情 報 学 基 礎 52－1
（1998. 11. 30）

シソーラスを用いた検索式拡張の評価

栗山和子
学術情報センター 研究開発部
kuriyama@rd.nacsis.ac.jp

概要. 本研究の目的は、シソーラスを用いた検索式の拡張を効率化することである。シソーラスや同義語・類義語辞書などの語彙的ツールを用いた検索式の拡張についての研究は、既に多数発表されている。一般的に、英文テキストに対して語彙的ツールを用いた検索式拡張では、自動的な検索式拡張は拡張なしの検索式よりも検索精度が下がるということが報告されている。

語彙的ツールを用いた検索式拡張による検索の効率化の失敗の原因としては、(1) 同義語の個数が十分ではない、(2) 語彙的ツールの見出し語の不足、(3) 検索式に追加された同義語の中に検索要求中に含まれる検索語の同義語として不適切なものが多く含まれる、の 3 点が考えられる。

本研究では、主に (3) の問題を解決するため、同義語による検索式の拡張に検索要求の context を反映させることを考える。検索式をデータベース中の文書の context に適した方法で拡張するために、検索要求中の単語・フレーズとシソーラスによって追加された同義語の共起頻度を利用する。具体的には、原検索式中に含まれる単語と、別の単語の同義語との共起頻度を調べ、共起頻度がある値以上である同義語を適切なものとして選択し、それを用いて検索式の拡張を行なう。

Query Expansion using Thesauri

Kazuko Kuriyama
National Center for Science Information Systems

Abstract. The purpose of this research is to improve retrieval performance using thesauri or other lexical tools for query expansion. Many studies on query expansion using lexical tools have been published. In general, it is reported that automatic query expansion using lexical tools degrade retrieval performance in English document retrieval. I should point out two problems as the cause of it, (1)the number of synonyms is insufficient; (2)lack of entry terms in the lexical tools; (3)the set of synonyms added to the query contains many inappropriate terms.

In this paper, I would like to consider how to reflect context of search request on query expansion by addition of synonyms to solve the second problem. For selecting appropriate words from the set of synonyms to expand query, co-occurrence relationship between the original terms and synonyms of the other terms is utilized.

1 はじめに

最近では、計算機上である程度容易に使用することができる日本語の辞書的資源が多数作られている。それらの語彙的ツールは情報検索にも使用可能であるが、本来のシソーラスあるいは類義語辞書のように厳密な階層構造や単語間の関係がうまく表現できていないものも多く、自動的な検索式拡張に利用するためには多少の工夫が必要である。

本研究では、語彙的ツールを用いた日本語の検索式拡張についての既存の研究を調べ、それらで述べられている問題点のいくつかを解決するための一つの方法として、シソーラスを用いて得られた同義語・類義語を検索要求中の単語との共起頻度によって選択する方法を提案する。

2 シソーラスによる検索式拡張

シソーラスや同義語·類義語辞書などの語彙的ツールを用いた検索式の拡張についての研究は、既に多数発表されている。

一般的に、英文テキストに対して語彙的ツールを用いた検索式拡張では、自動的な検索式拡張は拡張なしの検索式よりも検索精度が下がると言われている [15]。佐藤ほか [13] は、英語の大規模データベースに対して、既存のシソーラスを用いる代わりに、データベース中の文書から検索要求中の単語と共起する固有名詞を収集し、共起頻度の高い単語を用いて検索式を拡張しているが、拡張なしの場合よりも 10%程度の検索精度の改善が見られたと述べている。

また、同著者ら [11] は、同じデータベースに対して、英語のシソーラス WordNet を用いて検索式の拡張を行なっているが、検索式によっては検索精度が向上するものの、平均では拡張しない場合と同等であると述べている。同論文では、その原因として (1) シソーラス中の固有名詞や専門用語が十分でないこと、(2) 単語の多義性のために無関係な単語が多数展開されてしまうこと、を挙げている。

その他の語彙的ツールを用いた方法としては、EDR 電子化辞書を用いた方法 [10]、多様分類情報を用いた方法 [14]、類似性シソーラスを用いた方法 [12]、多言語キーワード・クラスタを用いた方法 [6] がある。

