単語共起によるクエリー展開を用いた大規模テキスト検索

5Ｐ－4

佐藤研治　赤峯 享　奥村 明俊

NEC 情報メディア研究所

1. はじめに

近年のネットワークの発展により、大規模なテキストデータベースを個人単位で使用する機会が生じてきた。このため大規模テキストに対する実用的な情報検索システムに対する要望が高まってきている。

本研究所では、GByte 級の英語テキストデータに対する情報検索システムを構築した。本システムでは、検索要求文は自然言語で入力し、その検索要求文からクエリーを自動生成し、そのクエリーとドキュメント中の単語とのマッチングを行うことで検索を行う。

一般に、情報検索システムでは、検索要求文中に含まれる単語のみを用いてクエリーを生成し、そのクエリーをそのまま用いて検索を行っても、検索者の必要としている文書が十分には検索できないことが知られている。これは、検索要求文と検索対象ドキュメントでは、異なる単語で同一の意味内容を表現していたり、検索要求文中の単語を抽出した時点で要求文の意味内容が損なわれたりすることが主な原因である。

クエリーが十分な検索を行うだけの単語をもっていない場合には、クエリー中に検索に必要な単語を追加するために、クエリー展開(Query Term Expansion)を行う必要がある。このクエリー展開の方法としては、シソーラスを用いて同義語等を展開する手法が知られている[1]。

本稿では、既存のシソーラスを用いるのではなく、ドキュメント中でクエリー単語と共起する単語を用いてクエリー展開を行う方法を提案する。そして、本手法を実際の情報検索システム中にインプリメントし、GByte 単位のドキュメントを用いて検索を行い、その検索精度の評価を行ったのでその評価結果についても報告を行う。

2. 単語共起によるクエリー展開

サーチャーが人手で検索クエリーを構築／修正する際によく用いる手法として、検索対象ドメイン内での固有名詞の展開がある。たとえば、計算機メーカーに関連する検索要求文であれば、"computer"や"hardware"といった単語から"IBM", "HP", "DEC", "Apple"等の単語に展開し、クエリーに追加する。このような展開を行うと、良い検索結果が得られることが知られている。

これらの固有名詞は、一般に、クエリー単語がドキュメント中に出現する位置の前後にしばしば現れるため、ドキュメント中での単語の共起情報を用いて収集することが可能である。

また、一般名詞は単語に多義性があり、検索要求文と異なる意味で使われているドキュメントも拾ってしまう可能性があるが、固有名詞であれば多義性が極めて低いため誤検索してしまうことが少ない。

共起情報は、クエリー単語と関連する固有名詞を同一文脈内で幅広く集めることを目的として、テキスト中でクエリー単語と同一段落内に出現する固有名詞を収集する。そして、それらの固有名詞の中で共起頻度の高いものをクエリー展開に用いる。

3. 実験システム

検索対象テキストは、Federal Register, IR-Digest, Virtual Worlds, Newsgroups からなる科学技術分野のテキスト約 1GByte を用いた。前もってこれらのテキストに対し、単

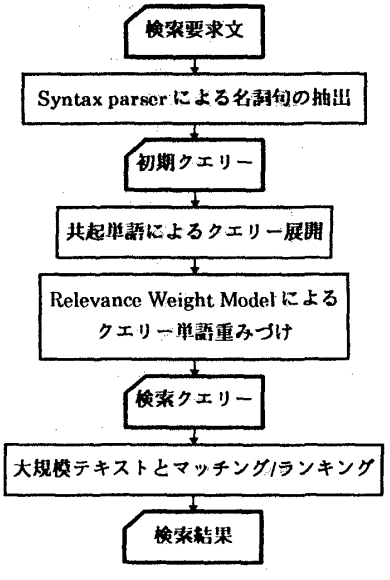

図１：検索システム全体の処理の流れ

Information Retrieval System Using Query Term Expansion Based on Word Co-occurrence Database
Kenji SATOH, Susumu AKAMINE and Akitoshi OKUMURA
Information Technology Research Labs. NEC Corp.

