組込みDBMSにおける全文検索用インデクス作成に関する研究

福田 肇† 白井靖人‡
†静岡大学大学院情報学研究科 ‡静岡大学情報学部

1 はじめに

近年、組込みシステムがデータ管理を行うことが多くなっている。例としてカーナビゲーションシステムや携帯電話などが挙げられる。組込みシステムがデータを検索する機会が増え、それにより全文検索機能を提供する組込みシステムが増加すると予想される。そこでDBMSが全文検索機能を提供することによって、開発効率が向上する可能性があると考えられる[1]。

組込みシステムにDBMSを適用するには、従来のDBMSと異なる点として組込み機器自体のリソースの制約やデータ処理のリアルタイム性などが挙げられる。

本研究では全文検索システムにおけるインデクスに注目し、主にリソース制約を満たすような全文検索用インデクス作成方法を検討し評価を行った。

2 全文検索

全文検索システムでは、検索を高速に行う手法としてインデクスを採用しているものが多い。

検索対象となるデータ群を予め走査をして、インデクスを作成する。データに直接アクセスするのではなく、検索キーワードに対応するインデクスにアクセスし、検索対象となるデータを絞り込んで結果を得ることで、より高速な検索を実現する。

全文検索システムでは、直接データにアクセスせず、インデクスを通じてデータを取得するため、インデクスの良し悪しによって検索精度や検索速度に影響が出る。全文検索システムの良し悪しは、インデクスの良し悪しによって決まる。

3 提案

全文検索システムにおいて、検索対象からインデクスタームとなる文字列の抽出を行う手法には、主に形態素解析とN-gramがよく用いられている。以下に各々の手法の特徴を示す。

形態素解析 形態素解析では、辞書を使用して検索対象の解析を行う。的確な検索やインデクスデータが小さくてすむなどの長所がある。一方で解析が辞書の品質に影響し、辞書に登録されていない単語を用いた検索を行うと検索漏れを生じたり、インデクス作成に時間がかかるなどの短所がある。

N-gram N-gramは検索対象データをN文字単位で分解する。辞書が不要であり、検索漏れが少なく、インデクス作成時間が早い長所がある。一方で、インデクスデータの肥大化や意図しない検索結果が生じることが多い。

そこで本研究では、それぞれ長所と短所をもったインデクスターム抽出方法である形態素解析とN-gramを組み合わせて利用することを提案する。両手法を組み合わせて利用することで、それぞれの短所をそれぞれの長所で補う。

インデクスタームを抽出する際には、基本は形態素解析を行い、形態素解析の辞書でうまく分解することの出来ない単語についてN-gramによって分解する。図1に例を示す。

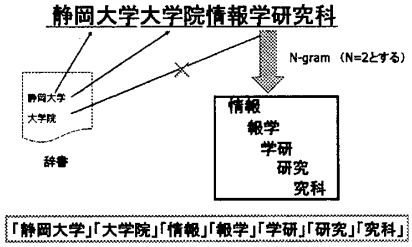

図1: インデクスターム抽出の例

図1の例では、「静岡大学大学院情報学研究科」に対して辞書に「静岡大学」「大学院」のみがヒットした場合を示している。したがって「静岡大学」と「大学院」は形態素解析での分解、残りの「情報学研究科」はN-gramでの分解を行う。その結果、抽出されるインデクスタームは「静岡大学」「大学院」「情報」「報学」「学研」「研究」「究科」の7つとなる(ここではN=2と仮定している)。

そこで本研究では、形態素解析による分解とN-gramによる分解の割合を変化させる方法として、形態素解析に用いる辞書のサイズ(品質)を変化させる。辞書サイズが大きい、すなわち辞書の品質が良い状態では、N-gramによる分解が少なくなる。逆に辞書サイズが小さい、すなわち辞書の品質が悪い状態では、N-gram

---

Full-Text Search Index Making on Embedded DBMS
†Hajime FUKUDA and Yasuto SHIRAI ‡
†Graduate School of Informatics , Shizuoka University , Hamamatsu , 432-8011 , Japan
‡Faculty of Informatics , Shizuoka University , Hamamatsu , 432-8011 , Japan