[10] では、検索要求中の語との概念関係（上位、下位、同等概念）を用いて検索式を拡張しているが、拡張なしに比べて検索精度の向上は少なかった。しかし、一般的なシソーラスには記述されていない名詞・動詞間の関係による動詞による拡張も行なっている。[14] では、ある観点での下での類義語の集合と類義語集合を代表する語（代表語）からなる多様分類情報を用いて、ある観点での類義語による拡張を行なっている。観点によるリンクである多様分類情報を提示することによって、検索者と検索システムとのオントロジーの統一を行なうことが可能になる。

[6] は、学術情報センターの学会発表データベースの著者キーワードから作成した多言語キーワード·クラスタを用いて言語横断検索 (CLIR) を行なっているが、日本語の単言語検索 (J–J task) においては、21 件の CLIR の検索課題の平均で 12.3%～14.0%検索性能が向上している。

シソーラスや概念辞書を用いた展開は検索要求中の一つ一つの単語に対して行なわれるため、検索要求全体が表わしている概念を拡張しているとは言えない。[12] ではテストコレクション BMIR-J1[3] を用いて類似性シソーラスを構築し、検索要求全体を表わす検索要求概念ベクトルに類似した語を検索語として追加している。この方法でも検索効率の改善は見られないが、テストコレクション中に特有の意味や概念を入れた検索がなされている。

上記の論文中でも述べられているが、語彙的ツールを用いた検索式拡張による検索性能の低下の原因としては、(1) 同義語の個数が十分ではない、(2) 語彙的ツールの見出し語の不足、(3) 検索式に追加された同義語の中に検索要求中に含まれる検索語の同義語として不適切なものが多く含

まれる、の3点が考えられる。本研究では、主に(3)の問題を解決するため、同義語による検索式の拡張に検索要求のcontextを反映させることを考える。

3 共出現を利用した同義語の選択

検索式をデータベース中の文書のcontextに適した方法で拡張するため、検索要求中の単語・フレーズとシソーラスによって追加された同義語の共起頻度を利用する。具体的には、原検索式中に含まれる単語と、別の単語の同義語との共出現を調べ、共起頻度がある閾値以上である同義語を適切なものとして選択し、それを用いて検索式を拡張する。

4 実験的評価

本研究での手法の振舞いを詳しく調べてみるために、1つの検索要求を事例として用いて、実験的に検索結果の評価を行なう。

4.1 テストコレクション NTCIR

評価には、学術振興会の未来開拓学術研究推進事業研究プロジェクト「高度分散情報資源活用のためのユービキタス情報システムに関する研究」の一環として、学術情報センター研究開発部が行なっている「情報検索システム評価用大規模テストコレクション構築プロジェクト」[5]において、同センターの学会発表データベースを用いて構築されたテストコレクション1(NTCIR:NACSIS Test Collection for IR Systems1)(テスト版)の一部を使用する。

NTCIR全体は日本語と英語の二言語のデータを用いて作成されているが、本研究で使用するシソーラス(後述)に含まれている同義語は日本語のみであるため、評価にはNTCIRの日本語部分のみ(日本語のタイトル、抄録、キーワード、会議名)を用いる。NTCIRの正解判定には、A(正解), B(部分的正解)の2段階があるが、ここではAとBの両方を正解とした。なお、検索については検索要求文だけを用い、より詳しい検索要求説明は使用しなかった。

使用した検索課題を以下に示す。

```
〈検索課題〉

〈タイトル〉
機械翻訳の評価
〈/タイトル〉

〈検索要求〉
機械翻訳における構造処理能力の評価
〈/検索要求〉

〈検索要求説明〉
機械翻訳以外の文献は除外する。
〈/検索要求説明〉
〈概念〉
a. 機械翻訳
b. 翻訳システム(aの上位)
c. 構造処理能力
〈/概念〉

〈分野〉
1. 電子・情報・制御
〈/分野〉

〈/検索課題〉
```

4.2 検索語の抽出

検索要求文の単語への分割には、日本語形態素解析システム Chasen Ver1.5 [7] を用いた。検索要求文から切り出した単語・フレーズからあらがじめ定義したストップ・フレーズを削除し、「名詞」、名詞の前に来る「形容詞」、「形容動詞」および「未定義語」を検索語とした。今回の検索実験では、検索語としては単語だけを使用し、フレーズは同義語との共起頻度を数えるときのみ使用した。

