全文検索エンジン Lux の開発
—高速かつスケールする全文検索エンジン—

1. 背景

近年、人類の創生する情報量は爆発的に増大している。SNS や Lifelog などの Web サービスにおける個々のログや企業内にあふれるデータなど、大量のデータが身の回りに存在している。さらに、マシンやセンサーなどのデータを加えると、その勢いはさらに凄まじく、情報爆発時代などとも呼ばれている。

2. 目的

そのような背景において、必要な情報を「探す」という行為の効率化が今後の重要な課題になってきている。企業という視点から言うと、各企業に「大量のデータを高速に検索する技術・ソフトウェア」が必要である、と言える。また、そのような技術・ソフトウェアはフリーでありオープンである（ソースが公開されている）必要がある。まず、フリーであることにより、資金が少ないベンチャー企業や中小企業においても利用できるということが挙げられる。また、オープンであることにより、多様化した各企業の需要に合わせて適宜カスタマイズできるようになるという利点も挙げられる。
つまり、オープンソースにおいて、拡張可能かつ高速な検索エンジンの開発が求められている。本プロジェクトでは、このような検索エンジンの開発を行うことを目的している。

3. 開発の内容

高速性、拡張性、スケール性を特徴とした検索エンジンの開発を行うことを目標とした。
開発項目としては、以下が挙げられる。

- 検索エンジンの内部データベースの開発
- 高速化(転置インデックスの圧縮、内部アルゴリズムの高速化)
- 基本機能の拡充
- 分散インデックスの開発
- （上記項目における）テスト、ベンチマーク
- （上記項目において）拡張性のある設計

検索エンジンの内部データベースの開発

検索エンジンのインデックスの情報などを格納するためのデータベースの開発を行った。データベースは、key-value 型の DBM と呼ばれるデータベースに似た構造をとり、全文検索における転置インデックスという索引構造を効率的に検索・保存できるように実装した。

高速化(転置インデックスの圧縮、内部アルゴリズムの高速化)

検索におけるハードディスクでの転送量を減らすために、転置インデックスの圧縮

を行う機能を開発した。圧縮アルゴリズムに関しては、広く知られているいくつかのアルゴリズムを実装し、現代のハードウェアの性能にうまく適合し、検索速度が向上するという目的に最適なものを選択した。また、インデックス構築方法において大きな問題となる、位置情報のリストの管理方法は in-place update という手法を独自に改良した戦略を実装し、高速な検索速度を保ちつつ、索引作成速度の向上を達成した。

基本機能の拡充

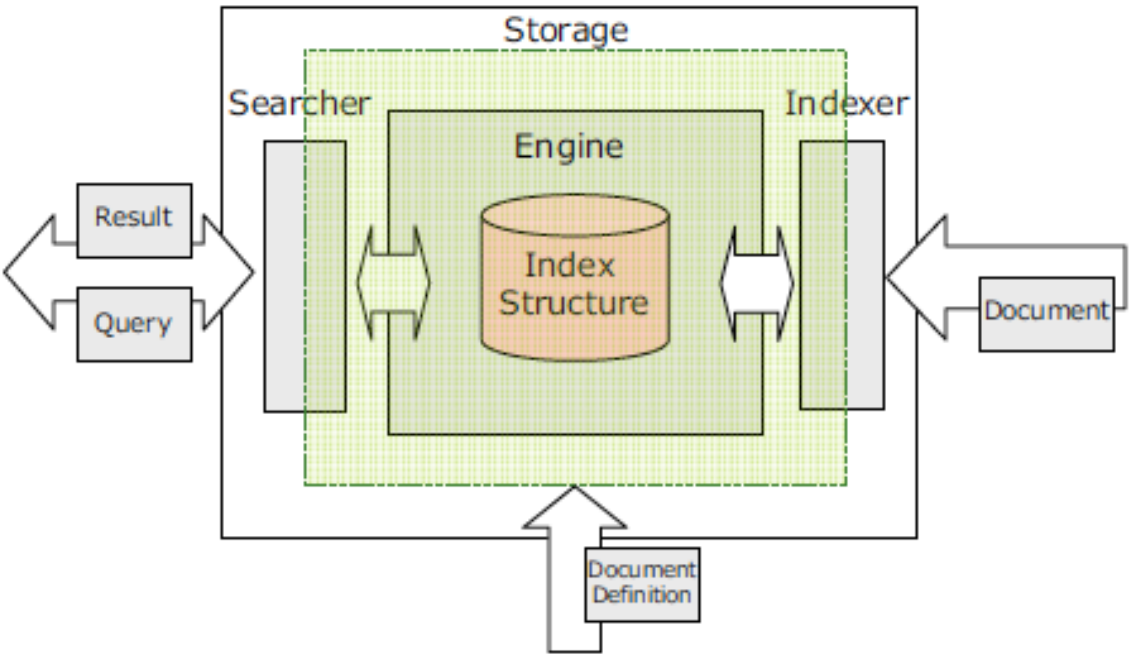
検索エンジンとして今後普及する上で必要と考えられる基本機能の開発を行った。指定した属性による絞り込みを行う属性インデックスと、定義ファイルにより設定を行えるようにする機能の追加を主に行った。

