組み合わせ型情報検索手法の検索実験による評価

1T－5　安形 輝　itasan@slis.flet.mita.keio.ac.jp
慶應義塾大学大学院文学研究科

1．はじめに

情報検索研究において、現在までに、複数の検索手法を組み合わせて検索効率の向上をはかる提案がイングベルセン(Ingwersen, Peter)[1]などの研究者によって行われてきた。これは複数の検索手法によって検索された文献の方が単一の検索手法でしか検索されなかった文献よりもレレバントである可能性が高いという仮定に基づくものである(図1)。しかし、実際に検索実験を行った研究はほとんどない。そこで組合わせ型検索手法を用いた検索実験を行い、その有効性について検討を行った。

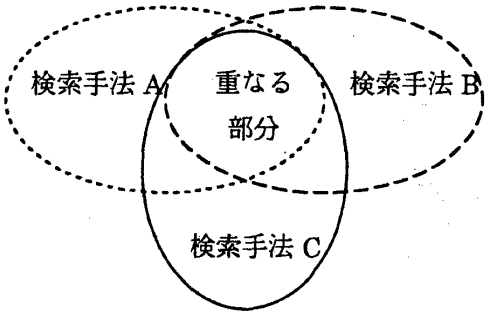

図1　三つの検索エンジンの組み合せ

2．検索手法

2.1　組み合わせ型検索手法

組み合わせ型検索手法における検索は、それぞれの検索手法による検索結果を重み付けし、組み合わせることによって行う。例えば、A,B,C の三つの検索手法を用いた場合の検索について考えると、個々の文献の重要度の算出は以下のようになる。検索結果は重要度の高い順に出力を行う。

$$W_i = \frac{W_{Ai}}{AveW_A} \times C_A + \frac{W_{Bi}}{AveW_B} \times C_B + \frac{W_{Ci}}{AveW_C} \times C_C \quad (1)$$

この式で W_{Ai} は手法Aの文献の重要度、$AveW_A$ は手法Aの文献の重要度の平均、C_A は手法Aに対する重み付け定数を表している。今回は、各手法に対する重み付け定数を 1 にした場合と、それぞれの検索手法の検索効率を正規化したものを重み付け定数とした場合(つまり、最適な場合)の二つについて実験を行った。組み合せる検索手法としては、それぞれ異なる観点からのマッチングを行うと考えられるベクトル空間型、クラスター型、論理型検索手法とした。

2.2　ベクトル空間型検索手法

ベクトル空間型検索手法では、文献と検索質問を索引語・検索語に対応したn次元の空間におけるベクトルとして表現し、それぞれのベクトル同士の類似度を算出することによって検索を行う。一般的にはベクトルとして表現する際に、語の重要度を表現するために、語の出現頻度などの情報を利用して、各語の重みを算出する。

ベクトル空間型検索手法の特徴としては、以下のようなものがあげられる。

1)従来のシステムで使われてきた論理型検索手法と比較し検索効率が高いこと
2)検索結果を類似度の高い順に出力するために順位付け出力が行われること、

ベクトル空間型における索引語の重み付けには、ソルトン(Salton, Gerald)の検索実験[2]で最も性能の良かった以下のような重み付けの公式を採用した。

$$Weight(Ti) = \frac{tf_i}{\max tf_i} \times \log\frac{N}{n} \times \frac{1}{\log N} \quad (2)$$

tfは文献中の語の出現頻度、Nは総文献数、nは語の出現文献数である。

文献ベクトルと検索質問ベクトルの間の類似度の算出にはコサイン相関関数を用いた。検索結果の出力は類似度の高い順に上位20件まで行った。

2.3　クラスター型検索手法

クラスター型検索手法では、あらかじめ文献同士の類似度を算出し、類似した文献群によって階層型クラスターを作成しておく。検索は検索質問とクラスター間の類似

Evaluation of Combined Retrieval Method: Information Retrieval Test
Teru Agata
Graduate School of Library and Information Science, Keio University
2-15-45 Mita, Minato-ku, Tokyo, 108, Japan

度を算出し、最も類似度の高いクラスター中に含まれる文献を出力することによって行われる。クラスター型検索手法は大きく階層型と非階層型に分けられるが、非階層型はベクトル空間型検索手法の一種と考えられるため、ここでは階層型を採用する。

クラスター型検索手法の特徴としては以下のようなことがあげられる。

1)検索質問と直接的な関係の見られない適合文献を検索することができること

2)自動分類から発展し性能面は比較的軽視されてきたため性能は良くないこと

クラスターの作成手法は、従来の研究においては、単一リンク、完全リンク、グループ平均、Wardの手法などが提案されてきた。グリフィス(Griffiths, A.)他[3]の研究では、単一リンクは他の三手法より性能が低いとされたが三手法の間では性能面での違いは見られなかったことから、ここでは、比較的簡単に実装可能な完全リンクを採用する。検索方法は、原則的にブルジン(Burgin, Robert)[4]の研究での完全リンクと同様の方法を用いる。2つの文献A,B間の相違性の算出には以下のようなMarczewski-Sterinhaus尺度を採用する。

$$1-\frac{A\cap B}{A+B-(A\cap B)} \quad (3)$$

ここで$A\cap B$はA,Bで共通な索引語を示す。実験では検索質問と類似度の高いクラスターから階層を上位に遡りクラスターを文献数が20件になるまで出力し順位付けを行った。文献の重要度はその文献の含まれるクラスターの類似度とした。

2.4 論理型検索手法

論理型検索手法は、従来商用データベースシステムで最もよく使われてきた手法である。検索質問はブール演算子を用いて記述し、検索は、検索質問中の語を文献中の索引語と対照させることによって行われる。

論理型は文献に対する重要度は付与されないため、全文献の重要度を1として計算を行っている。

3．検索実験

3.1 実験データ

検索実験用のデータベースは、情報検索分野の英語文献の書誌データ 1698 件から構成され、用意された検索質問についてのレレバンス判定が行われている。

3.2 検索手法の評価

それぞれの検索手法を用いた検索結果については再現率・精度を算出した。検索効率の評価では比較のために以下のようにE尺度も算出した。この式において、Pは精度、Rは再現率を示す。

$$E=1-\frac{2PR}{P+R} \quad (4)$$

4．おわりに

ここでは実際に組み合わせ型検索手法を使った簡単な検索実験について報告を行った。今後の方向性として考えられるものは以下の通りである。

1)レレバンスフィードバックを用いた検索手法を使った組み合わせ型検索

2)インターネット上の複数の検索エンジンを統合するメタ検索エンジン

3)複数のデータベースを組合わせる検索

5．謝辞

本研究を進めるにあたり、懇切なるご指導をいただいた慶應義塾大学文学部の上田修一教授に深く感謝する。

＜引用文献＞

1) Ingwersen, P. “Cognitive perspectives of information retrieval interaction: Elements of cognitive IR theory”. Journal of Documentation, vol.51, no.1, p.3-50(1996)

2)Salton, G.;Buckley, C. "Term-weighting approaches in automatic text retrieval". Information Processing & Management. vol.24, no.5, p.513-523 (1988)

3)Griffith, A.; Robinson, L.A.; Willett, P. “Hierarchic agglomerative clustering methods for automatic document classification”. Journal of Documentation, vol.40, no.3, p.175-205(1984)

4) Burgin, Robert. “The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity”. Journal of the American Society for Information Science, vol.46, no.8, p.562-572(1995)