上記の検索課題の検索要求文から抽出した単語・フレーズは、{機械翻訳, 構造, 処理, 能力, 評価, 構造処理能力}である。

4.3 検索語の展開

検索語の同義語・類義語展開には、Express Finder/シソーラス辞書[9]、および、相澤ら[2],[6]による多言語キーワード・クラスタを用いた。

Express Finder/シソーラス辞書（以下、NTTシソーラス）は、NTT-ATが提供する「システムシソーラス辞書」と、ユーザが内容を定義する「ユーザーシソーラス辞書」から成る。「システムシソーラス辞書」は、一般的な文書、新聞、雑誌などで使用される基本普通名詞と一般語から成る「基本用語シソーラス辞書」と、既存の専門用語辞典等から収集した専門用語から成る14ジャンルの専門分野別の「専門用語シソーラス辞書」の合計15種類のシソーラス辞書から構成されている。本研究では、「システムシソーラス辞書」のうち、「基本用語シソーラス辞書」、および、「企業名」、「機関・団体・学校名」、「コンピュータ・情報・通信用語」、「医学・薬学用語」の4つのジャンルの「専門用語シソーラス辞書」の合わせて5種類のシソーラス辞書を使用する。

多言語キーワード・クラスタは、学術情報センターの学会発表データベース中の文書において、著者が付与したキーワードの中から英語キーワードと日本語キーワードのペアをとり、日本語キーワードと英語キーワードが一致したもののうちの一部を抽出し、クラスタリングを行なったものである。ここでは、[6]で述べられている**K3**と**K3–2**を使用する。

K3は同文献で使用されている中で最も大きなクラスタであり、数多くの単語を含んでいるため、同義語の網羅性という点から同文献の他のクラスタに比べて検索精度の平均の向上に最も効果があると報告されているが、同義語として関連の低い一般的な語も多く含まれてしまうという欠点がある。**K3–2**は**K3**からクラスタのグラフの形状を用いて関連の低い語を排除しているため、一般的な語は少なくなってしまい、検索精度の平均としては効果が出ていないが、上位の部分に対して有効である。本研究では、専門用語の網羅性と一般性という点から、この性質の異なる2つのクラスタを使用する。

NTTシソーラスの用語は、既存の文書や専門用語辞典などから確立された専門用語を収集されているため、新聞や一般雑誌等での各分野での検索には有効であると考えられるが、辞書等に収集されていない最新の専門用語や訳語がないためそのまま用いられることがある専門用語の原綴が頻繁に用いられる学術文献の検索ではどの程度有効であるかは未知である。一方、多言語キーワード・クラスタは、学会発表論文について著者が付与したキーワードを用いて作成されているため、最新の専門用語や原綴などを多く含んでいるが、キーワードとして付与されるような高度な専門用語ではない一般的な語をほとんど含んでいない。本研究では一般語の網羅性と専門用語の最新性という面からNTTシソーラスと多言語キーワード・クラスタが補い合えるのではないかと考え、両方を使用することにした。

NTTシソーラスと多言語キーワード・クラスタを用いて展開された同義語には単語以外にフレーズも含まれるが、ここでは同義語については単語とフレーズを区別しない。

上記の検索語に対してNTTシソーラスと多言語キーワード・クラスタ**K3**、**K3–2**を用いると、以下のような同義語・類義語が展開された。区別のために、それぞれの語について、NTTシソーラスをL_{th}、クラスタ**K3**をL_{k3}、**K3–2**をL_{k32}とする。

機械翻訳 → L_{th} = { mechanical translation , 自動翻訳},

L_{k3} = { 自動翻訳, 変換システム, 翻訳システム, 機械翻役, 機会翻訳, 機能語表現, machine translation , machine trauslation , machine traslation , machine franslation , machine tromslation , translation system , machine trans lation , mcchine translation , maching translation , automatic traslation , machire translation , mchine translation , machine tramslation , machive translation , machine transtation , machine , machine traus lation , madrine translation , machine transraltion , machichin transtation , transformation system , machine

translating , machine translalion , machin translation },

L_{k32} = { 自動翻訳, machine translation , machine trauslation , machine tromslation , machine traslation , machine franslation }

