7N-1

オープンソースの全文検索システムの速度性能比較

早坂 良太　　林 貴宏　　尾内 理紀夫

電気通信大学

1. はじめに

近年，日本語を扱うことができるオープンソースの全文検索システムの開発が盛んになっている．それらのシステムを使用することで，デスクトップ検索などの個人用途から大規模な検索エンジンまで，様々な要求を満たすことが可能となる．いくつかの全文検索システムの中から自分の求めるシステムを選択する場合，速度性能は重要な指針となりうるが，どのシステムが高速であるかは実際に動作させてみなくては明らかではない．

本稿では Namazu(*1)，Lucene(*2)，Senna(*3)，Estraier(*4)，Hyper Estraier(*5)の 5 つのオープンソースの全文検索システムについてインデクシング速度・検索速度を比較した結果について述べる．

2. 調査対象の全文検索システム

2.1 Namazu(*1)

Namazu は日本では個人ユーザを中心に広く使われている全文検索システムである．文字列の分割には KAKASI(*6)，ChaSen(*7)，MeCab(*8)などの形態素解析器による分かち書きを用いる．

2.2 Lucene(*2)

Lucene は Apache プロジェクトの 1 つで，Java で書かれた全文検索システムである．文字列の分割には形態素解析器 Sen(*9)による分かち書きと，N-gram 分割からどちらかを選択して用いる．

2.3 Senna(*3)

Senna は組み込み型の全文検索エンジンであり，MySQL などに組み込んで用いることができる．文字列の分割には MeCab による分かち書きと N-gram 分割のどちらかを選択して用いる．

2.4 Estraier(*4)

Estraier は QDBM というデータベースライブラリをコアとして作成された全文検索システムである．文字列の分割には MeCab による分かち書きを用いる．

2.5 Hyper Estraier(*5)

Hyper Estraier は Estraier の後継の全文検索システムである．コアの部分には Estraier と同じく QDBM が使われている．文字列の分割には N-gram を用いる．

3. 実験

3.1 実験方法

本実験では，5 つのシステムのインデクシング時間・検索時間を測定した．インデクシング時間の測定は，文書数を 1 万から 500 万まで増やしながら行った．ただし，1 万から 10 万までは 1 万ずつ，10 万から 100 万までは 10 万ずつ，100 万から 500 万までは 100 万ずつ追加登録を行った．文書は全て 1 つのディレクトリ内に置き，ファイルパスを確認して，未登録の文書のみをインデックスに登録した．検索時間の測定は文書が追加される度に行った．また，検索クエリは“ブログ”“酢”“実験”の 3 種類を使用した．

実験で使用した計算機は CPU が Intel Xeon 3.0GHz，メモリ 4GB である．インデックスに登録する文書には，著者らの研究室で開発した画像付きブログ検索エンジン もぶろげっと[1]のデータベース内の文書を使用した．各システムのバージョンは以下の通りである．

- Namazu 2.0.13 + MeCab 0.81
- Lucene 1.4.3 + Sen 1.2.1
- Senna rev57 + MeCab 0.81 + MySQL 5.0.15
- Estraier 1.2.28（QDBM 1.8.24）+ MeCab 0.81
- Hyper Estraier 0.9.1

3.2 インデクシング時間の測定結果

実験結果を図 1 に示す．図 1 は登録された文書数と，インデクシング時間との関係を表している．

図 1 より，インデクシングについて 5 つの全文検索システムの中で最も高速なのは Hyper Estraier となった．

Namazu は文書数 30 万のとき，他のシステムに比べて 5 倍以上の時間がかかってしまったため，以後の実験を行わなかった．

Senna は文書数 40 万までしか測定しなかった．これは他のシステムに比べて検索時間が 100 倍以上かかってしまうからである．（3.3 で述べる）

Estraier は Hyper Estraier の次に高速だったが，文書数 500 万の段階で Lucene とほぼ同じインデクシング時間となっている．

Speed performance comparisons of open source full-text search systems

Ryota Hayasaka, Takahiro Hayashi, and Rikio Onai

The University of Electro-Communications

(*1) “Namazu”, http://www.namazu.org/

(*2) “Lucene”, http://lucene.apache.org/

(*3) “Senna”, http://qwik.jp/senna/

(*4) “Estraier”, http://estraier.sourceforge.net/

(*5) “Hyper Estraier”, http://hyperestraier.sourceforge.net/

(*6) “KAKASI”, http://kakasi.namazu.org/

(*7) “ChaSen”, http://chasen.naist.jp/hiki/ChaSen/

(*8) “MeCab”, http://chasen.org/~taku/software/mecab/

(*9) “Sen”, http://ultimania.org/sen/

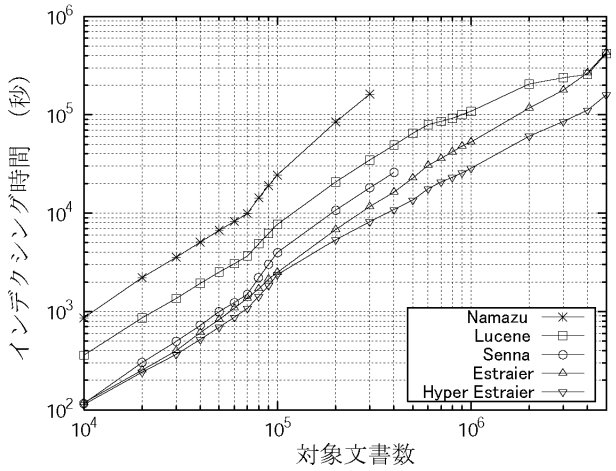

図 1　インデクシング時間の測定結果

図 1 のグラフでは 5 つのシステムすべてにおいて，文書数 10 万前後で傾きが大きくなっている．これは文書数 10 万前後で登録した文書のサイズ（文字数）が平均的な文書のサイズよりも大きいために，インデクシングに時間がかかっているものと考えられる．

