n-gram型大規模全文検索方式の開発
5T－2　－ インクリメンタル型n-gramインデクス方式 －

菅谷　奈津子　川口　久光　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. 概　要

パソコン等の普及に伴い，それらによって作成される電子化文書が急増している。また，特許公報も電子化され，全文検索対象のデータベース(DB)は大規模となる一方である。そのため，このような大規模 DB を対象とした場合でも，所望の文書を高速に探し出すことができる全文検索システムへのニーズが高まっている。

今回，n 文字の連続する文字列(n-gram)に対するインデクスを用いて検索を行う n-gram インデクス方式[1]において，検索時間の長大化と総インデクス容量の巨大化という相反する問題点を解決する方法を検討した。その結果，インデクス容量が大きい n-gram のみ，その長さを動的に拡張するインクリメンタル型 n-gram インデクス方式を開発することができた。本稿では，その基本方式と評価結果について報告する。

2. n-gram インデクス方式とその問題点

n-gram インデクス方式は，登録時に文書中の全ての n-gram についてその出現位置をインデクスとして格納しておき，検索時に検索ターム中の n-gram に対しそのインデクスを参照し，検索ターム中の位置関係とインデクス中の位置関係が等しいかどうかを判定(隣接判定と呼ぶ)することによって，検索タームが出現する文書を探し出す方式である。図 1 に 1-gram インデクス方式の例を示す。本方式は，テキストを走査することなしに，インデクスの読込みと隣接判定だけで検索が行えるため，大規模な文書 DB に適用した場合でも高速な全文検索を実現できる可能性がある。

しかし，本方式には以下の問題がある。単一文字(1-gram)による方式では，単一文字の出現頻度が高いため，個々のインデクス容量が大きくなる。このためインデクスの読込みに時間が掛かるばかりでなく，隣接判定処理も増えるため，検索に時間が掛かってしまう。この問題を回避するには，n の値を大きくして個々のインデクス容量を小さくして，インデクスの読込みや隣接判定を高速化する必要がある。しかし，n 文字未満の検索タームでも検索できるように，n 文字未満のインデクスも全て作成しておかなければならないため，総インデクス容量が膨大になってしまう。

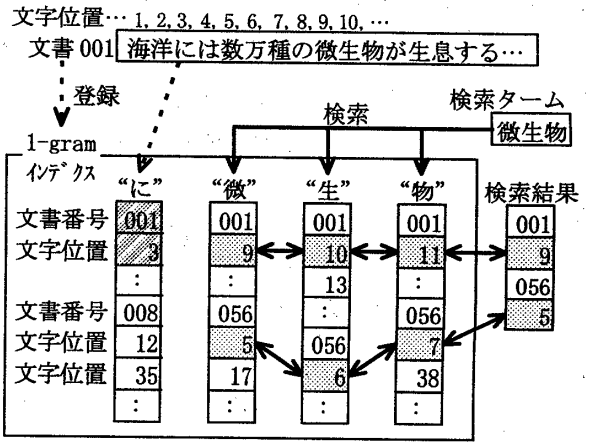

図 1　1-gram インデクス方式

3. インクリメンタル型 n-gram インデクス方式

図 2 にインクリメンタル型 n-gram インデクス方式の概要を示す。出現頻度が高い n-gram はインデクス容量が大きいため，インデクスの読込みや隣接判定に時間が掛かる。これを改善するため，本方式では，インデクス容量がある基準値(基準インデクスサイズと呼ぶ)を超えた n-gram に対してのみ，n の値を増やして容量の小さなインデ

A Full-Text Search System for Large Japanese Text Bases using n-gram Indexing Method
-- Incremental n-gram Indexing Method adapting to Index Size --
†Natsuko SUGAYA, Hisamitsu KAWAGUCHI, Atsushi HATAKEYAMA, Katsumi TADA, Kanji KATO
‡Information Systems R&D Division, Hitachi,Ltd.