構造 → L_{th} = { configuration , constitution , mechanism , structure , texture , しくみ, ストラクチャ, ストラクチャー, ストラクチュア, メカ, メカニズム, 機構, 構成, 仕組, 仕組み, 組み立て, 組織, 組成, 組立, 組立て},

L_{k3} = { 構造化, cad データ, 構造データ, 情報構造化, 情報の構造化, 構造化データ, 格納庫, 階層構造化, 内容把握, 構造体, 構成, pdb , structure , structuring , cad data , structured data , information structuring , structuralization of information , structured , information structurization , description style , structureal data , structurization , structur , information orgarization },

L_{k32} = { structure }

処理 → L_{th} = { DP , data processing , film processing , information processing , processing , treatment , ＤＰ, データ処理, ハンドリング, フィルム処理, プロセス, 扱い, 現像処理, 治療, 取り扱, 取り扱い, 取り計らい, 取扱, 取扱い, 取計い, 取計らい, 処置, 処理過程, 情報処理, 措置, 操作, 療法},

L_{k3} = { },

L_{k32} = { }

能力 → L_{th} = { ability , capacity , competence , 伎倆, 器量, 技術, 技能, 技量, 技倆, 才幹, 才能, 実力, 手並, 手並み, 手腕, 受容能力, 体力, 力量, 腕まえ, 腕前 },

L_{k3} = { ability },

L_{k32} = { }

評価 → L_{th} = { assessment , evaluation , アセスメント, 格付, 格付け, 査定, 値踏, 値踏み, 評定, 品定, 品評 },

L_{k3} = { 性能評価, 性能, 性能解析, 性能予測, 評価システム, 関係代数, パフォマンス, 伝送効率, 見積り, 演奏, 評価実験, 能力評価, 性能設計, 推定, 評価式, 定量評価, 性能検証, 性能分析, 評定, 近似比解析, 性能見積り, 推定問題, パフォマンス測定, 性能診断, シミュレション性能評価, ap 性能, 視聴率, システム解析, 実用評価, 見積, 見積り手法, 演奏傾向, 評価問題, 見積り技術, ボクセル空間, 層状否定, performance evaluation , evaluation , performance , performance analysis , estimation , performance estimation , relational algebra , peformance evaluation , evaluation system , estimate , rating , evalution , performability , valuation , performance design , performance prediction , perfomance evaluation , performance evalvation , experimental evaluation , evaluation of capability , performince evaluation , evaluation of system , performance verification , performance eraluation , prediction of performance , optimization of performance , performance evaluatin , transmit efficlency , performanca evaluation , trnsmission efficiency , performance evaluatia , perfornence evdludtion , pertormance evaluation , performance evalnation , peformance analysis , estimating performance , perfor mance eualuation , performanceevaluation , evaluartion , preformance evaluation , perfromance evaluation , performance evalution , performance estimate , stratified negation , evalualion , performance test , pertormance analysis , relational algebla , evaluation formulae , evaluation of performance , performance analisys , persormance analisys , pertormance evaluatean , protofype design , performance evabuation , design of performance , pertornance evaluation , pefo'rmance , quantitative estimation , appraisal , performance evaluations , performance eialuation , transmission efficiency },

L_{k32} = { 性能評価, 性能解析, 見積り, 評価実験, 推定, 能力評価, evaluation , performance evaluation , performance analysis , estimation , peformance evaluation , evalution , estimate , experimental evaluation , valuation , perfomance evaluation , performability , performance evalvation }

4.4 同義語の選択

展開された同義語・類義語は、元の検索語（単語・フレーズ）以外の単語との文書中での共出現の回数を数える。単語と単語の距離は、日本語の場合には1文字を1単位、英語の場合には1 word

を1単位とした。元の検索語と同義語・類義語との距離が100単位以内である場合に共出現しているとし、共出現を何回含んでも1文書を1回と数えた。

使用した文書は学術情報センターの学会発表データベースに含まれる約30万件の学会発表データのうち、日本語のタイトル、発表者、抄録、キーワード、会議名の部分である。

同義語を選択するための閾値としては、共出現する元の検索語の個数とその回数を用いた。本論文では、**N1F1**:どれか一つの検索語と1回以上、**N1F5**:どれか1つの検索語と5回以上、**N3F1**:どれか3つの検索語と1回以上、**N3F5**:どれか3つの検索語と5回以上、という4種類の値に対して、その基準を満たした同義語・類義語を検索式に追加する語として選択した。

