疑似語句抽出による大規模日本語全文検索方式

森田幸伯, 和田久美子, 池田恵美
沖電気工業 (株) 研究開発本部
メディアネットワーク研究所 μプロジェクト
〒108 東京都港区芝浦 4-10-3
E-mail: {kwada, ikeda, morita}@okilab.oki.co.jp
Tel: 03-3454-2111 内線 2733 Fax: 03-3798-3290

概要

本稿では, 字種に基づく擬似語句抽出を用いた日本語全文検索方式を提案する. また, 従来日本語に対する全文検索方式として広く用いられている N グラム方式との簡単な比較を行う.

キーワード

全文検索, 字種区切り, N グラム

Abstract

In this paper, we propose a indexing method of full-text retrieval for Japanese that extract quasi word using character class information (such as "katakana", "kanji", etc.). And we compare our method to n-gram index method that is widely used for Japanese full-text retrieval.

keyword:

Full text search, Full text retrieval, quasi word index method, n-gram index method

1. はじめに

電子メディアの著しい普及に伴い, 様々な分野で大規模な文書の電子化が急速に進んでいる. 次世代電子図書館システムの構築においては, 多様な要求に答えるべく多くの検索機能の研究が行われている [石本 97]. そのなかで, 大規模電子化文書に対する効率のよいテキスト検索技術は不可欠である.

全文検索技術は, テキストの形態やデータベース構造等に依存せず実用的な検索を実現できる点で非常に有効であると考えられる. 一方で, 検索対象が非常に大規模になった場合, 検索者による絞り込み検索には限界がある. これに対するひとつの解決策として, 単語ごとの統計情報や重要語の推定技術等を用いた高度検索技術との融合が求められている.

日本語文書に対する全文検索の手法としては, 従来 N グラム方式と呼ばれる方式に基づくものが多く用いられてきた. これは, 文章中に現れる文字並びをもとに, 固定長の N 文字に対する出現位置情報をすべて格納するものである.

N グラム方式が日本語文書に対する全文検索システムにおいて多く用いられてきた理由として, 日本語の文章がわかち書きされないという性質が挙げられる (文章中の単語を抽出するためには, 何らかの日本語処理が必要となる). N グラム方式によるシステムにおいては, N の値が増大すると取り扱う文字の組合わせが爆発的に増大するので, 全文検索用索引生成時の計算量や索引格納領域の増大を招く. そのため, 実用システムでは一般に N の値として 4 未満程度を取ることが多い.

本稿では, 文字種や幾らかのヒューリスティックスをもとに文書中の語句を自動的に抽出し, 抽出された疑似単語をもとに索引を生成する全文検索システムについて述べる. 本方式は, N グラム方式において N の大きさが比較的大きく, 可変である場合であると見なすことができる.

本稿では, サンプルデータに基づき, N グラム方式と本方式に対し, 簡単な比較を行なう.

2. N グラム方式

N グラム方式では, 予め対象とするテキスト（以下本文テキストと呼ぶ）に出現する各 N 文字の文字列に対してその位置情報を索引に登録する.

検索語が指定された場合, 検索語をN文字に区切り, 各 N 文字の索引を検索し, N 文字の文字列が出現する位置情報の集合を求め, 各 N 文字の文字列が正しく隣り合ったものかの判定 (隣接判定) を行う.（図 1 ）

N が小さい時は, 少ない文字数に対する位置情報を保持するため, 各 N グラムに対する位置情報が非常に大きくなる. 一般に隣接判定により隣接した位置情報が存在しなければ, 解候補から落とされる. N が小さければ, 隣接判定の回数も多くなり, 解に反映されない位置情報を多く含むことになり, 計算量は増加する.

一方, N が大きい場合には, N 文字の組み合わせが爆発してしまう. このため, 索引に登録される語（この場合 N 文字の文字列）即ちエントリ語の語数が増大し, 索引が膨大になる.

