電子書籍の検索機能の改善

木下研究室　鴫原　善寿 (201002713)

1　まえがき

スマートフォン等の急速な普及は, 世界の移動体通信事業及びその関連産業の視界を一変させつつある. なかでもスマートフォンの普及とともに現れた電子書籍にはさまざまな会社が力をいれている. 電子書籍には以下のメリットがある.

資源 物理的な資源がいらず, 絶版がなくなる. 本棚もいらない.

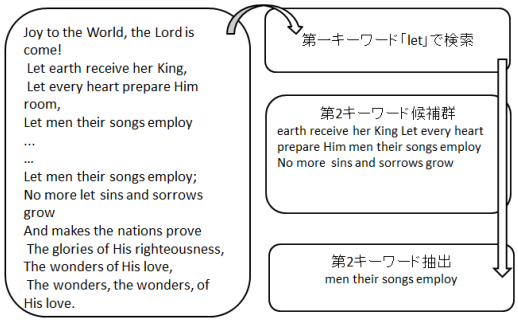
機能 文字の大きさやレイアウト変更だけでなく書籍内の検索も可能.

通信 インターネットでの購入, 配信. 語彙の検索も容易.

問題点として, 検索機能をつかったところ二語以上の検索ができなかった.web 上の検索システムでは検索対象が web ページであるため検索対象が日々更新されること, また第 2 キーワードが検索回数などで変わることがある. 以上の問題点を改善するための電子書籍の検索システムとして, 一冊の書籍から検索ワード (第 1 キーワード) と最も関係のある単語を第 2 キーワードとし抽出するアルゴリズムを提案する. 一般的に検索機能で扱うアルゴリズムと今回提案するアルゴリズムの 2 つのプログラムを用意し, 後者の方が単語の抽出の精度が高いことを示すことを目的とする.

2　提案手法

本研究では第 2 キーワードとなる単語の抽出まで行い, 抽出方法として tf-idf 法の応用である okapi-BM25 を使用し, 再現率と適合率によって抽出する単語の数を決め抽出をする.



2.1　検索方法

perl を使用した検索手順を以下に記す.

1 検索対象となるテキストをあらかじめ用意する.

2 tf-idf 法のプログラムと応用である okapi-BM25 のプログラムを用意する. 以下プログラム A,B とする. プログラム A で扱う tf-idf 法の式は

$$tf \times idf \tag{1}$$

となる. ある単語の文書内での出現頻度を t f（t erm f requency）といい, ある単語が含まれている文書の頻度を df(document frequency) という. 単語の出現頻度が高いことと, その出現頻度が高い単語が含まれる文書頻度が少ないという二つの条件を満たすことで単語の重み（重要性, 優先度）を特徴付ける.tf(単語の出現頻度)idf(df の逆数) とする. ここで, 大域的重み付けとなる idf(inverse document frequency）は,df の逆数であり,df との対数をとったものとする. 対数をとることによって, ほとんどの文書で出現する単語の影響を小さくする. よって idf は以下の式

$$idf = log(N/df) + 1 \tag{2}$$

で関係づけられる. プログラム B で扱う okapi-BM25 の式は

$$idf = \frac{(K1+1)tf}{k1((1-b)+b\frac{dl}{avdl}tf} log\frac{N-df+0.5}{df-0.5} \tag{3}$$

となる. 式（2）に比べて, 式（3）では文字列の平均値をとっている.tf-idf 法では文書長が長いものほど tf の値が大きくなりすぎる傾向があるため, 多くの文書を解析しつつ絶対評価的にスコアリングする場合, 特徴語のスコアにムラが出るという欠点がある.okapi-BM25 は文書の長さの平均化をはかり, スコアリングする文書が大体どれくらいの長さなのかを比率的に計算することで特徴語のスコアのムラを小さくしている.

3 コマンドプロンプト上で用意したテキストにプログラム A,B を使用し, 検索ワードを含む文を取り出し文字列, 行数を取り出し計算する.

4 各々の検索結果を表, グラフにまとめる.

3　考察

プログラム A, プログラム B での検索結果が, 検索対象のテキストの文字列の長さ, 行における文字列の長さにばらつきがでる事によってどの程度の誤差が生じたかを数値化し表にする. 検索結果の表を比べることで,okapi-BM25 のプログラムのほうがより優れていることを示す.