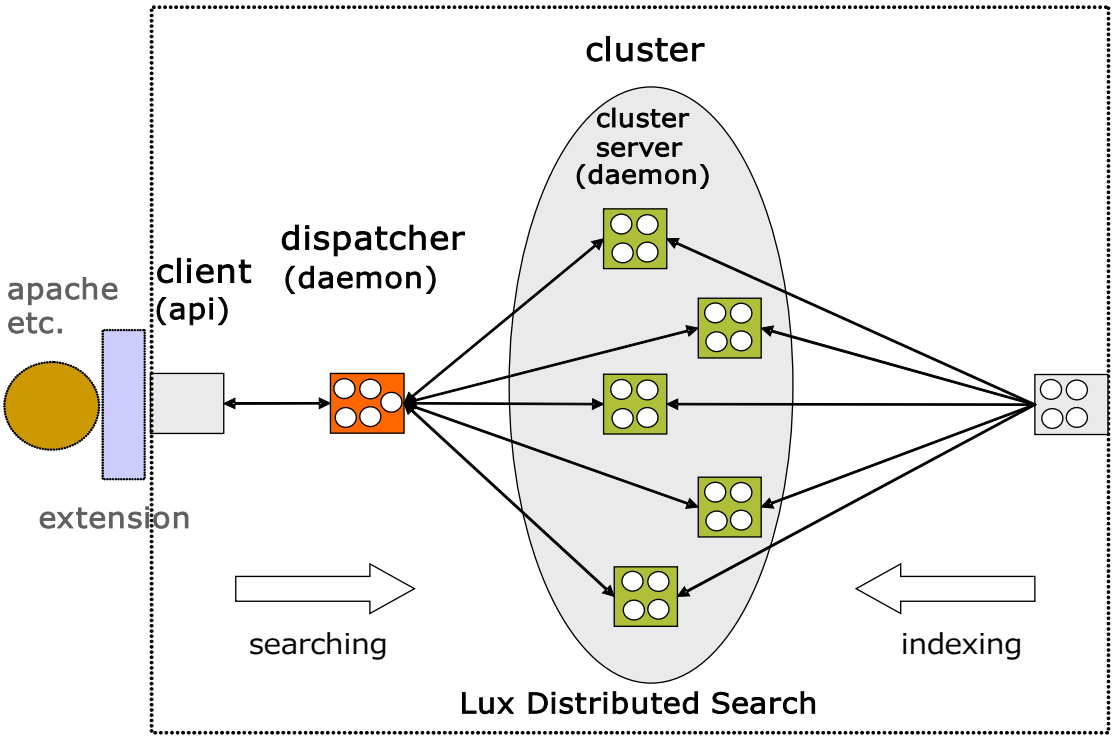
分散インデックスの開発

検索エンジンにおけるスケール性を確保するために、分散インデックスの機能を開発した。文書数が増えても性能が極端に劣化せず、マシン台数を増やせば扱える文書数も可能な限り線形に増加するという目標を掲げて開発を行った。また、ベンチマークにより、これらの目標がほぼ達成できていることを実証した。

動作環境

オペレーティングシステム: Linux, MacOSX (Unix-based Operating System)
コンパイラ: g++ (GNU C++)
必要なライブラリ: Lux IO, MeCab, Zlib, Libevent, GoogleProtocolBuffer





4. 従来の技術(または機能)との相違

Lux は、高速性とスケール性を追求しつつ、拡張性を持たせたアーキテクチャーとなっている。高速性は、その時代のハードウェアなどの環境に大きく依存するものであり、内部のアルゴリズムやソフトウェア構成は時代とともに変化していく必要がある。そのような状況において、柔軟に対応できる検索エンジンとなっている。
類似ソフトウェアとの比較表は以下の通りである。

	H.E	Senna	Lucene	Lux
性能	○	◎	△	◎
分散、スケールアウト	△	×	×(○)	◎
拡張性	×	×	○	○
使いやすさ	◎	△	△(○)	△
機能	○	△	◎	△

5. 期待される効果
有償の全文検索エンジンを利用していた企業において、Lux を利用することにより無償で高速な全文検索機能を利用できるようになる。また、大量のデータを高速に検索可能にすることで、今までその部分が技術的に難しく実現できなった事業へも比較的容易に参入できるようになり、新たなビジネスが生まれる可能性が期待できる。

6. 普及(または活用)の見通し
開発期間中にできなかったドキュメント作成やコミュニティの形成を早急に行う必要がある。機能としての全文検索を必要とする企業、開発者はとても多く、それらをきちんとやることが今後の普及につながると考えている。また、内部データベースの Lux IO はウェブページを英語で書いたことにより、海外の開発者からの問い合わせが非常に多い。よって、今後は英語でドキュメントを書くことによって、より世界へ普及させていける可能性を感じている。
現在は一人で開発を行っているが、今後はコミュニティを作り、複数人で開発していくことを予定している。

7. 開発者名(所属)
山田浩之(株式会社メタキャスト、慶應義塾大学理工学研究科)

(参考)開発者URL
http://luxse.sourceforge.net/
http://luxio.sourceforge.net/