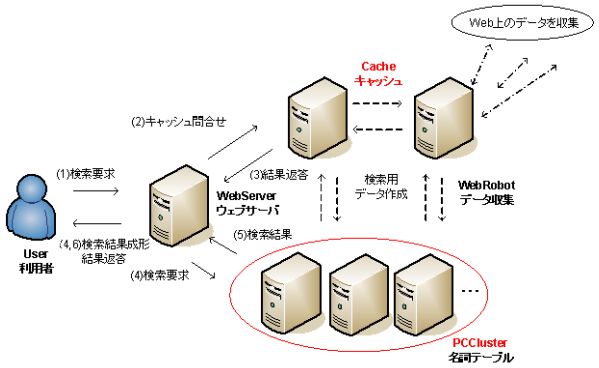
静岡理工科大学 理工学部 情報システム学科 卒業論文
「日本語形態素解析の分散及び小規模サーチエンジンシステムに関する研究」
0613091　遠山 瞬［担当教員:幸谷智紀］
2010 年 2 月 18 日

## 1.目的

本研究では、数年前から研究を続けている小規模サーチエンジンシステムの完成を目指す。昨年までに UI の改善と名詞テーブル(転置インデックス)の実装によって実用が可能なレベルに近づいたが、検索速度面で十分とはいえなかった。そこで今年は、より高速な検索を実現するために複数の計算機を使用し、処理の並列分散処理の実現を目指す。また、検索自体を先に行うキャッシュ機構を実装し、検索速度の更なる向上を図り、実用レベルのサーチエンジンシステムとして運用可能にすることを目的に研究を進める。

## 2.システム概要

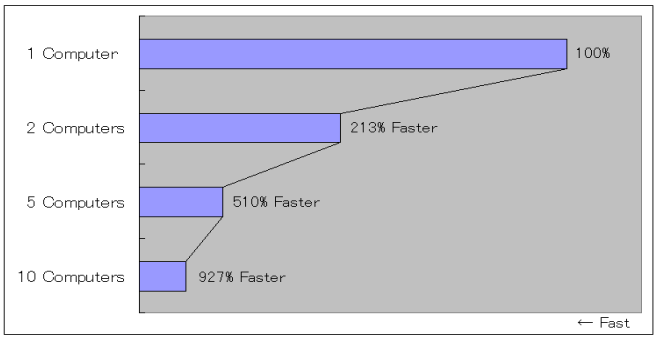
図は本研究で扱う検索システムの全体図である。
赤く示した部分は昨年との変更点を表す。




ユーザの入力を Web サーバが受け取り、キャッシュ、名詞テーブルと順番に検索を実行する。検索に利用するデータは WebRobot が随時収集し、PCCluster が名詞テーブルを作成、検索を可能な状態にする。また、名詞テーブルの作成に並行してキャッシュの作成も随時行う。

## 3.名詞テーブルと並列化

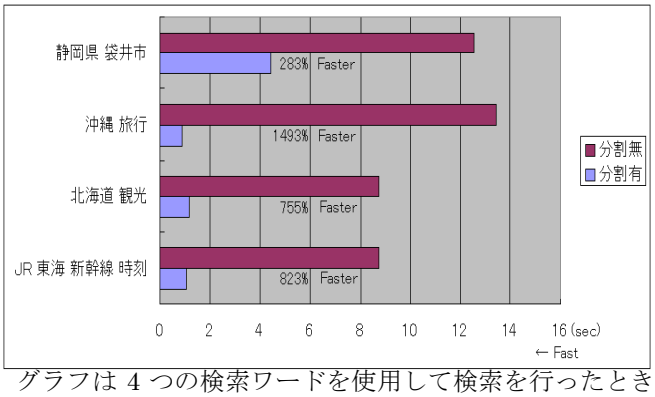
名詞テーブルは一般的に言うところの転置インデックス、逆引き索引のことである。これを使用することで必要なデータを検索するだけで結果を得ることができる。しかし、利用するには予め名詞テーブルを作成しておく必要があり、この作成には相応の時間が必要になる。単語の品詞判別、形態素解析が低速なためである。そこで、処理を複数の計算機で並列実行させることで高速化を図ることにした。下に処理時間の差を示す。




グラフは 1000URL を処理したときの時間の差を表す。台数が増えればその分だけ処理時間は短縮されていることがわかり、並列処理の効果は疑いようがないと言える。

## 4.名詞テーブルの分割

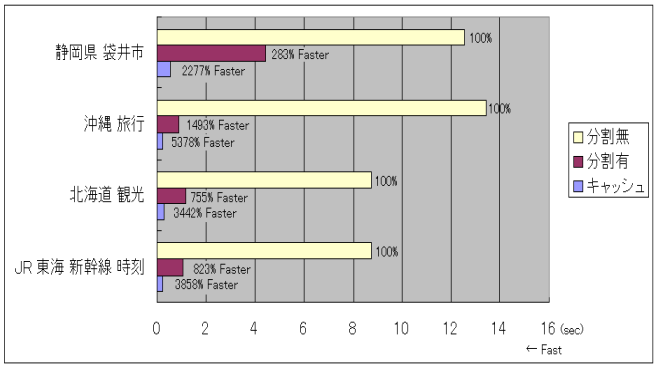
名詞テーブルの並列化に伴い、名詞テーブルそのものも分割し、10 台の計算機に格納するように変更した。昨年度は 1 台の計算機に全ての名詞テーブルを格納していたが、過度な負荷が原因と思われる処理の遅延が発生した。負荷の分散を図り、高速化するために名詞テーブルの分割を行い、検索速度の比較を行った。




グラフは 4 つの検索ワードを使用して検索を行ったときの処理時間の差を示す。全ての検索ワードにおいて名詞テーブルの分割を行った時のほうが検索時間は短いことがわかる。負荷の分散の効果は十分にあったと言える。

## 5.キャッシュの実装

全体的な検索速度は大幅に向上したが、まだ、十分な検索速度を確保したとは言えない。僅か 1000 件程度のデータ検索で数秒間もかかっているようでは、これからデータが増えたときに満足できるレスポンスを得ることは難しい。そこで、検索を事前にサーバ内で実行し、結果のみを返すというキャッシュの実装を行った。キャッシュを使用したときの検索時間を下に示す。




全ての検索ワードにおいて劇的な高速化を図ることができた。キャッシュが存在するという前提があってのものだが効果は非常に大きいと言える。

## 6.感想・考察

昨年と比較して飛躍的に検索速度が向上し、実用に耐えうるレベルになった。データの更新やバックアップ等の機能を追加すれば、さらに良いものになると思う。課題は残されているが、研究は大幅に進歩したと私は考えている。