小規模 Web 収集システム向け日本語形態素解析の素朴な分散処理について*

遠山瞬†　幸谷智紀‡
静岡理工科大学§

1. 初めに

我々は膨大な Web 上のテキストデータを自動的に収集し [1)2)]，日本語形態素解析ソフトウェアを用いて名詞のみ取り出して検索する小規模なシステムを作成してきた。しかし日本語形態素解析部分の速度が非常に遅く，名詞テーブル作成に要する時間を短縮する必要があることが昨年度までのベンチマークテストの結果，判明している。そこで，日本語形態素解析部分を PC クラスタ上で素朴に並列分散化し，テーブル作成時間の短縮を目指すことにした。本講演ではその結果について述べる。

2. 小規模サーチエンジンシステムの概要

我々が昨年度まで小規模な Web 検索システムを Fig.1 に示す。

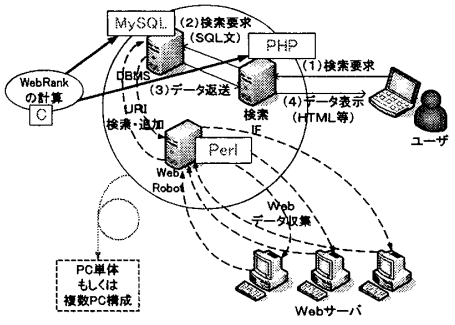

Fig. 1: 小規模サーチエンジンのシステム

Web Robot が自動収集した Web コンテンツデータを Web データベースサーバに蓄え，ユーザの検索要求に対して結果を返すようになっている。我々が昨年まで作成した検索インターフェースではこの部分に日本語形態素解析ソフトウェア (MeCab) を使用し，名詞のみを検索するように改良した。さらに検索の高速化を図るため，名詞ごとにインデックステーブル (名詞テーブル) を作成し (Fig.2)，生のコンテンツデータを直接検索対象とした場合に比べて約半分の検索時間になることを確認している。しかし名詞テーブルを作成する時間が膨大になるため，昨年までの段階ではこの部分の高速化が課題として残っていた。

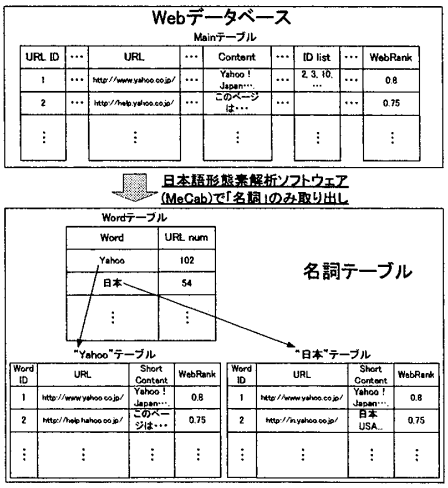

Fig. 2: 日本語形態素解析部分

3. MeCab による日本語形態素解析の並列分散化

以上述べてきたように，我々の小規模サーチエンジンシステムでは名詞テーブル作成作業の短縮化が課題となっていた。そこで，この作業を素朴に並列分散処理化することにした。その方法を Fig.3 に示す。

今回は CentOS 5.3 を導入した Pentium IV 2.8GHz (P4) PC を 11 台用意し，NFS/NIS サーバを介して/home および/usr/local 以下を NFS 共有したクラスタの上で実験を行っている。この上で，日本語形態素解析部分のみを Fig.2 に示すように PC クラスタ上で割り当てら

*On Simply-distributed Processing of Japanese Morphological Analysis for Small-sized Web Crawling System
†Shun TOHYAMA
‡Tomonori KOUYA
§Shizuoka Institute of Science and Technology