3.3 検索時間の測定結果

実験結果を図 2～4 に示す．各図はそれぞれ “ブログ”，“酢”，“実験”をクエリとした場合の検索時間を各システムに対して測定した結果を表している．Namazu は他のシステムに比べてインデクシングに時間がかかってしまうため，検索時間の測定は行わなかった．

図 2～4 より，検索について文書数 500 万までの範囲では 5 つのシステムの中で最も高速なのは Lucene となった．しかし，Estraier は文書数や検索結果数の増加にかかわらずほぼ一定時間で検索できている．よって，本実験で扱った文書数よりもさらに多くの文書数を扱う場合では Estraier が最も高速になると予測できる．

検索クエリによって測定時間が異なるのは検索結果数（検索クエリに適合する文書数）が異なるのが原因である．

Hyper Estraier では 1 文字のクエリに対しての検索時間は，同じような検索結果数の 2 文字以上のクエリに対しての検索時間よりも増加してしまう．これは Hyper Estraier における N-gram 分割の文字数が 2 であるため，1 文字のクエリによる検索は，その文字を含む全ての 2 文字の文字列を検索することになってしまうのが原因であると考えられる．

4. おわりに

本研究では 5 つのオープンソースの全文検索システムの速度性能比較を行った．その結果，インデクシングについては Hyper Estraier が最速となった．また，検索については文書数 500 万までの範囲では Lucene が最速となることを確認した．また，500 万以上の文書数では Estraier が最速になるという結論を得た．

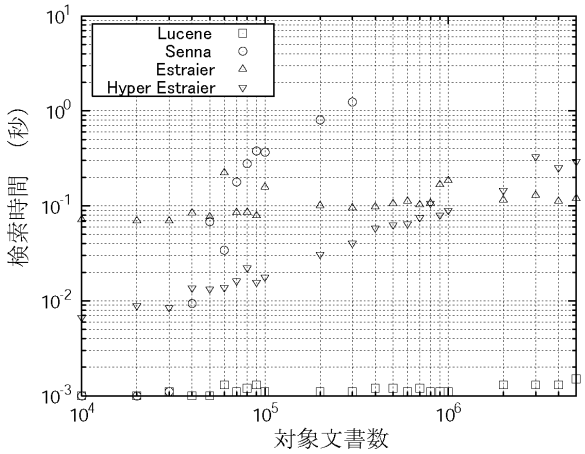

図 2　“ブログ”をクエリとしたときの検索時間

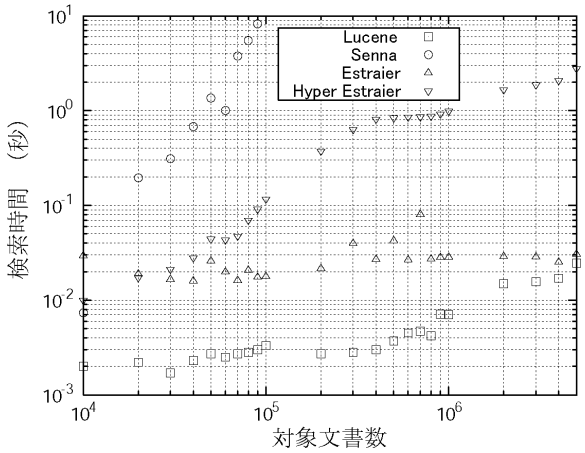

図 3　“酢” をクエリとしたときの検索時間

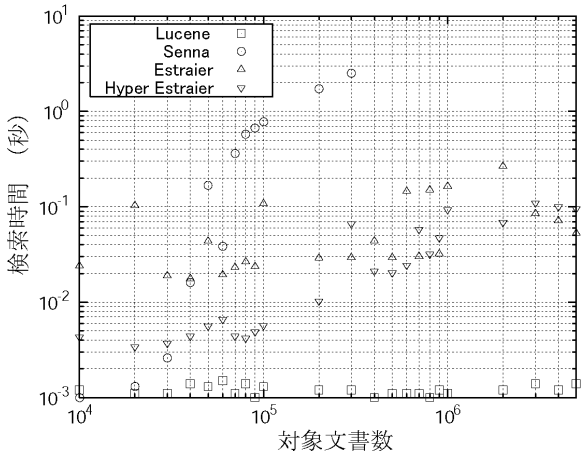

図 4　“実験” をクエリとしたときの検索時間

参考文献

[1] 井原伸介，林貴宏，尾内理紀夫 ： “もぶろげっと: 画像情報を含む blog 記事検索システム”，第 13 回 インタラクティブシステムとソフトウェアに関するワークショップ (WISS2005) 論文集，pp.69-74，2005.12