図 2 に特許データをサンプルデータとした場合の, 1 グラムから 4 グラムのエントリ語の量を示す. N が大きくなると, 指数的に増大している. 任意文字列に対して漏れの無い検索を行なうためには, 少なくとも各文字に対する位置情報が必要となる. 少なくとも 1 グラムではそれに近い位置情報を持つことが少なくない. 従って, 位置情報は, 本文サイズのオーダーに近いと考えることができる. 従って, エントリ語が数バイトで表現できると仮定すれば, 索引におけるエントリ語の格納容量は, 位置情報（ほぼ本文サイズ）に比べると 4 グラムの場合でも一桁程度小さい.

いくつかの文献によれば, N 文字未満の文字列の検索のために, N 文字未満の文字列の位置情報も重複して保持する必要があるとされており, その場合, 索引量は本文テキストサイズに匹敵するオーダで増加するので, 大規模全文検索システムとしては実用的ではなくなる. この問題に対しては, 次の節で述べる語句切り

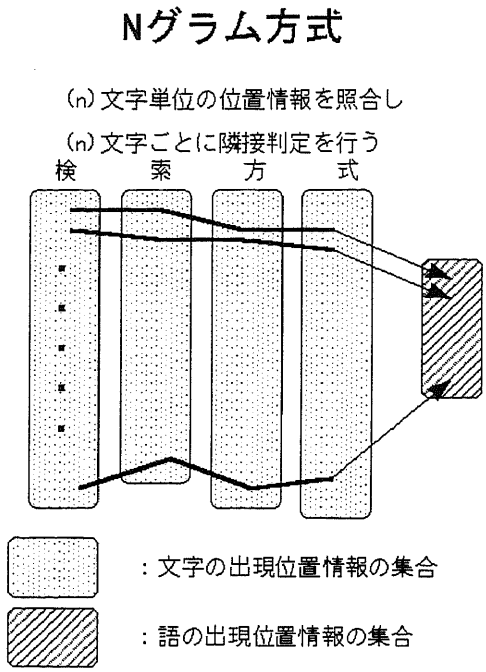

図1

出し方式で行っているように, 索引のエントリ語の前方一致検索を可能にする構成をとることにより回避することができる.

このように, N グラム方式では, N が小さければ, 解以外の位置情報を読み上げる可能性が高くなり, N が大きくなるとエントリ数が増大し索引量が増えるという性質がある.

3. 擬似語句方式

擬似語句方式では, 本文を文字種を基本とし, それに若干のヒューリティックスを加えることにより, 高速に語句抽出を行う. 文字種としては, 「漢字」「平仮名」「片仮名」「数字」「記号」「句読点」などからなる. 実際には, 文字種により単純に切り出すのではなく, 『「句読点」は登録しない』や, 『「漢字」のあとに「平仮名」がきた場合切らない』といったヒューリスティックスを導入する. このヒューリスティックスは, 対象とするテキストの性質や, 本文照合の機会があるかなどのシステム構成により適切なものを定める.

例えば, 「電子メディアの著しい普及に伴い」であれば

電子／メディア／著しい／普及／伴い

と抽出される. さらに, 漏れのない検索を行なうために, 索引に対しては, 抽出された疑似語句（切出し語句）だけでなく, 以下のような展開語句も登録する. 例えば. 「全文検索方式」が切出し語句とすれば, 「文検索方式」「検索方式」「索方式」「方式」「式」を登録する. 索引は, 前方一致で検索するので, 一文字専用の索引を作成する必要はない. これら語句に対して位置情報が登録される.

図3の構成では, 索引の前方一致的な検索を行なうことができる. 即ち, 前方一致検索が必要である場合, 例えば, 「検索*」のような検索を行なう必要がある場合は, 「策」の下のノードの位置情報すべてをマージすれば良い.