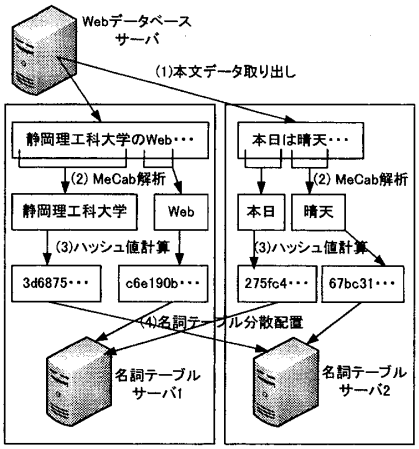

Fig. 3: 日本語形態素解析の並列分散処理

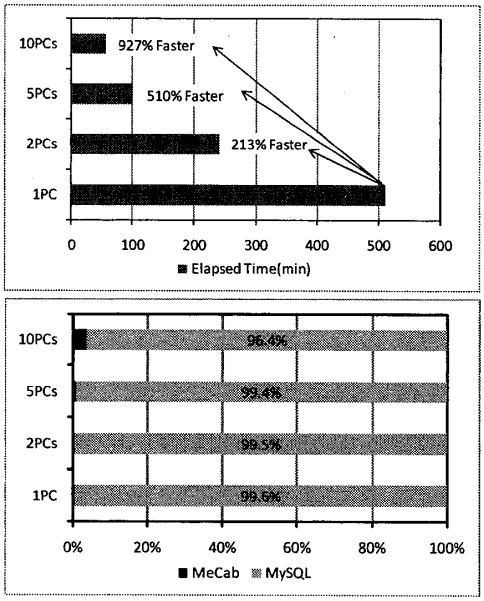

Fig. 4: 並列処理時間 (上) と処理時間に占める MeCab/MySQL の割合 (下)

れる名詞数が平等になるように並列分散処理を行った。名詞テーブル作成作業をなるべく平等に割り振るため，MD5 による名詞のエンコードを行い，MySQL によるデータベースは各マシンのローカルなハードディスク上に作成するようにしてある。そのため，Web 収集データを格納したデータベースサーバマシンへの負荷集中がそれほど無ければ，MeCab プロセス及び MySQL データベース処理に要する時間をリニアに減らすことが期待できる。

4. ベンチマーク結果

以上の構想に基づいて名詞テーブル作成処理の並列分散化を実現した。ここでは実際のところどれほどの効果があったのかを検証してみる。Web データは URL1000 件のページデータを使い，名詞テーブル作成に 1,2,5,10 台のマシンをそれぞれ割り当てて処理した時の処理時間の比較を行った。その結果を Fig.4 に示す。

Fig.4 上のグラフより，台数が増えれば増えるほど処理時間は短くなっており，ほぼリニアに台数効果が現れていることが分かる。なお Web データを格納してあるデータベースサーバマシンへの負荷を考慮し，名詞テーブル作成スクリプトには数十秒程度のウェイトをかけていることを考えると，名詞テーブル作成にかかる時間の大きさがよく分かる。

Fig.4 下のグラフでは，名詞テーブル作成全体における，MeCab の平均処理時間と MySQL の平均処理時間の割合の比較を行っている。当初は MeCab が処理の足を引っ張り速度を低下させていると予想していたが，実際には殆ど影響がないことが分かる。これは偏に MySQL データベースの処理における I/O のボトルネックよるものである。また，並列処理の台数が増えるほどに MeCab の時間の割合が増加しているが，これは Web データベースサーバへの同時アクセスによる負荷の増大によるものと考えられる。しかし 10 台程度の並列分散化では処理時間全体にかかる時間に占める比率は小さい。

5. 結論と今後の課題

以上の結果から，小規模な Web データ収集システム向けとしては，素朴な分散処理によって日本語形態素解析を実用的なレベルまで高速化可能であることが示された。処理時間の殆どはデータベース処理，ことに I/O のボトルネックによるものであることも明らかになった。

今後の課題は，このシステムを利用してより大規模な日本語データ解析を行うことである。そのためにインストールと保守作業がしやすいよう，スクリプトとドキュメントのパッケージングを行い，より広い応用を行っていきたい。

参考文献

1) 幸谷智紀, 小規模な分散 Web ロボットの最適化に関する一考察, 第 71 回情報処理学会全国大会講演集, 2009.
2) 竹口友大・幸谷智紀, ランク機能付きサーチエンジンの開発および I/O ボトルネック対策, 第 70 回情報処理学会全国大会講演集, 2008.