疑似語句抽出による大規模日本語全文検索方式

4N－1

池田恵美 和田久美子 菅井猛 奥村晃 森田幸伯
{ikeda,kwada,sugai,okumura.morita}@okilab.oki.co.jp
沖電気工業(株) メディアネットワーク研究所

1　はじめに

近年の電子メディアの著しい普及に伴い，大量の文書が電子化されるようになり，これらを効率よく検索するテキスト検索技術のニーズが高くなってきている．とりわけ，文書全体を検索対象とする全文検索技術は，テキストの形態や構造等に依存しないため，電子図書館システムなど様々なシステムにおいて不可欠の技術となることが考えられる．

我々は，字種やヒューリスティクスをもとに文書中から自動的に語句を抽出し，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を開発した．本発表では，本方式で作成したシステムの評価を報告する．

2　索引方式

2.1　n-gram 方式

日本語を対象とした全文検索技術の手法として，現在最も主流なインデックス方式は，n-gram 方式である．本文から，固定長の n 文字の連続する文字列を抽出し，これに対応する出現位置情報をすべて格納するものである．n の値を大きくするほど，エントリ文字列の出現頻度は小さくなり，隣接判定処理も少なくなるが，n 文字の組み合わせで生成されるエントリ数は巨大になる．これに対し，菅谷 [3] は，出現頻度の大きいエントリに対しては n の値を増やし，エントリの持つ出現位置情報数を平均化した．また，赤峯 [4] は，文字種ごとに n の値を設定し，語句の前後の情報をもつ縮退文脈により，隣接判定処理を軽減させている．

2.2　疑似語句方式

これに対し，疑似語句方式では，字種とヒューリスティクスを用いて語句を切り出す．

例：システムの評価を報告する．

システムの | 評価を | 報告 | する |．

また，このように切り出された疑似語句から，１文字ずつずらして得られる展開語句もエントリとして登録する．これにより，漏れのない任意の文字列検索を可能とする．本方式は可変長の n-gram 方式とみなすことができる．

先述の２方式では，検索の高速化のため n 以下の 1-gram,2-gram,…の索引を持ち，同じ位置情報を重複して持つ．一方，本方式では可変長の索引語を１つの索引で扱うため，位置情報を重複して持つ必要がない．

そして，疑似語句と展開語句の情報を保持する．

3　システム構成

本システムは，C/S 方式で実現され，複数の検索サーバを管理するインデックスマネージャが存在する．マネージャはサーバの実行管理や，索引の生成処理を行う．また，クライアントからの要求に従い，適切なサーバに処理を渡す．サーバの索引のロードに関する詳細は，和田 [1] で報告する．

本システムは論理演算機能（AND,OR, 近傍演算, ワイルドカード検索 etc.）やランキング演算機能を有し，SQL ライクな仕様となっている．

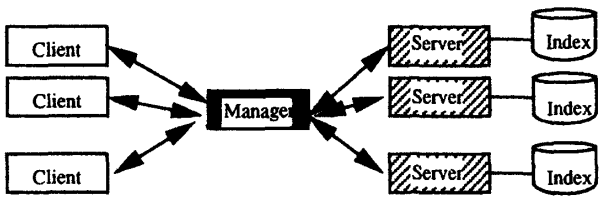

図 1: システム構成図

4　評価

本方式で作成したシステムの評価を示す．

An indexing scheme of Japanese full text search using quasi word extraction method

Emi Ikeda, Kumiko Wada, Takeshi Sugai, Akira Okumura and Yukihiro Morita
Oki Electric Industry Co., Ltd., Media Network Laboratories

HP9000/777(メモリ 128MB) を用い，特許公開広報より 10MB,60MB,100MB,500MB のデータを抽出し，生成した索引で測定した．

4.1 索引

索引には，本文照合を前提としたものと，そうでないものとあり，前者の索引に対し，後者の索引のサイズは大きくなる．一方，後者は本文照合を行わないため，本文データを併せ持つ必要はない．本発表では，後者のタイプについて評価した．よって，索引には出現する全ての語を登録している．

疑似語句方式で，500MB のデータから抽出した語の平均長は，約 3 文字弱であった．抽出語長は，抽出規則に依存するが，本方式では，特殊文字や記号などからなる疑似語句長は短く，漢字，片仮名などから成る疑似語句長は長く切り出される傾向にある（平均 4 文字弱）．

表 1によると，索引において，位置情報数が支配的であるため，本文サイズが大きくなるにつれ，索引サイズは，ほぼ単調減少していくのがわかる．

位置情報部分の圧縮により，さらに小さくすることが可能であると思われる．

表 1: 索引ファイルのサイズ比較

本文サイズ (MB)	10	60	100	500
索引サイズ (倍)	3.13	2.61	2.51	2.52

4.2 検索速度

500MB のデータで作成した索引に対し，理化学用語をサンプル（サンプル数 2,029）として検索時間を測定した（表 2参照）．なお，以下の結果は，索引の参照開始から解の生成処理（文書ＩＤのランキング処理を含む）終了までの時間であり，クライアントとの通信時間は含まない．

表 2: 平均検索速度

クエリ平均文字数	平均切り出し数	平均検索時間 (msec)
3.70	1.24	304.47

クエリには名詞が指定される傾向があり，理化学用語をサンプルとして得られた検索速度の値から，実用に耐え得るものと思われる．

検索速度は，索引の参照回数と位置情報の隣接判定処理の回数が大きく影響する．即ち，クエリの分割数（切り出し数）が多いほど索引の参照回数は増加し，切り出し語長が短いほど，多くのエントリを参照することになる．

図 2に文字数ごとの検索時間と切り出し数の平均値を示す．検索時間は対数表示（単位:msec）である．

一般に，語ベースの索引方式では，エントリが長いほど解の絞り込みは強くなるが，図 2のような結果になったのは，今回使用したサンプルが,「酸化｜ケイ｜素」「Ａ｜／｜Ｄ｜コンバータ」など疑似語句方式で短いエントリに切り出されたものが多かったためと思われる．

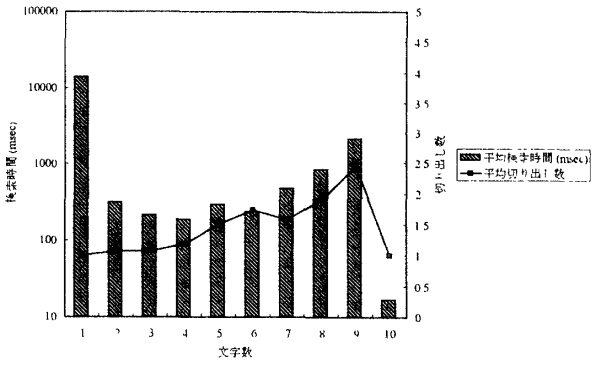

図 2: 文字数ごとの検索時間

5 終わりに

以上，疑似語句方式で実現した全文検索システムについて，報告した．今後は，疑似語句方式の高精度化，ランキング処理の強化を検討し，SQL/MM 対応のシステムを構築する予定である．

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，行っている．

参考文献

[1] 和田他: “索引の動的ロードによる全文検索方式の高速化”, 第 55 回情報処理全国大会,4N-2

[2] 森田他: “疑似語句抽出による大規模日本語全文検索方式”, 第 10 回ディジタル図書館ワークショップ,
URL http://www.DL.ulis.ac.jp/DLworkshop/

[3] 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 インデックス方式-”, 第 5 3 回情処全大会,5T-2,pp.3-235～236,1996

[4] 赤峯他: “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”, ADBS96,pp.35～42,1996