EDR電子化辞書を用いたクエリー拡張による検索支援

太田 千晶, 奥村 学
{chiaki, oku}@jaist. ac. jp
北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

本研究では, 電子化辞書中の概念関係を用いてクエリー拡張を行い, その結果とユーザから入力される情報の両者を利用して, よりユーザの要求に近い情報を発掘することを目的とする. 本稿では, このうち前者について述べる.

これまでのシソーラスを用いたクエリー拡張の研究は様々であるが, 単語の表記を基づく拡張（類語辞書の使用など）には限界があり, 単語の概念等を用いた深層的な関係を用いたものがより有効であると考えられる [Ellen 95][下畑 96].

そこで本稿では, 大規模なエントリを持ち, 様々な概念間関係を記述した辞書を持つ, 電子化辞書 EDR を用いた拡張の方法および, 概念階層による重みを加味した検索文書スコアリングの方法を提案する. 一般的には, クエリー拡張により適合率は非常に低下するが, この方法により各手法で拡張なしの場合を上回る適合率を得ることができた.

以下, この手法について詳しく述べるとともに, BMIR-J1 を用いた本手法の評価実験の結果を示し, 考察を行う.

2 概念間関係を用いたクエリー拡張の手法

2.1 電子化辞書 EDR

今回利用する電子化辞書 EDR では, 各単語にその概念を表すとされる 6 桁の英数字である概念識別子が割り当てられている. 電子化辞書 EDR は, 概念見出し辞書, 概念体系辞書, 概念記述辞書という 3 種の概念間関係を表す辞書の他, 単語辞書, 共起辞書, 対訳辞書など複数の辞書から構成されているが, これらはこの概念識別子を基に関連付けられている. 本稿では, これらのうち 3 つの概念辞書および日本語単語辞書を使用し, 以下のような拡張を行った.

2.2 クエリー拡張の手法

2.2.1 同等概念からの拡張

本手法は, 初期クエリーと同じ概念識別子を持つ単語のレコードに含まれるものをクエリー候補として獲得する. この方法で得られる単語は, 初期クエリーの類義語または言い替え語であると考えられ, また表記の揺れにも対応することが可能である.

表 1: 同等概念からの拡張例

初期クエリー	獲得クエリー候補
フェスティバル	祭, フェスチバル, フェスチヴァル . . . etc

2.2.2 上位・下位概念からの拡張

概念辞書のうちの概念体系辞書は, 2 つの概念間の関係のうち上下関係を記述したもので, グラフ構造を構成する. よって, このグラフ上のある概念識別子から上または下につながる概念識別子を辿って行くことにより, 初期クエリーの抽象度を高めた単語, あるいは具体度を高めた単語を得ることが可能となる. 本手法はこれらを関連語として, クエリー候補とするものである.

なお, EDR の上位下位関係を表すグラフを分析すると, 概念階層間の抽象化（具体化）の進み方は一定ではなく, 最下数階層に 一般的に使われる単語が持つ概念識別子が集中しており, それ以上の階層に上がると抽象化の進み方が非常に大きくなるという特徴を得た. よって, 上位・下位概念をすべて辿り, それら全てからクエリー候補を生成することは望ましくないと考えられる. よって, 本稿では上から何階層目までの概念識別子を拡張に使用するかという閾値を求めるための実験を行った. これについては, 4.3 で述べる.

2.2.3 概念記述辞書からの拡張

概念記述辞書は, 2 つの概念間関係のうち上下関係以外の関係を記述したもので, 現在のところ 動詞的概念と名詞的概念の関係についてのみ記述されている. これらの 2 つの概念間の関係は 8 種類の概念関係子であらわされているが, 今回は概念関係子の情報は使用せず, 概念識別子のみを利用している.

具体的には, クエリーが名詞であるか動詞であるかを日本語単語辞書のレコードから判断し, それとその単語の持つ概念識別子から対となる概念識別子を得, それを基にクエリー候補となる語を獲得した.

本手法により, 一般的な辞書では得られない動詞・名詞間の連想を実現することができ, またさらに, 一般的にクエリー拡張においては名詞のみが利用されている中で, 動詞の情報も有用に利用することができると考えられる.

