シソーラスによるクエリー展開を用いた大規模テキスト検索 *

5P－5　赤峯享　佐藤研治　奥村明俊 †

ＮＥＣ 情報メディア研究所 ‡

e-mail: {akamine,satoh,okumura}@hum.cl.nec.co.jp

1　はじめに

新聞・論文・特許文・社内文書等の大量の文書の電子化、及び、コンピューター・ネットワークの普及により、サーチャー等の専門職でない一般のユーザが直接、大量の文書を検索することが可能な環境が整いつつある。それにつれて、一般ユーザが手軽に文書を検索できるシステムの要望が高まっている。

筆者らは、自然言語で記述された検索要求書からクエリーを作成し、そのクエリーと文書中の単語のマッチングを行うことで検索を行う英文書検索システムを構築した。自然言語文で、検索を行うメリットとしては、(1) 誰でも簡単に記述できる、(2) 検索対象の文書を正確に記述できる (検索結果の評価が可能である)、(3) 類似した文書を検索するシステムへの拡張が容易である (例えば、自分が現在書いている論文と類似した文書を検索する)、等があげられる。しかしながら、検索要求書中に記述された単語と実際に文書中に含まれる単語の間には、同じ意味でも表現にズレが生じるため、検索要求書中に記述された単語のみでクエリーを作成したのでは、適切な文書が十分に検索できないという問題が生じた。

この問題を解決する方法として、シソーラスや共起情報を利用して検索入力の単語を展開することで検索を行う方法が提案さている。しかしながら、実用規模の文書に対する評価について、ほとんど報告されていない。今回、ギガバイト単位の英文書に対して、汎用のシソーラスの同意語・下位概念語等を利用してクエリーの展開を行い、検索精度の評価実験を行ったので報告する。

2　シソーラスを用いたクエリーの展開

同意語をクエリーに追加することで、検索洩れを少なくすることは情報検索において頻繁に行われている。しかしながら、既存の汎用の知識ベース (シソーラス) を用いて自動的にクエリーを展開することで、どの程度の検索精度が向上するのかについての客観的な報告は、なされていない。

今回、この評価を行うためのシソーラスとして WordNet[2] を用いた。WordNet は、プリンストン大学で作られた英語のシソーラスであり、約９万語 (その内名詞は約６万語) について、同意語、反意語、下位語、上位語等が与えられている。例えば、“cat” には、６つの意味があり、表１に示すような同意語、上位語、下位語が登録されている。

また、実際の評価実験では、クエリーが過剰に展開されることを防ぐために、名詞の同意語と下位語のみを利用した。例えば、“cat” は同意語では、{“cat”, “true cat”, “caterpillar”, “cat-o'-nine-tails”, “big cat”, “computerized axial tomography”} に展開され、下位語では、{“domestic cat”, “wildcat”, “tiger”, “lion”, … } に展開される。

3　検索方式

検索方式の概略の流れを図１に示す。検索は以下の手順で行った。具体的な方式については、同様の方式を用いた参考文献 [1] を参照されたい。

1. 自然言語で書かれた検索要求書から名詞句及び名詞の単語を初期クエリーとして抽出する。
2. 初期クエリーの同意語 (下位語) を WordNet を利用して作成し、クエリーに追加して検索用のクエリーを作成する。
3. 人手で作成された正解 (不正解) データを元にクエリー中の各単語に重み付けを行う。
4. 単語のベクトル空間モデルを用いて、検索文書とのマッチングを行い、ランク付けされた検索文書を出力する。

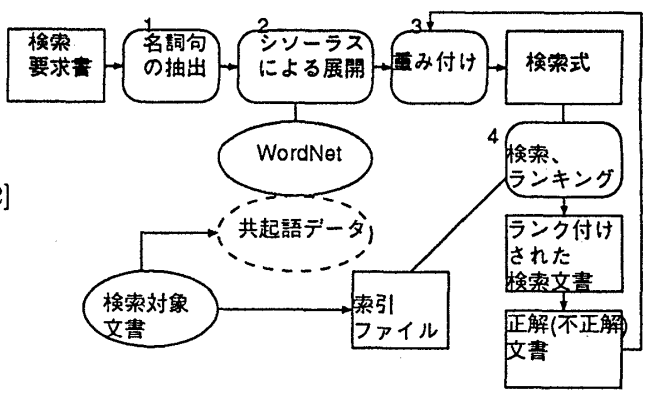

図 1: 検索方式

*Information Retrieval system with Query Term Expansion using WordNet

†Susumu AKAMINE, Kenji SATO and Akitoshi OKUMURA

‡Information Technology Labs.,NEC Corp.

