n-gram型大規模全文検索方式の開発
5T－3　－文字種適応型n-gramインデクス方式－

川口 久光　菅谷 奈津子　畠山 敦　多田 勝己　加藤 寛次†
(株)日立製作所 情報・通信開発本部‡

1.はじめに

電子化文書情報が急激な勢いで増加するに従い，大量の文書情報の中から所望の文書を迅速に探す検索システムへのニーズが高まってきている[1]。

これに応えるため，登録文書のテキストからn文字の連続する文字列(以下，n-gramと呼ぶ)を抽出し，そのインデクスを参照して全文検索を行うn-gramインデクス方式の検討を行ってきた[2]。

本稿では，n-gramインデクス方式において，総インデクス容量を削減するために開発した，抽出n-gramの種類を抑制する文字種適応型n-gram抽出方式と，インデクス情報としての文書識別子とn-gram出現位置を可変長形式で格納する可変長インデクス方式について報告する。

2．文字種適応型n-gram抽出方式

n-gramインデクス方式では，nを大きく取りn-gramの種類を増加させ，個々のインデクスの容量を小さく抑えることにより，高速な検索が実現できる。しかし，n-gramの種類が増加すると，総インデクス容量が増大するという問題が生じる。

従来の単純に全n-gramを抽出する方式について，3-gramを例に図１を用いて説明する。ここでは，“み取り”や“発した”などのように，検索タームとしてほとんど使用されない付属語を含むものが多く抽出されている。すなわち，自立語の一部と付属語の一部を含む検索タームに使われないn-gramが抽出される結果になっている。検索タームに指定されるのは自立語あるいは複数の自立語から構成される複合語が大半であり，さらに，自立語は同一の文字種で構成されることが多い。したがって，同じ種類の文字で構成されるn-gramだけを抽出することにより，自立語の一部と付属語の一部から構成されるn-gramの抽出を回避することができる。

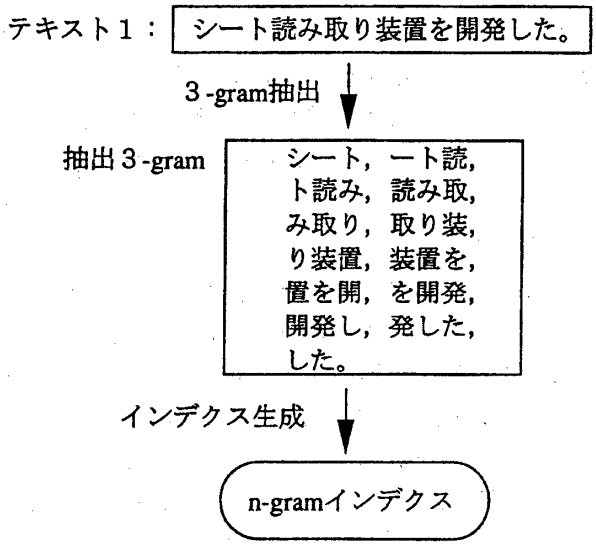

図１　単純方式によるに３-gramの抽出例

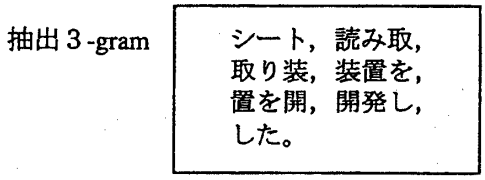

図２　文字種適応方式による３-gramの抽出例

しかし，本図のテキスト１の中の文字列“読み取り装置”のように送り仮名が付く場合は，“読”，“み”，“取”，“り”，“装置”のように文字種の切れ目で小さく分割されてしまうため，検索性能が劣化する。この問題に対処するために，漢字の後のひらがな一文字だけを漢字とみなして文字種の切れ目でn-gramの分割を行い，送り仮名も含めてn-gramの抽出を行う(図２)。

このように文字種に応じてn-gramの抽出を行うことにより，検索時間の劣下を抑えながらインデクス容量を削減し，最適化されたn-gramインデクスを生成することができるようになる。

A Full-Text Search System for Large JapaneseText Bases using n-gram Indexing Method
－ Adaptive n-gram Indexing Method to Character String Type －
† Hisamitsu KAWAGUCHI, Natsuko SUGAYA, Atsushi HATAKEYAMA, Katsumi TADA, Kanji KATO
‡ Information Systems R&D Division, Hitachi,Ltd.