表 2: 概念記述辞書からの拡張例

初期クエリー	獲得クエリー候補
アサガオ	植える, 成長する, 咲く, 芽吹く . . . etc

3 クエリー候補の獲得

以上の 4 手法をもとに, クエリー候補を生成する過程を図 1 に示す.

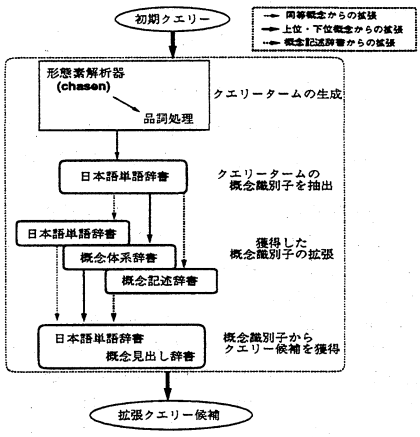

図 1: クエリー候補の生成過程

今回は以上の 4 手法で得られた概念識別子を持つ日本語単語辞書中のレコードから, クエリー単語として, **単語見出し**, **概念見出し**, **概念説明**をクエリー候補として獲得した. なお, 概念説明については, 一文で書かれており, そのままクエリー候補とするのは不適であるため, 形態素解析を行って, 有用な自律語のみをクエリー候補として獲得している.

4 評価実験

4.1 評価方法

上記の 4 手法を評価するため検索システムを実装し, 実験を行った.
実験には, 情報検索システム評価用ベンチマーク BMIR-J1 を使用した. この BMIR-J1 では検索文書 600 件とクエリー 60 種およびその正解が与えられる. 本稿では, これらの検索文書 600 件すべてを形態素解析器 [松本 96] にかけ, 必要な自立語[1]のみを抽出して, 検索用データベースを作成した. また, 用意されているクエリー 60 種は 6 種類のファンクションに分類されており [BM 96], 今回使用したクエリーはこのうち「キーワードの存在確認, あるいは, キーワードのシソーラスによる展開語の存在確認」に用いる「基本機能」ファンクションに属するクエリー 8 種[2]とした.

また実験は, 本手法を用いて拡張されたクエリーセットを入力とし, tf・idf 法を用いてスコアリングされた文書集合を出力とする検索システムの上で行った.

$$w_{dt} = \frac{1 + log(tf_{dt})}{log(length(d))} * log\frac{N}{n_t} \quad (1)$$

ただし, w_{dt}は単語 t に対する文書 d の重要度, tf_{dt}は d における t の出現頻度, $length(d)$ は d の長さ（文字数）, Nは総テキスト数, n_tは t を含む文書の数である [木谷 96]. 実際の文書スコアは, 式 (1) によって得たある文書 d の単語ベクトルと, 同様に生成されたクエリーセットの単語ベクトルの内積計算によって求められる.

また, 検索システムの出力評価は, 式 (2), 式 (3) に定義した recall, precision を用いて行った.

$$\text{recall} = \frac{\text{システムの出力中の正解文書数}}{\text{全正解文書数}} * 100 \quad (2)$$

$$\text{precision} = \frac{\text{システムの出力中の正解文書数}}{\text{システムが出力した文書数}} * 100 \quad (3)$$

4.2 各手法の評価・考察

各手法における上記の実験結果を図 2 に示す.

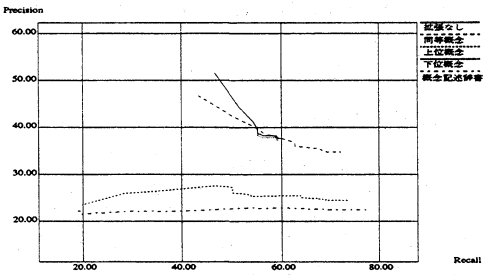

図 2: 4 手法による実験結果

考察

- 同等概念からの拡張
  拡張なしと比較して Recall 値はかなり上昇しているが, precision 値において拡張なしを上回る成果をあげているのは一部分のみである. precision 値を下げたひとつの理由としては, 今回使用したクエリーは具体的な企業名を求められているものが 2-3 含まれていたが, EDR ではそれに対応できなかったこ