クスを作成する。こうすることにより，常に容量の小さなインデクスの読込みと隣接判定で検索が実現できるため高速な検索が可能となるとともに，総インデクス容量の増加も防ぐことができる。基準インデクスサイズには，マシン性能やディスク性能を考慮して算出した，目標検索時間を達成できる最大のインデクス容量を用いる。また，本方式では登録時や検索時にインデクスを参照するためにトライ[2]を用いる。トライとは検索対象のキーワードに共通な前方部分文字列を共通の節で括り出して作られる木構造である。

本方式では，登録時にインデクスを作成するとともに，インデクス容量が基準インデクスサイズを超えたか否かを判定する。ここで，基準インデクスサイズを超えた場合のみ，その n-gram に続く文字を文書中の 2-gram をキーワードとして作成したトライを参照して調べる。そして，検出された文字のインデクスと基準インデクスサイズを超えた n-gram のインデクスで隣接判定を行うことにより，n-gram の長さの拡張を行う。

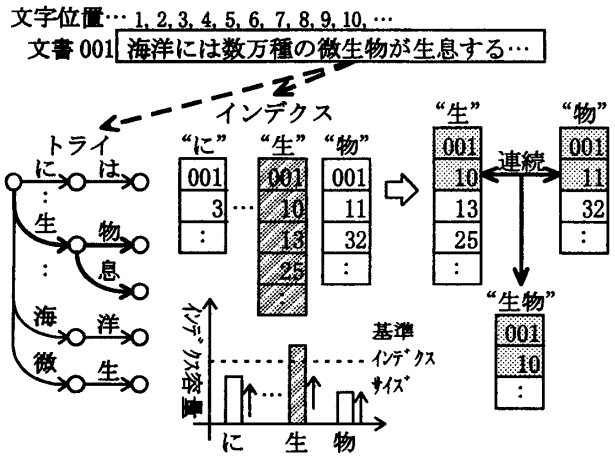

図 2　インクリメンタル型 n-gramインデクス方式

4. 評　価

ワークステーション 3050RX/340(ディスク 2GB)で，新聞記事 100 万件を対象に，目標検索時間を 1 秒として評価を行った。

(1)総インデクス容量

図 3 に示すように，1-gram から 3-gram までのインデクスを持つ 1〜3-gram 型と比べ，総インデクス容量を 26%削減することができた。

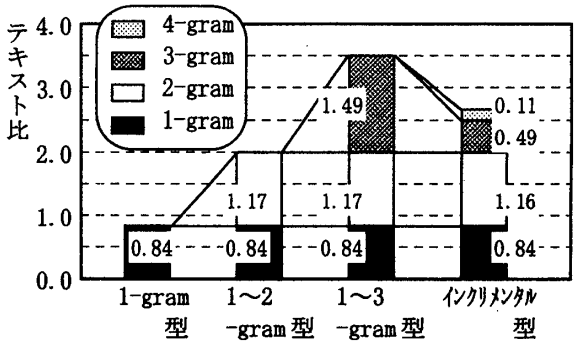

図 3　総インデクス容量

(2)検索時間

図 4 に示すように，1〜2-gram 型で時間が掛かる検索タームに対しても，目標検索時間 1 秒を満たすことができた。

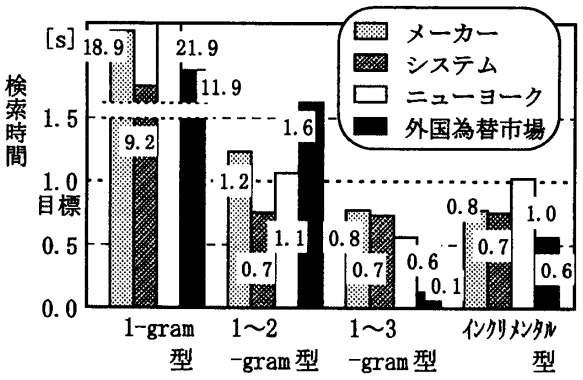

図 4　検索時間

5. まとめ

出現頻度が高い n-gram の長さを動的に拡張するインクリメンタル型 n-gram インデクス方式を開発した。評価の結果，総インデクス容量を 1〜3-gram 型の 74%に削減でき，1〜2-gram 型で時間が掛かる検索タームでも目標検索時間 1 秒を達成することができた。

参考文献

[1] 菊池，「日本語文書用高速全文検索の一手法」，情処情報学基礎研報，Vol.92，No.32，25-2，pp.9-16(1992.5)

[2] 青江，「トライとその応用」，情報処理，Vol.34，No.2，pp.244-251(1993.2)