表 1: WordNe における cat に対する同意語、上位語、下位語

	同意語	上位語	下位語
意味 1	cat, true cat	feline, felid	domestic cat ／ wildcat
意味 2	cat	gossip, ／ woman, …	
意味 3	caterpillar, cat	tractor	
意味 4	cat-o'-nine-tails, cat	whip	
意味 5	big cat	feline, felid	tiger,lion,leopard,jaguar …
意味 6	computerized axial tomography	x-ray	

4 実験

検索対象の文書は科学技術分野の約 2 ギガバイト (約 60 万文書) を用いた。50 の検索要求書に対して、「展開なしクエリー (平均 40 単語)」、「同意語で展開したクエリー (平均 110 単語)」、「下位語で展開したクエリー (平均 470 単語)」、「同意語及び下位語で展開したクエリー (平均 540 単語)」に対して、正解／不正解データを利用して重要度の重み付けを行い、検索を行った。なお、これらの英文書や正解データは、Text Retrieval Conference(TREC)[3] に参加したサイトに配付されたデータである。

検索結果は、上位 1000 文書に対して、0.1 刻の Recall(再現率) に対する Precision(適合率) を用いて評価した。その結果、個々の検索については、約 4 割の検索要求対して、同意語／下位語で展開することで検索精度が向上した。しかし、平均では展開を行ったものと行わないもの間で、検索精度にほとんど差は付かなかった。50 個の検索結果を平均した評価結果を表 2 に示す。

表 2: 検索結果 (Recall に対する Precision の値)

Recall	展開なし	同意語	下位語	同意と下位
0.0	0.544	0.545	0.548	0.553
0.1	0.411	0.399	0.405	0.404
0.2	0.339	0.326	0.336	0.331
0.3	0.282	0.253	0.273	0.265
0.4	0.246	0.222	0.235	0.225
0.5	0.209	0.195	0.196	0.192
0.6	0.181	0.187	0.173	0.173
0.7	0.128	0.123	0.131	0.131
0.8	0.084	0.078	0.087	0.093
0.9	0.048	0.041	0.046	0.049
1.0	0.002	0.002	0.002	0.002

5 考察

シソーラスを用いて、クエリーを同意語／下位語に展開することで、検索をおこなったが、期待した検索結果の向上は、得られなかった。理由としては、(1) シソーラス中には、キーワードとなりやすい固有名詞や専門用語があまり含まれていなかったこと、(2) 単語の多義性のために無関係な単語を展開してしまい (例えば、"caterpillar" の意味の "cat" を同意語で展開した結果、"domestic cat", "tiger", "lion" 等の無関係な語を大量に生成してしまい)、正解データからのフィードバックだけでは十分な単語の絞り込みができなかったこと、があげられる。

単語の多義性の中から適切な語義を選択するために、同一段落中の共起情報を用いた展開方式 [1] と組み合わせる方法が考えられる。これは、同意語で展開した単語の中で検索クエリーとして適切な語義の単語は、他の初期クエリー中の単語との共起頻度が高いと仮定し、共起頻度が高いものだけを検索クエリーとして追加を行う方法である。今後、この方法で展開する単語の絞り込みを行い、評価を行う予定である。

6 おわりに

ユーザの検索入力と実際の文書中の表現のズレを吸収するために、シソーラス (WordNet) を用いた同意語の展開を行い、検索を行う方式の評価実験を行った。その結果、同意語、下位語を用いた展開に関しては、約 4 割の検索について検索精度が向上したが、平均では、検索精度の向上は見られなかった。これは、単語の多義性等の理由のために過剰にクエリーを展開してしまったためであると考えられる。今後は、シソーラスと共起情報等を組み合わせることで、単語の多義性の絞り込みを行う方式の評価を行う予定である。

参考文献

[1] 佐藤研治他, " 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 ", 第 52 回情処全国大会

[2] "ftp://clarity.princeton.edu/pub/wordnet"

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