[1]名詞, 動詞, 未定義語

[2]基本機能のクエリー 10 種のうち, 正解数が 5-30 の範囲外にある, ベンチマーク作成者によって「不適」とされるものを除いた 8 種

とがあげられる．また単語見出しのみと概念情報のみによる拡張効果を比較したところ（図3），全体的な precision 値を下げている原因は概念情報であるが，recall 値の上昇には概念情報が役に立っていることがわかる．

- 上位概念からの拡張
  これについては，4.3 で詳しく述べる．
- 下位概念からの拡張
  拡張なしとほぼ同様の結果を得た．先にも述べた最下数階層において情報が密であるという特徴にあったように，実際に単語を検索してみると一般的に使われるほとんどの単語において下位概念は存在しない場合が多い．今回使用したクエリーにおいても同様で，8 種のクエリーのうち下位概念を持つものは 1 種のみであった．この特徴は，別に行ったクエリー 40 種を用いた実験でも同様に見られた．実際に獲得されたクエリー候補は，初期クエリーをより具体的に表現した単語であり，本手法による獲得単語はクエリー拡張に適すると考えてよいだろう．
- 概念記述辞書からの拡張
  拡張なしを大きく下回る精度であったが，この一番の原因は本手法により獲得されたクエリー候補が膨大な数に及んだことであると考えられる．
  しかし，実際に検索結果を分析した結果，他の手法で獲得できなかった正解が本手法によるクエリー候補によって獲得できているという例がいくつかみられ，また recall 値は拡張なしを約 20%上回る結果を得ている．このことや先にも示した例からわかるように，本手法により獲得されるクエリー候補は他の手法ではほとんど得ることのできないクエリー拡張にとって非常に有用な語であると考えられる．しかし，少しでも関係のあると思われる概念識別子全てがエントリに存在するため非常に獲得候補数が多いという欠点を持ち，それらの中でどれが重要であるかという判断なしには，本手法が実際に精度の向上に貢献することはできないと考えられる．図 4 に示した概念情報のみを拡張に使用した場合の結果が比較的良いことから，これらの概念情報を有効に利用する方法や、また検索文書内の情報を利用する方法等を今後検討していく予定である．

4.3 概念階層を用いた閾値実験

先にも述べたように，上位概念からの拡張において上位概念識別子を全て拡張に用いるのは不適であると考えられる．そこで，さらに以下のような方法で，拡張に用いるのに最適な概念階層を決定するための実験を行った．この実験の結果を図 5 に示す．

1. 全ての概念識別子に共通な最上位概念識別子（3aa966）は拡張に使用しない．
2. 1. の処理後，$0 \leq n \leq N$において，最上位から数えて n番目以下の概念識別子のみを使用してクエリー候補を獲得し，各々のクエリーセットを準備する．
3. 各クエリーセットにおいて，4.1 で述べたのと同様の実験を行う．
   ただし，概念階層別に重みを決定し，その重みを各クエリーの tf・idf 法に基づいて算出されるスコアに乗じ，それをそのスコアとする．

考察

- 最上位から n階層目までの概念識別子を削除して，残る概念識別子を拡張に使った場合のものを L_n と呼ぶとすると，図 5 の L'_0 および L_0 より，3. で提案した階層重みを加味したスコア計算の効果が，非常に大きいことがわかる (precision 値は約 30%上昇)．なお今回は，初期クエリーから数えて m番目の概念階層にある概念からの拡張クエリーに対する階層重み w_m を，$w_m = 0.5^m$ としている．
- 概念階層別に 3. のスコアリングを用いて評価した結果が L_0-L_{12} であるが，L_7の時初めて拡張なしを上回り，L_9をピークにそれ以降は除々に減少傾向にある．また，そこで，L_7-L_9の場合をさらに考察してみると，L_7が最も高い recall 値を得ており，実際に人手で上位概念から得られる情報を調べてみた結果（表 3）からも，n=7 あるいは n=6 くらいまでが拡張として用いる許容範囲ではないかと判断される．よって以上の結果より，上位概念からの拡張を行う場合，n=7 を閾値として拡張を行うのが最適ではないかと考える．

表 3：上位概念階層別獲得単語例

n	例 1	例 2
9	-	メーカー
8	菓子	業種や社名で捉えた会社
7	お菓子	会社
6	食べるもの	経済組織
5	飲食物	組織のいろいろ
4	機能で捉えた具体物	組織
3	静物	自立活動体
2	具体物	主体
1	もの	もの
0	物事	物事