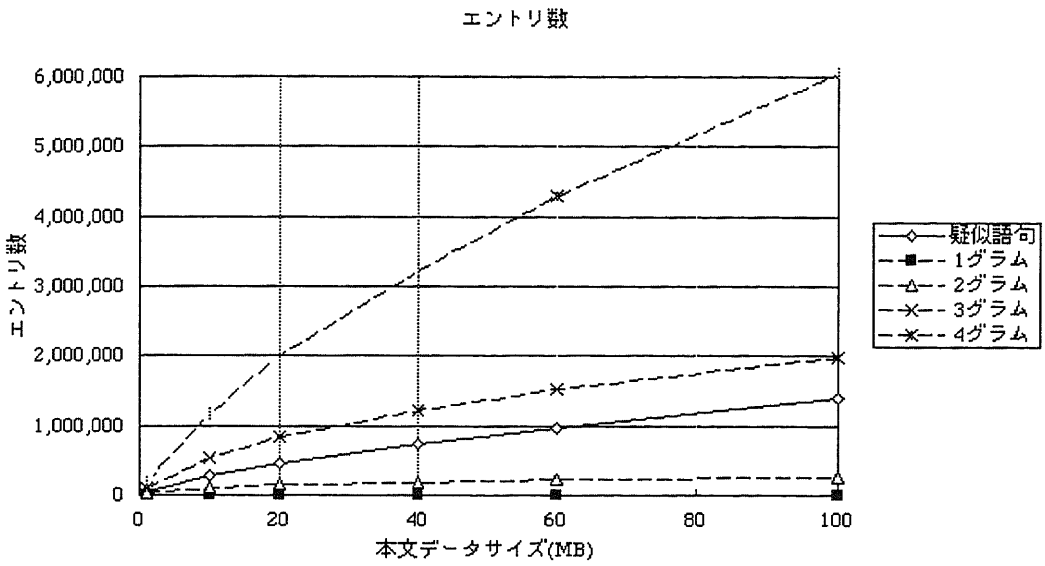

図2

索引の前方一致検索を可能にすることで, 前節で議論となった, 位置情報の重複保持は不要となる. また, 索引の前方一致検索は, 検索語の最後に切り出された語句に対してのみに行なえば良い. 隣接判定とはことなり, マージ処理自体で不必要となるデータは存在しないため, (重複して位置情報を保持する場合に比べ) 読み込むデータ量は, 同じである. 大規模データを想定した場合, データの読み込み時間が支配的要因になる可能性が高いため, マージ処理によるオーバーヘッドはそれほど深刻ではないと予想される.

[菊地 92] によれば, 漢字に関しては 2 文字でかなり強い絞り込みが期待できる. 従って, 我々は先頭 M 文字によるハッシュ表を用いることにより, 単純な木構造の場合より, 高速に索引に対して前方一致検索を行なう構成を採用している.

勿論抽出された疑似語句間の隣接関係は, N グラム同様判定する必要がある. 疑似語句方式は, N が比較的大きめで可変の場合と見做すことができる.

疑似語句方式のエントリ量については, 図 2 に示した. おおよそ, 3 グラムに近い量になっている.

検索速度は, 検索語が幾つに分割されたか (以下分割数と呼ぶ) が大きな要素となっている. 分割数の回数だけ索引を参照し, 位置情報の隣接判定を行うことになる. N グラムの場合は, 文字数だけで決定される. つまり, M 文字の検索語に対しては, M/ N を整数に切り上げた数の N グラムに分割され, 検索される. しかしながら, 擬似語句方式の場合には, 切り出し処理を行うため, 検索語に依存してしまう. 一般に検索者がどのような検索語を用いるのかの予測は難しいため, 分割数の期待値を求めることは困難である.

サンプルデータに対する疑似語句方式の抽出語長の平均は, 7 バイトであり, これは, エントリ語の数が 3 グラムを下回っていたことを考えると若干長い. もし, 検索語が, 本文データと同じ言語的性質をもつとすると, 検索語の分割された語句の平均もこの値に近いものとなる筈である. その意味で大雑把に言えば, 擬似語句方式の方が分割数が若干少ないことが予想される. 特に, 検索語を長い場合に有効である.

図 4 に, エントリあたりの位置情報の平均を示す. これは, 一つのエントリを検索した場合に候補となる位置情報の数の平均を示している. この数が小さい程絞り込みが強く行われていることを示している. 擬似語句方式は, 傾きが小さく, 大量のデータに対しても, 平均位置情報数が比較的小さくなる傾向がある.