基準**N3F1**によって選択されたものは**4.3**で挙げた同義語のうちの下線部のものである。(**4.3**参照)

4.5 実験

検索対象は学会発表データベースの日本語のタイトル、発表者、抄録、キーワード、会議名である。**4.1**で示した検索要求文から抽出した検索語に対して**4.4**で述べた4種類の閾値**N1F1**,**N1F5**,**N3F1**,**N3F5**によってそれぞれ選択した同義語を追加して検索式を作成し、検索を行なった。拡張した検索式の中では元の検索語と追加した同義語を一つのグループとして重み付けを行ない、検索式を作成した。すなわち、元の検索語と同義語のいずれかがデータベース中に存在しても同じ一つの語が存在する場合と同じ重みになるようにした。これによって、データベース中に小数しか含まれていない語が同義語として追加されても、その語の出現する文献の重みだけが大きくなってしまうことがなくなる。

検索式は、検索要求文から抽出した検索語だけを用いた「拡張なし検索式」($E0$)、**N1F1**,**N1F5**,**N3F1**,**N3F5**のそれぞれについて「全ての同義語・類義語で拡張した検索式」($E11$, $E15$, $E31$, $E35$)、「NTTシソーラスから選択した同義語で拡張した検索式」($E11_{th}$, $E15_{th}$, $E31_{th}$, $E35_{th}$)、「クラスタ**K3**から選択した同義語で拡張した検索式」($E11_{k3}$, $E15_{k3}$, $E31_{k3}$, $E35_{k3}$)、「クラスタ**K32**から選択した同義語で拡張した検索式」($E11_{k32}$, $E15_{k32}$, $E31_{k32}$, $E35_{k32}$)の17種類をを作成した。検索システムとしてはOpenText6 (OpenText社、カナダ)を用い、検索結果のランク付けには、同システムのRankMode "Relevance1"を用いた。正解判定には、NTCIR (テスト版)の正解文書リストを使用した。

評価の計算には、TREC[4]使用されている評価プログラムを用いた。検索結果は上位1000件とし、0.1刻みのRecallに対するPrecisionを用いて評価した。TRECの評価プログラムでは、平均は11 pointの平均ではなく、補間を行なわない全ての適合文献の精度の平均である。RecallとPrecisionの定義は以下の式の通りである。

$$\text{Recall} = \frac{\text{全検索結果のうちの正解文書数}}{\text{全正解文書数}} \times 100 \ ,$$

$$\text{Precision} = \frac{\text{全検索結果のうちの正解文書数}}{\text{全検索結果文書数 (1000件)}} \times 100$$

N1F1,**N1F5**,**N3F1**,**N3F5**の評価結果をそれぞれ表1, 表2, 表3, 表4に示す。表中のrclはRecall、aveは平均、%impは%improved(ベースラインからの向上率)を表わす。

表1. 評価結果 (Recall/Precision)

rcl	$E0$	$E11$	$E11_{th}$	$E11_{k3}$	$E11_{k32}$
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
0.1	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	0.2857	0.7500	0.8000	0.2857	0.2222
0.3	0.2857	0.5714	0.8000	0.2847	0.2222
0.4	0.1875	0.1714	0.1333	0.1304	0.1463
0.5	0.1707	0.1522	0.1250	0.0833	0.1176
0.6	0.1231	0.0419	0.0988	0.0734	0.1176
0.7	0.0602	0.0330	0.0641	0.0446	0.0440
0.8	0.0550	0.0185	0.0243	0.0420	0.0440
0.9	0.0327	0.0000	0.0000	0.0259	0.0253
1.0	0.0000	0.0000	0.0000	0.0000	0.0000
ave	0.2663	0.3069	0.3299	0.2425	0.2431
%imp	0	15.2	23.8	−8.9	−8.7

