4E－7

全文検索システムにおける文字成分表の作成手法

藤井　洋一、望月　泰行、鈴木　克志、丸山　冬樹
三菱電機(株)パーソナル情報機器開発研究所

１．はじめに

近年、全文検索システムの開発が活性化している。全文検索の大きな流れとしては、キーワードの自動抽出による検索と、文字成分表を利用した検索対象絞り込み後、実際に検索をおこなう２つが挙げられる。我々はこの２つの方法に関して研究開発をおこなっているが、今回、文字成分表の作成方法に関して大量文書を利用した実験をおこない、評価をおこなった。

２．評価方法の概要

従来の文字成分表の構成は、文字単位の情報を格納したテーブルを作成する場合が主であるが、これでは十分な絞り込みをおこなうことができず、複数文字連続の情報を格納することが、研究されてきた[1],[2]。しかし、単純に２文字３文字と、文字を増やそうとすると、文字成分表の大きさが巨大になって、現実的なシステムとしては、導入が難しいといったことがわかってきている。そこで今回、日本語の文書に限って、文字の出現分布を分析し、大量の文書での絞り込みの可能性について、検討をおこなった。今回利用した文書は、「朝日新聞」記事1年分(朝日新聞社提供）、65447記事(1記事1ファイル)で、トータルサイズ：78612515バイト(１記事平均：1201バイト)である。また、評価方法としては、上記新聞記事から実際に出現する２文字連続の情報を抽出し分析し、4000個のキーワードを検索実験用のキーとして、検索対象の絞り込みの可能性を評価した。

A Signature File Create Method for Full-text Search

Youichi FUJII, Yasuyuki MOCHIZUKI, Katsushi SUZUKI, Fuyuki MARUYAMA
Mitsubishi Electric Corp.

３．２文字連続出現頻度

２文字連続の出現頻度を計算したのが表1-1,2である。表1-1は総出現回数を出現可能なパターンで割ったものであり、表1-2は総出現回数を、実際に出現したパターンで割ったものである。

表1-1 平均出現回数

	句読点	記号	数字	Alph	平仮名	片仮名	特殊	１水準	２水準
句読点	5960.3	118.0	2913.3	72.0	513.4	335.1	0.0	74.9	0.1
記号	72.4	1.3	25.8	1.5	13.6	11.4	0.0	0.6	0.0
数字	1473.0	25.0	2766.6	1.6	19.8	25.1	0.0	11.2	0.0
Alph	20.2	1.3	5.2	33.7	3.2	0.6	0.0	0.1	0.0
平仮名	2541.0	11.5	155.6	3.0	829.0	27.8	0.0	13.5	0.0
片仮名	80.7	13.0	11.4	0.3	31.0	198.5	0.0	0.8	0.0
特殊	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
１水準	27.9	0.5	4.5	0.0	17.9	0.5	0.0	0.9	0.0
２水準	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

表1-2　存在した文字列のみの平均出現回数

	句読点	記号	数字	Alph	平仮名	片仮名	特殊	１水準	２水準
句読点	17880.8	1506.5	5593.5	177.2	1368.9	784.5	1.3	199.1	2.9
記号	1125.8	178.3	261.0	41.4	261.4	156.3	1.8	20.6	1.9
数字	2828.1	361.9	7082.6	7.9	112.9	75.6	1.0	142.6	1.3
Alph	44.9	42.7	22.1	115.1	34.0	5.9	1.6	6.8	1.0
平仮名	6592.8	209.6	439.3	22.8	1701.1	83.1	1.2	66.7	1.9
片仮名	189.1	274.0	36.8	4.9	130.7	370.3	1.6	13.0	1.8
特殊	2.2	1.5	2.0	1.1	1.7	1.0	1.0	3.0	-
１水準	99.0	28.7	35.4	3.2	132.0	7.5	1.0	35.0	3.1
２水準	2.5	2.6	2.2	1.0	1.8	3.5	-	3.1	3.4