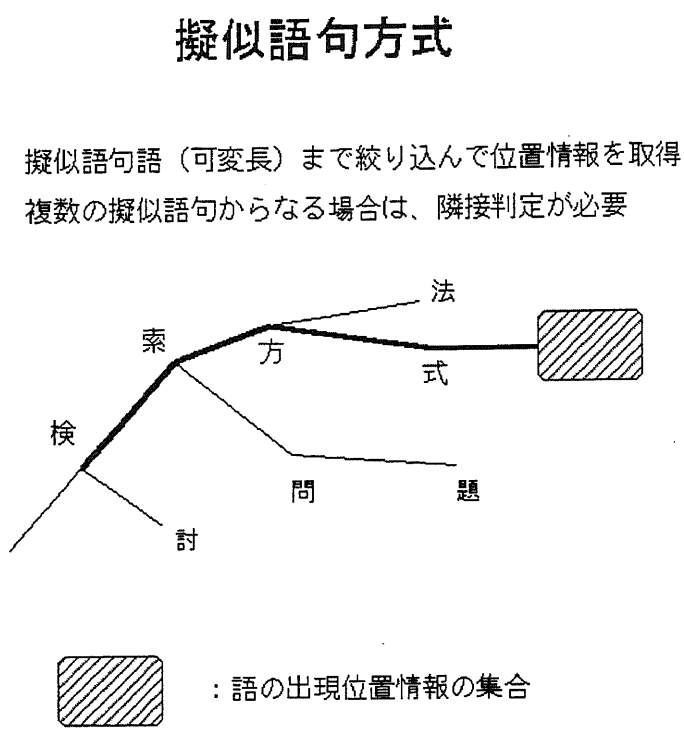

図 3

4. 関連技術

擬似語句方式は, 文字列毎に可変長のエントリ語に対する索引を作成する. 一方, N グラム索引に対して, N を固定ではなく, 可変にすることにより, 効率をあげる研究がなされている.

[赤峰 96] らは, フレキシブル文字例インバージョン法として, 文字種により, N グラムの N を変える方式を提案している. 片仮名などは比較的長く用いられまた文字種も少ないため, N が大きい方が有効である. このように文字種毎に適切な N が採れるようにしている. しかし, 文字種毎には固定された N が採用されている. 擬似語句は, 切り出しに字種に基づくヒューリスティックスを用いるている. 従って, 単純に N の制限を外したフレキシブル文字列インバージョン法とも等しくならない. また, フレキシブル文字列インバージョン法では, 1, 2グラムのエントリ語を同時に保持しているが, 本稿で想定している擬似語句方式では, 索引の前方一致での検索を行うことにより, 切り出し語句と展開語句のみの登録で, すべての文字の 1 グラム情報を保持する必要は無い.

フレキシブル文字列インバージョン法では, 絞り込みを強くするために, 語句の前後の情報を持つ縮退文脈も提案している. この情報を付加することにより, エントリあたりの位置情報の数が減少し, 隣接判定処理時に不要となる位置情報が少なくなり計算量が少なくなる.

[菅谷 96] では, インクリメンタル型 N グラム方式を提案している. これは, 高頻度の語句に対しては, N を増加させたエントリ語を新たに作るという方法で N を文字列毎に可変にしている. エントリ語あたりの位置情報の数に上限を設けることになり, 適切なエントリ語だけを保持するという意味で最適化がなされている. しかし, N を増加させる際, 索引の部分的な再構成が必要となるため, 更新が多い場合には索引生成の処理量も増加し, N をそれほど大きくすることは効率的できないと思われる.

