平成15年度未踏ソフトウェア創造事業

# 圧縮方式高速全文検索エンジンの開発

開発者: 松永 拓

Raijin

## コンパクトなFull-Text Index Search Engine

検索エンジンにとって重要な要素のひとつに、任意の部分文字列を検索可能である、という項目が挙げられる。これは、文章内に出現する如何なる文字列を検索キーワードに入力しても正しい検索結果を得ることができるということである。この当たり前であるように思える要素が、現在普及しているほとんどの中小規模な検索エンジンでは、重要視されていない。

このような検索エンジンではインデックスを構築する手法として形態素解析などを用い文章から単語を抽出するKeyword Indexを採用している。しかし近年、Full-Text Indexと呼ばれる方式が注目されている。 Full-Text Indexでは、keyword Indexのように単語にぶつ切りせずにインデックスを構築する為、任意の部分文字列で検索することが可能であり、自然言語の文法に頼らないため、多言語に対応できる。しかし、インデックスサイズが大きくなるなどKeyword Indexに比べコストが大きい。そこで、本プロジェクトでは圧縮アルゴリズムを応用し、コストが低いFull-Text Index型の検索エンジンを開発した。


本プロジェクトの特徴は以下の通り

- 任意の部分文字列で検索可能
- 多言語に対応
- 省スペースなIndex
- Indexから本文テキストを復元可能
  この特徴を生かし、検索結果の表示に検索語周辺の文章テキストを表示可能