表1-1からわかることは、それぞれ平仮名、片仮名同士が連続して出現する可能性は非常に高いということである。また、漢字連続の出現頻度は表1-1からは、ほとんどないということになってしまうが、これは、表1-2からわかるように、実際には現れないパターンが多くあることがわかり、パターンが現れるのは、第１水準でも、1/40程度である。このことは、漢字文字列の検索には２文字連続の情報をある程度落としても有効な検索がおこなえる可能性を示している。

４．文字成分表絞り込み評価

上記の文字分布を参考に、幾つかの文字成分表を作成した場合の絞り込みの効果について、評価した。

文字成分表の作成にあたっては、１文字コードの

文字成分表による絞り込みに加えて、２文字連続のテーブルをどのようにマッピングするかということを考えた。その結果を示したのが表2-1である。データは、正解件数/絞り込み件数x100である。

表2-1 文字コードのタイプによらないマッピング

	フル	12ビット	8ビット	4ビット	3ビット
	91.4	91.0	57.8	32.9	32.2
A	88.0	87.8	73.0	57.1	55.8
B	98.0	97.6	49.2	24.4	23.8

(A: 仮名文字だけのキー、B: 漢字だけのキー)

表2-2 文字コードタイプによるマッピング

	フル	12ビット	8ビット	4ビット	3ビット
	----	91.1	62.8	36.8	35.8
A		87.9	87.9	87.9	87.9
B		97.6	51.0	24.6	23.9

(A: 仮名文字だけのキー、B: 漢字だけのキー)

表2-1は、全体をフル×フル(16ビット×16ビット)で１対１にマップしたものと、下位12ビットずつ、8ビットずつ、４ビットずつ、３ビットずつにしたものである。

また、表2-2はそのうちの数字、アルファベット、平仮名、片仮名についてはフル×フルでマッピングし直した結果である。

この２つの表から、１文字(16ビット)のうち、12ビット程度の情報を持ったテーブルを作成すれば、90%程度まで絞りこむことができることがわかる。このことは、２文字コード連続に対して、フルに情報をマッピングしても、12ビット分でマッピングしてもほぼ変わらないことと、12ビット分マップすれば、絞込みの結果がおおむね検索結果と一致するであろうということが予測できる。また、アルファベット、数字、仮名に関する情報だけを別のテーブルとして生成することで、5%程度絞込みの効果をあげることができることがわかる。

一方、表3-1,2は、今回、実験に使ったキーワードに対して、絞り込み件数の平均の一覧である。

表3-1 表2-1の平均検索（絞込み）件数

	フル	12ビット	8ビット	4ビット	3ビット
	101.4	101.9	160.6	282.0	288.2
A	553.1	554.0	666.2	851.3	871.7
B	60.5	60.7	120.4	243.1	248.4

(A: 仮名文字だけのキー、B: 漢字だけのキー)

表3-2 表2-2の平均検索（絞込み）件数

	フル	12ビット	8ビット	4ビット	3ビット	正解
	----	101.8	147.6	252.2	259.1	92.7
A		553.1	553.1	553.1	553.1	486.5
B		60.7	240.0	248.0	248.0	59.3

(A: 仮名文字だけのキー、B: 漢字だけのキー)

この一覧表から、実際に検索結果として得ようとする結果の数に対して4ビット以下の情報では、検索結果が急激に増えて行くことが分かると共に、特に仮名文字だけからなる検索キーによる絞り込みが困難となり、実際のシステムを構築する場合、仮名文字情報に、重みを持たせた文字成分表を作成することに意味があることが分かる。

5. おわりに

今回、全文検索の絞り込み効果について、２文字連続のマッピングをおこなうことで、ある程度の絞込み効果を実現することが、確認できた。

しかし、今回、新聞記事を利用したということで、幾つかの問題点が考えられる。

1. 新聞記事は、使用する単語の制限もあり、一般の技術文書で、同様の絞込み効果が得られるか。
2. 英単語が頻繁に現れるような技術文書で、英単語の検索要求があった場合同様の絞込み効果が得られるか。

これらのことについて、一般の技術文書を利用して、絞込みの効果について検証することが今後の課題である。

参考文献

[1] 橋本,村井,東谷: 全文検索における高効率プリサーチファイルの一検討、春信全大、1992

[2] 岩崎,小川: テキストデータベースのための文字成分表によるプリサーチ、情処全大(45回)、1992