語の出現位置まで検索できるインデクスを用意した。このインデクスのインバーティッドファイルの容量は、約860Mbyteである。

検索システム全体の処理の流れを図1に示す。システムはまず、自然言語で書かれた検索要求文から初期クエリーを生成する。初期クエリーは、Syntax parserを用いて、検索要求文から特定の予約語(document, relevance 等)以外の名詞句全てを抜き出すことで生成される。

クエリー展開の処理では、初期クエリーの単語と同一段落内で共起する固有名詞(先頭が大文字の英単語)をドキュメントより収集する。共起単語を収集した後は、共起頻度の高い(ドキュメント中で同一段落に出現する確率が0.1以上)固有名詞を選択し、元クエリーに追加する。

この共起情報および下記の重み計算で用いるドキュメントは、検索対象と類似したドメインのドキュメントで600Mbyteある。このドキュメントと評価で用いたドキュメントとの間には重なりはない。

クエリー拡張された単語に対し、正解データ(別ドキュメントセットに対するもの)の情報を用いて、以下のRelevance Weight Modelに基づく重みを与えている。[2]

$$w_i = \frac{r_i/(R-r_i)}{(n_i-r_i)/(N-n_i-R+r_i)}$$

N：全ドキュメントの数
n_i：単語iを含むドキュメントの数
R：正解ドキュメントの数
r_i：単語iを含む正解ドキュメントの数

重みづけを行って生成されたクエリーに対して、検索を行う。検索は、クエリー単語とドキュメントのマッチングを行い、単語の重みによってドキュメントをランキングすることによって行う。ランキング式はドキュメント中に1回以上出現したクエリー単語の重みの和である

この実験で用いた検索要求文、検索対象ドキュメント、および、重み計算と評価で用いる正解データは全てText Retrieval Conference(TREC)に参加したサイトに配布されたデータを用いている[3]。TRECでは、多数のサイトが同一データに対し検索結果を提出し、その結果の和集合を人手で判断することで正解データを作成している。このデータを用いることで、GByte単位のドキュメントに対する検索精度を評価することが可能になっている。

4. 評価

評価実験は50文の検索要求に対して、クエリー展開を行ったものと行わないもののそれぞれについて、処理を行った。検索結果はランキングの上位1000ドキュメントまで順位をつけて出力し、その出力結果中に含まれる正解ドキュメントの数で評価を行った。

評価は、情報検索の分野でよく用いられているrecall-precisionを50の検索要求で平均したもので行った。この評価結果のグラフを図2に示す。このrecall-precisionのグラフは、全体が上に行くほど良い検索精度を示す。

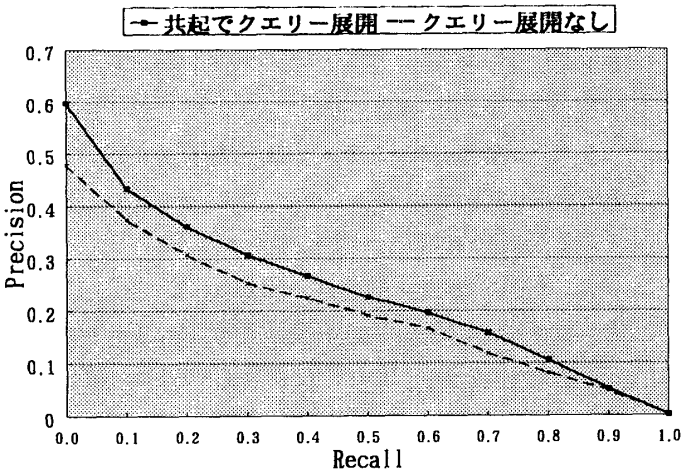

図2：評価結果

共起情報を用いてクエリー展開を行うと10%程度の検索精度の改善が見られた。これは、固有名詞がドキュメントの選択性が非常によく、更にRelevance Weight Modelによる重みと固有名詞の整合性が良かった為と思われる。

5. おわりに

単語共起を用いたクエリー単語展開法を、GByte単位の大規模テキスト検索システムにインプリメントし、その有効性を実証した。今後は、共起情報と既存シソーラスを組み合わせたクエリー展開法等を、大規模検索システムにインプリメントし手法の有効性を評価していく予定である。

参考文献

[1] W.B.Frakes 他,"Information Retrieval", Prentice Hall

[2] D. Harman , "The Second Text Retrieval conference (TREC-2)", NIST Special Publication 500-215

[3] D. Harman, "Overview of the Third Text Retrieval Conference(TREC-3)", NIST Special Publication 500-225