による分解が多くなる。このように2つの手法の割合を変化させることによって「ちょうどよい」分解の割合を見つけ出すことを目標とする。

4 実験

4.1 評価用システムの実装

本稿で提案した手法を評価するため実験として、今回は提案した手法の動作を確認するためPC上で実験を行った。データとして数値地図25000(地名・公共施設)[2]における公共施設名データを用い、形態素解析器としてSen[3]を利用した。Senで用いる辞書は、公共施設名データをSenで形態素解析し、単語に分割したものを利用した。

実験では単語数を減らす方法として、IPADIC[4]における形態素生起コストに注目し、昇順、降順、ランダムの3通りで辞書の単語数を1000件ずつ減らした。評価対象はインデクシング時間、インデクスのメモリ使用量、検索におけるヒット件数を測定した。また、検索キーワードは「京都」とした。

4.2 結果と評価

以下に実験結果を示す。なお、図2、3、4における「完全な辞書」とは単語を減らす前の辞書のことを指している。

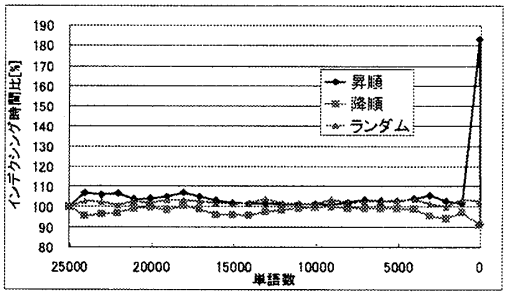

図2: 辞書の単語数に対するインデクシング時間

図2は辞書の単語数を減少させていったときのインデクシング時間を「完全な辞書」のときとの比であらわしたものである。どの方法でも形態素解析が行われている場合は辞書の品質に対してインデクシング時間は大きく関係していないことが分かる。

図3は辞書の単語数を減少させていったときのインデクスのメモリ使用量を「完全な辞書」のときとの比であらわしたものである。単語数を昇順で減少させていった場合、単語数が減少してもメモリの使用量は他の方法と比べてあまり変化しない。

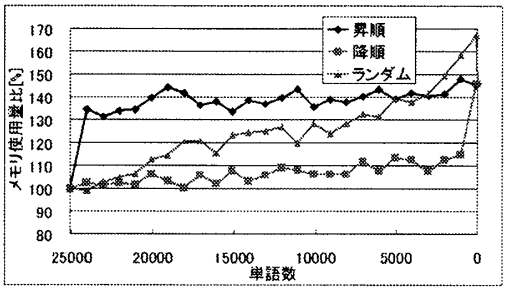

図3: 辞書の単語数に対するメモリ使用量

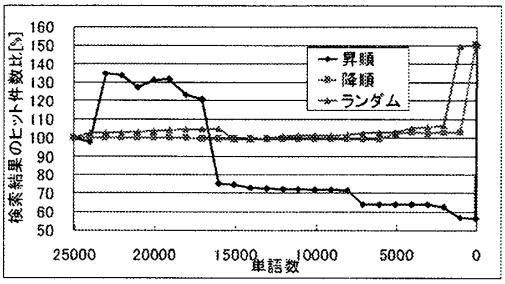

図4: 辞書の単語数に対する検索結果

図4は辞書の単語数を減少させていったときの検索キーワード「京都」の検索結果の件数を「完全な辞書」のときとの比であらわしたものである。単語数を降順で減少させていくと検索結果の件数は、単語数によって大きく左右される。一方で、他の方法では単語数による変化はほぼ無い。

以上の結果より、形態素生起コストの小さなもの、すなわち頻出する単語のみを辞書として用いることでインデクスのサイズを抑えつつ良い検索が行えると考えられる。

5 まとめと今後の課題

形態素解析に用いる辞書の単語を頻出するものだけにして、N-gramと組合わせることでインデクスのサイズを抑えつつ良い検索が出来ることが確認できた。

今後の課題としては、インデクスの登録、削除等の処理に対するリアルタイム性を確保することである。

参考文献

[1] 井上尚樹，田原靖大，福島大雅： 組込みDBMS向け全文検索インデクス更新方法の開発と評価，DEWS2007，E7-10，(2007).

[2] 数値地図25000(地名・公共施設)，国土地理院，http://www.gsi.go.jp/ .

[3] Sen , http://ultimania.org/sen/ .

[4] IPADIC , http://chasen.naist.jp/stable/ipadic/ .