日本語辞書におけるハッシュ関数の評価

4K-2

和田良一　江村里志　青木豊　本間真人

松下電器産業（株）無線研究所

1．はじめに

日本語処理においては、日本語辞書の検索処理が頻繁に発生する。一般に、大量のデータを検索するためには、ハッシング技法が用いられる。これまでに種々のハッシュ関数が提案され、評価されてきた。しかし、日本語を対象としたハッシュ関数の評価に関する研究例は少ない。そこで我々は、代表的な４つのハッシュ関数について、日本語辞書の見出し語を対象に、シミュレーション実験を行ったので、その結果について報告する。

2．シミュレーション方法

ハッシュ関数をソフトウェアでシミュレートするにあたり、以下の方法を採った。

1)対象とする日本語辞書の見出し語は、約４４０００語からなり，コード体系はＪＩＳコードである。

2)ハッシュ関数として次の４つの関数を選んだ。

①代数符号化法（多項式除算法）

除数の多項式には次の多項式を用いた

16ビット：$x^{16}+x^{12}+x^3+x+1$

32ビット：$x^{32}+x^{26}+x^{23}+x^{22}+x^{16}+x^{12}+x^{11}+x^{10}+x^8+x^7+x^5+x^4+x^2+x+1$

②乗算法

$H(k)=[M*\{k*\alpha\}]$

k:キー　　H(k):ハッシュ値

M:表サイズ　$\alpha=(\sqrt{5}-1)/2$

[ ]:整数部　{ }:小数部

③平方採中法

キーを平方しその中央を採る

④除算法

キーを表サイズで割り、その剰余を採る

3)表サイズを2^{32}＝４Ｇおよび2^{16}＝６４Ｋとする。ただし、除算法では表サイズが２のべき乗の場合、キーの下位ビットを採るだけになってしまうので、2^{16}以下で最も大きい素数である65521 とその２乗を表サイズにした。

4)ハッシュ関数の評価に次のような分散係数を用いた。

$$Bh(k,M)=\sum_{i=0}^{M-1}(n_i*(n_i-1)/2)/k$$

k:ファイルの大きさ　M:表サイズ

n_i:番地ｉへ変換されるキーの個数

ハッシュ関数を実用するには、Bh(k,M)が表占有率(k/M)の２分の１以下であることが望ましいとされている。[1]

5)辞書の見出し語は自由長であるため、乗算法、平方採中法では、統合法により固定長に変換する必要がある。統合法の適用の時期については次の３通り考えられる。

つまり見出し語を固定長の切片に区切り、

(A) 各切片の排他的論理和を採ってから、ハッシュ関数を適用する。

(B) 各切片に対してハッシュ関数を適用してから、各々の排他的論理和を採る。

(C) 前回までのハッシュ途中結果と、次の切片の排他的論理和を採ったものに対してハッシュ関数を適用する。

3．シミュレーション結果

表に４つのハッシュ関数について分散係数を示す。

望ましいとされる、表占有率の２分の１は、

16ビット　$(44109/2^{16})/2 = 3.365*10^{-1}$

32ビット　$(44109/2^{32})/2 = 5.135*10^{-6}$

である。

また図１には各ハッシュ関数の分散係数と表占有率の２分の１との関係を、図２には各ハッシュ関数の衝突数をそれぞれ示す。

Evaluation of Hashing Function for Japanese Dictionary

Ryouichi WADA, Satoshi EMURA, Yutaka AOKI, Masato HOMMA

MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD. WIRELESS RESEARCH LAB.

調査した４つのハッシュ関数のうち、最も衝突が少なかったもの、つまり分散係数が小さかったのは代数符号化法であり、除算法がこれに続く。この傾向は表サイズが大きいほど顕著になる。

乗算法と平方採中法では統合法の時期が大きく影響を及ぼす。

乗算法では例えば、キーを32ビットとした場合、これはＪＩＳコードでは２文字分となる。従って、前述の（Ａ）や（Ｂ）では、例えば「わくわく」「ふわふわ」といった単語に対しては全てハッシュ結果が０となる等、統合法により好ましくない結果になる。従って、（Ｃ）のような工夫が必要である。平方採中法についても同様のことが言える。

４．まとめ

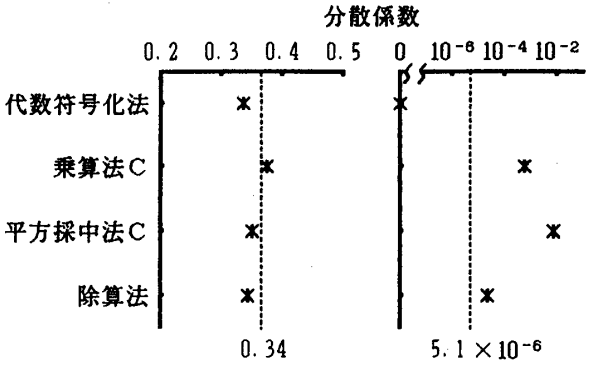
日本語辞書の見出し語を対象としたハッシュ関数の評価を行った。代数符号化法と除算法が優れていて、乗算法や平方採中法では統合法による影響で衝突が少し多い。特に、日本語キーの場合１語の長さが長いため統合法の使用には注意が必要である。今後は、これらのハッシュ関数の実現容易性などを検討して行きたい。

ハッシュ関数		表サイズ 2^{16} (65521)	表サイズ 2^{32} (65521^2)
表占有率／２		０．３６５	５．１３５＊１０$^{-6}$
代数符号化法		０．３３５	０
乗算法	A	１．０７４	７．４４０＊１０$^{-1}$
	B	１．０７４	７．４１０＊１０$^{-1}$
	C	０．３７４	５．８９４＊１０$^{-4}$
平方採中法	A	１．９６８	１．１６２
	B	１．３６８	７．４８４＊１０$^{-1}$
	C	０．３５０	７．４５９＊１０$^{-3}$
除算法		（０．３４３）	（２．２６７＊１０$^{-5}$）

表　４つのハッシュ関数の分散係数

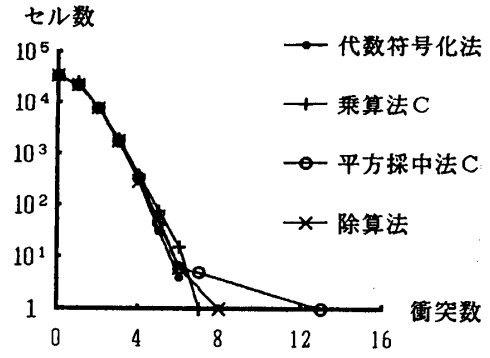
参考文献

[1]西原：ハッシングの技法と応用，情報処理，Vol.21，No.9(1980)．

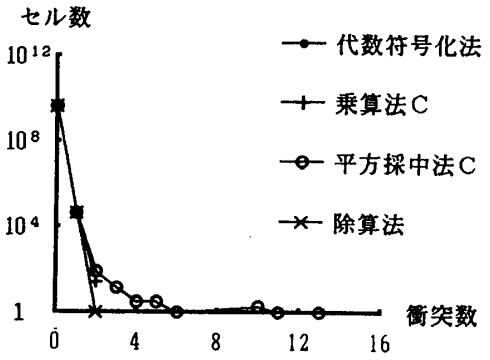


（ａ）　１６ビット　　（ｂ）　３２ビット

図１　分散係数



（ａ）１６ビット



（ｂ）３２ビット

図２　衝突数