5 おわりに

本稿において提案したクエリー拡張および検索文書のスコアリング方法を用いた評価実験において，3 手法（同等概念，上位概念，下位概念からの拡張）では各々拡張しない場合を上回る精度が得られることがわかった．また，概念記述辞書からの拡張においては拡張なしの場合を下回る結果しか得ることができなかったものの，この手法から獲得できるクエリー候補は非常に有用であり，今後獲得されるクエリー候補の重要度の決定方法を模索していくことにより，よりよい結果を得ることができると期待される．
今回の実験に際しては、実験に用いるクエリーセット数が

少ない、拡張クエリーを全てOR演算子で結合した検索を行っているという問題点があり、今後はこれらに対処してシステムを修正し、その評価実験を行う必要があると考えられる.

さらに, 実際のユーザの検索支援の場において, 本手法によって獲得されたクエリーをどのようにユーザに提示していくかという大きな問題が残されており, これを今後検討していく予定である. またそれと同時に、獲得されたクエリー候補を介してさらにユーザから何らかの入力を得ることで, さらにDATAMINING に有効なインタラクティブなクエリー拡張の手法を提案していく必要があると考えている.

謝辞

本研究では, 株式会社 日本経済新聞の協力によって, 社団法人 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが, 1993年9月1日から12月31日の日本経済新聞記事を基に構築した情報検索評価用データベース（テスト版）を使用した. この使用を許可して下さった同グループに感謝致します.

参考文献

[Ellen 95] Ellen M. Voorhees, "on Expanding Query Vectors with Lexically Related Words" *TREC3*, pp.223–231, 1995

[下畑 96] 下畑 光夫 坂本 仁, "多様分類情報による検索語拡張", 情処研報 *96-NL-115-19*, 1996

[木谷 96] 木谷 強 高木 徹 木原 誠 関根 道隆, "フルテキストと抽出キーワードを利用した情報検索", 情処研報 *96-NL-115-18*, 1996

[松本 96] 松本 裕治 他, 「茶筌 `Version 1.0b7`」, *http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html*, 1996

[BM 96] 情報検索システム評価用データベース構築ワーキンググループ, "情報検索システム評価用ベンチマーク Ver.1.0 解説書", 1996

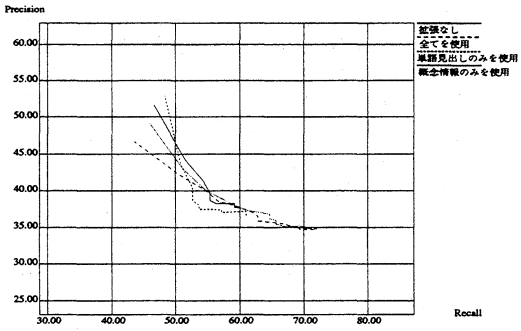

図 3: 「同等概念からの拡張」の実験結果

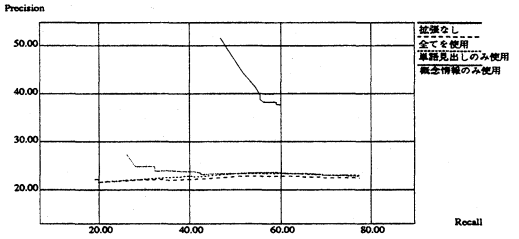

図 4: 「概念記述辞書からの拡張」の実験結果

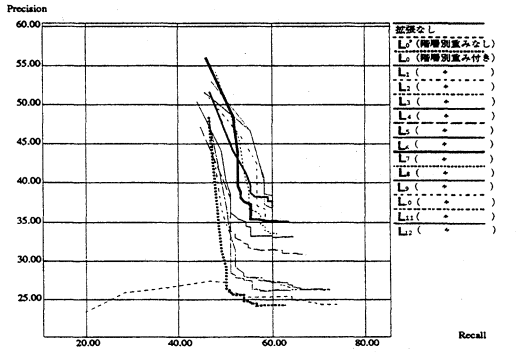

図 5: 「上位概念からの拡張」の閾値実験結果