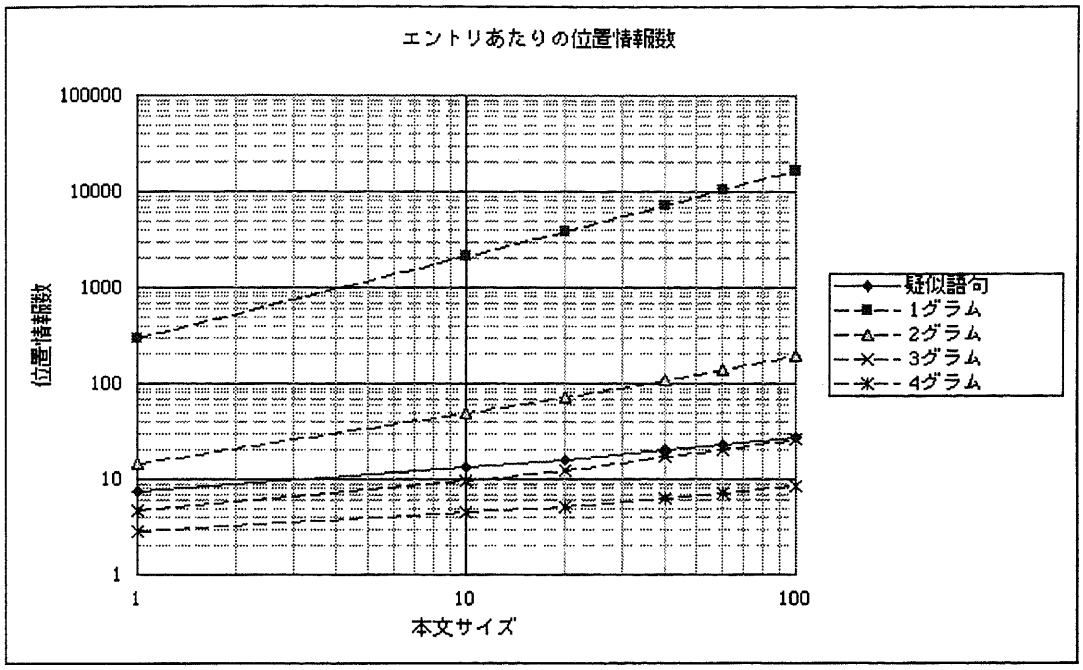

図 4

5. おわりに

全文検索方式の評価を行うことは以下のような依存性があり容易ではない.

(1) 対象としたデータ依存性

(2) 検索語に対する依存性

(3) 提供検索機能に対する依存性

例えば, (1) は, 対象とするデータが用語の誤記や統制をどの程度制御しているか, あるいは, 用語の偏りがどのくらいあるかにより性能が異なることを意味している. また, 利用者が検索語として指定するのが, 名詞だけなのか, 単語にならない文字列を指定することもあるのかなどにより性能が左右される (2). (3) としては, 検索漏れをどの程度重視するかなど設計思想により大きく性能が変わることや, 近傍検索をどのような単位で行えるかなど細かい仕様が方式により異なり, それを導入すると性能が大きく変わる場合などがあげられる.

本稿では, 疑似語句抽出による全文検索方式を提案し, 主にサンプルデータからのデータ量の側面から, N グラムと比較した. かなり荒い比較ではあったが, 特に大規模システムにおいては, データ量が支配的になる場面が多いため, 一つの目安として有効であると考えられる. 今後, より詳細な評価を行なって行きたいと考えている.

また, これまで行われて来た様々な研究成果から, 文字種をもとにした語句抽出によってもある程度日本語の単語に近いものを抽出できることがわかっている. このことから, 本方式は, 索引生成処理および検索処理において実用的計算量で処理可能ならば, 文字並びを扱う N グラム方式に比べて高度検索との結合などの検討を行なって行きたい.

本研究は, 日本情報処理開発協会殿の次世代電子図書館システム研究開発事業の一環として行われており、次世代電子図書館アーキテクチャに準拠して、プロトタイプシステムに組み込む予定である。

参考文献

[赤峰 96] 赤峰, 福島, 「高速全文検索のためのフレキシブル文字列インバージョン法」, アドバンスドデータベースシンポジウム 96, pp.35-42, (1996.12).

[石本 97] 石本, 福島, 「次世代電子図書館プロジェクトの概要」, 第１０回ディジタル図書館ワークショップ, (1997.7)

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

[菅谷 96] 菅谷, 川口, 畠山, 多田, 加藤, 「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 インデックス方式 -」, 情報処理学会第 53 回全国大会, 5T-2, pp.3-235,236, (1996)