3．可変長インデクス方式

インデクスは文書識別子とn-gram出現位置で構成されており，通常，これらは，４バイトの固定長で表わす。しかし，このままではインデクスが大きくなるため，それらを可変長コードで表わして容量の削減を図る。この方式では，図３に示すように文書番号とn-gram出現位置をともに１バイト，２バイトおよび４バイトの可変長コードで表わし，直前の文書識別子あるいはn-gram出現位置の差分値をその中に格納する。

4．評　価

新聞記事８万件を用いた場合の抽出n-gram数を表１に，総インデクス容量を表２に示す。

（1）文字種適応型n-gram抽出方式を用いることにより，種類数は2-gramおよび3-gramで，それぞれ64%および47%に削減できた。

（2）可変長インデクス方式では，テキスト容量に対するインデクス容量の比は，1-gram，2-gramおよび3-gramで，それぞれ0.84，1.2および1.6となり，４バイトの固定長の場合に対して，60%～80%程度削減できた。

5．まとめ

文字種適応型n-gram抽出方式と可変長インデクス方式を開発することにより，n-gramインデクスにおける総インデクス容量を大幅に削減できる見通しを得た。

参考文献

[1] 菊池，他「日本語文書用高速全文検索の一法」，情処情報学基礎研報，Vol.92，No.32，25-2，pp.9-16，1992.

[2] 菅谷，他「n-gram型大規模全文検索方式の開発：インクリメンタル型n-gramインデクス方式」，第53回情処全大，5T-02，1996.

表１　抽出n-gram数

（）内は無条件抽出方式に対する割合

抽出方式	1 - gram	2 - gram	3 - gram
無条件抽出（全n-gramを抽出）	4,379 (100%)	482,910 (100%)	3,782,630 (100%)
文字種適応抽出	4,379 (100%)	311,314 (64%)	1,760,799 (47%)

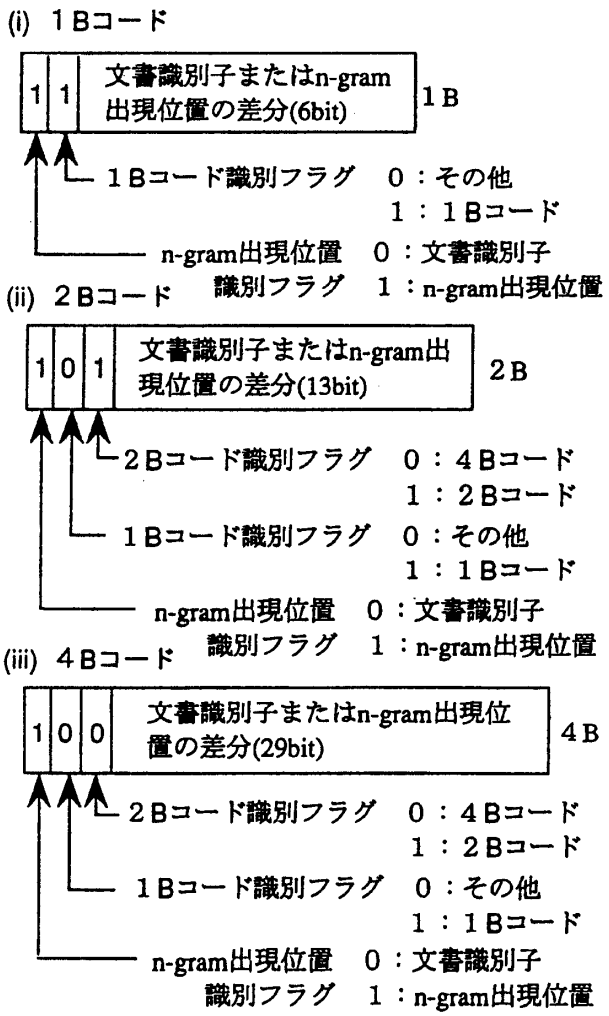

図３　文書識別子とn-gram出現位置の構成

表２ 可変長方式の総インデクス容量

単位：[MB]

コード種		1-gram	2-gram	3-gram
１Bコード	文書識別子	13.3	18.6	10.3
	n-gram出現位置	28.1	24.4	23.2
	合計	41.4	43.0	33.5
２Bコード	文書識別子	1.0	21.2	39.5
	n-gram出現位置	30.1	36.8	38.5
	合計	31.2	58.0	78.0
４Bコード	文書識別子	0.017	3.0	21.6
	n-gram出現位置	0.002	0.1	0.1
	合計	0.022	3.1	21.7
総インデクス容量		72.6	104.0	133.2
テキスト比		0.84	1.2	1.6
固定長形式(４B)との比		0.32	0.36	0.43