索するかは，情報検索の分野で重要な研究課題の１つである[1].

本稿では，用例文の絞り込みを高速化する手法として，文番号ベクトルを用いた手法を提案し，その応用として構築した多属性情報（表記・品詞・概念）を用いた用例検索システムを紹介する．そして，約２１万の用例文に対する実験結果より，従来の手法に比べて１．６～４倍高速化することが分かった.

2．用例文の検索手法

2．1　文番号ベクトルの導入

従来の手法では，索引表から該当する索引の文番号列を検索し，文番号を照合することによって共通の用例文を抽出する．従って，絞り込み速度は索引数と対応する文番号列の長さに比例する[2].

本手法では，共通文番号の絞り込みに文番号ベクトルを用いる．文番号ベクトルは，全用例文数が t のとき，長さが t のビット列で，文番号に対応するビットを１にしたものである．これによって，用例文の絞り込みは，文番号ベクトルの論理積で実行できるが，大規模文書ではこのベクトルが非常に長く，かつスパースになる.

そこで本稿では，文番号ベクトルを多段階に圧縮するデータ構造と，それに対する検索法を提案する.

2．2　多段階に圧縮したデータ構造

まず，文番号ベクトルを図１（ａ）のように長さが一定（dとする）のブロックに分割し，ブロック番号を割り当てる．次に，すべてのビットが０のブロックを削除することにより圧縮する（図１（ｂ））．このベクトルをレベル１とするとき，レベル２のベクトルは，レベル１のブロック番号に対応したビット位置に１をたて，レベル１の場合と同様にブロック分割とブロック削除を行ったビット列である．（図１（ｃ））．この操作を，レベル３以降のベクトルに対しても同様に行えば，$\lceil \log_d(t-1) \rceil$ †レベルでブロックの個数が１になる．このベクトルをルートベクトルと呼ぶ（図１（ｄ））.

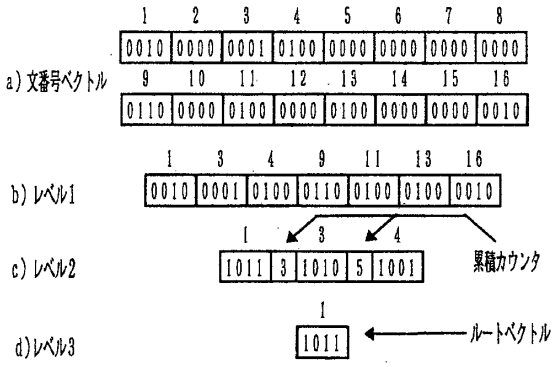

図１　文番号ベクトルの圧縮例（$d=4$）

また，レベル１と各レベルの最初のブロック及びルートベクトルを除くブロックに対しては，累積カウンタを設ける．これは，ブロック内のビットに対応する１レベル上のブロックが，そのレベルの何番目にあたるかを求めるのに使用する．なお，b 番目のブロックで i 番目のビットに対応する１レベル上のブロック番号は，$(b-1)d+i$ となる.

2．3　絞り込みアルゴリズム

k 個の索引が指定されたと仮定するとき，絞り込みは，圧縮されたベクトルを段階的に展開して進められる.

An Efficient Method for Retrieving Example Sentences and its Application
Shoji Mizobuchi, Masao Fuketa, Masami Shishibori, Jun-ichi Aoe
Faculity of Engineering, The University of Tokushima

† $\lceil x \rceil$ は x 以上の最小の整数をあらわす.

まず，k個のルートベクトルに対して論理積を計算し，ブロック番号を求める．次に，k個のベクトルに対して，ブロック番号に対応するブロックを1つ上のレベルから取り出し，その論理積とブロック番号を計算する．これをさらに1つ上のレベルについても同様の処理を行い，レベルが1になるまで繰り返すことによって共通文番号を絞り込む．

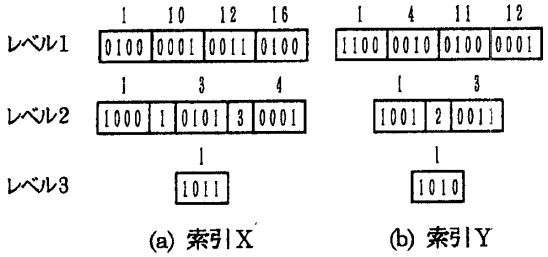

図2　索引X,Yに対する多段階ベクトル

次に，図2の索引XとYの多段階ベクトルに対して，共通文番号を絞り込む．図2は，全用例文数が64でブロックの長さが4の場合のベクトルである．まず，ルートベクトルの論理積の結果は，1010になるので，レベル2のブロック番号1，3のブロックを取り出す．それぞれ論理積を計算すると，1000，0001になるので，レベル1のブロック番号1，12のブロックを取り出す．それぞれ論理積を計算すると，0100，0001になり，最終的に共通文番号は2，48となる．

3．多属性用例検索システム

3．1　システムの概要

本手法の応用として，多属性用例検索システムを構築した．本システムは，表記・品詞・概念の3種類を属性とし，それらに該当する用例文を検索する．

図3に示すようにユーザの検索要求は，インタフェース部に送られる．ここでは，属性に対応するベクトルのIDをインデックスDBから取得し，これを絞り込みエンジンに送る．絞り込みエンジンでは，ベクトルDBからIDに対応するベクトルを取り出し，共通文番号の絞り込みを実行する．そして，絞り込んだ共通文番号に対応する用例文をテキストDBから取り出し画面に表示する．

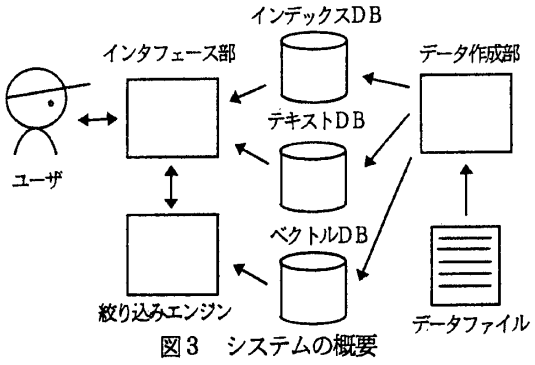

図3　システムの概要

3．2　評価

本手法の有効性を確認するため，EDRの日本語コーパスから抽出した約21万の用例文を用い，種々の索引集合に対する検索時間を測定した．

表1は，指定した索引の文番号列の長さと検索時間の関係を示す．実験結果から従来手法より検索時間が高速化されていることが分かる．

表1　検索時間

索引集合		S1	S2	S3	S4	S5
文番号列の長さ	1～10	1	0	0	0	0
	10～100	0	1	0	3	0
	100～1000	0	0	2	3	0
	1000～1万	0	1	2	1	2
	1万～10万	0	1	0	1	0
	10万～	1	1	2	0	4
本手法の時間		0.05	0.19	0.23	0.06	0.41
従来法の時間		0.23	0.31	0.32	0.15	1.09
倍率		4.6	1.6	1.39	2.5	2.6

4．おわりに

本稿では，文番号ベクトルを用いた効率的な用例検索手法を提案した．さらに本手法を用いた多属性用例検索システムを構築し，その有効性を確認した．

今後は，さらに大規模なデータに対して，本手法を適用する予定である．

参考文献

[1] 加藤他：大規模文書情報システム用テキストサーチマシーンの研究,情処情報学基礎研資,14-6,pp.1-8(1989)

[2] 隅田他：翻訳支援のための類似用例の実用的検索法，信学論，Vol.J74-D-II,No.10,pp.1437-1447(1991)