表 2. 評価結果 (Recall/Precision)

rcl	$E0$	$E15$	$E15_{th}$	$E15_{k3}$	$E15_{k32}$
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
0.1	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	0.2857	0.7500	0.8000	0.2857	0.2222
0.3	0.2857	0.5714	0.8000	0.2847	0.2222
0.4	0.1875	0.1714	0.1333	0.1304	0.1463
0.5	0.1707	0.1522	0.1250	0.0833	0.1045
0.6	0.1231	0.0419	0.1000	0.0741	0.0930
0.7	0.0602	0.0331	0.0641	0.0446	0.0442
0.8	0.0550	0.0185	0.0244	0.0420	0.0442
0.9	0.0327	0.0000	0.0000	0.0260	0.0253
1.0	0.0000	0.0000	0.0000	0.0000	0.0000
ave	0.2663	0.3070	0.3300	0.2426	0.2412
%imp	0	15.2	23.9	−8.9	−9.4

表 3. 評価結果 (Recall/Precision)

rcl	$E0$	$E31$	$E31_{th}$	$E31_{k3}$	$E31_{k32}$
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
0.1	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	0.2857	0.7500	0.8000	0.2857	0.2222
0.3	0.2857	0.5714	0.8000	0.2847	0.2222
0.4	0.1875	0.1714	0.1333	0.1304	0.1463
0.5	0.1707	0.1522	0.1250	0.0833	0.1045
0.6	0.1231	0.0421	0.0988	0.0741	0.0930
0.7	0.0602	0.0333	0.0641	0.0446	0.0442
0.8	0.0550	0.0186	0.0244	0.0420	0.0442
0.9	0.0327	0.0000	0.0000	0.0259	0.0253
1.0	0.0000	0.0000	0.0000	0.0000	0.0000
ave	0.2663	0.3070	0.3299	0.2426	0.2412
%imp	0	15.2	23.8	−8.9	−9.4

表 4. 評価結果 (Recall/Precision)

rcl	$E0$	$E35$	$E35_{th}$	$E35_{k3}$	$E35_{k32}$
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
0.1	1.0000	1.0000	1.0000	1.0000	1.0000
0.2	0.2857	0.7500	0.8000	0.2857	0.2222
0.3	0.2857	0.5714	0.8000	0.2847	0.2222
0.4	0.1875	0.1765	0.1429	0.1304	0.1463
0.5	0.1707	0.1556	0.1321	0.0843	0.1061
0.6	0.1231	0.0442	0.1053	0.0755	0.0941
0.7	0.0602	0.0352	0.0671	0.0455	0.0449
0.8	0.0550	0.0198	0.0253	0.0426	0.0449
0.9	0.0327	0.0000	0.0000	0.0270	0.0256
1.0	0.0000	0.0000	0.0000	0.0000	0.0000
ave	0.2663	0.3082	0.3339	0.2430	0.2416
%imp	0	15.6	25.3	−8.7	−9.2

4.6 考察

シソーラスおよびキーワード・クラスタを用いて展開された同義語・類義語を共起頻度を利用して取捨選択し、拡張した検索式を用いた検索の事例を示した。この検索課題の例では、拡張なしの場合に比べて全ての同義語を追加した検索式とシソーラスから得られた同義語を追加した検索式については平均の検索精度が向上しているが、キーワード・クラスタから得られた同義語を全て追加した検索式は少し低下している。この検索課題の例では、クラスタからの同義語に関しては、適切でない語を排除するのと同時に適切な語まで排除してしまっているからではないかと思われる。

1例だけの事例分析ではあるが、本研究の目的とした、多過ぎる同義語から適切な同義語を選択するという点からは、選択の効果が見られた。

5 おわりに

情報検索ではいろいろな検索要求に対処できることが重要である。今回の事例分析で、シソーラスから展開された同義語を共出現を利用して選択して検索式を拡張する手法の振舞いが把握できた。今後は、テストコレクション全体を用いて評価を行ないたい。その際、本手法で考慮すべきこととしては、以下のようなことが考えられる。

- 検索要求から切り出された単語には検索要求の主要概念を表わしている語とそうでない語がある。
- データベース中での共起頻度が高いほど重要な単語であるとは言えない。

謝辞

NTT アドバンステクノロジには NTT シソーラスの使用をご承諾いただきました。学術情報センター 相澤彰子助教授には、多言語キーワード・クラスタの使用をご承諾いただきました。深く感謝致します。また、以下の方からは、ご助言・ご助力をいただきました。感謝致します。学術情報センター 安達淳教授、大山敬三教授、神門典子助教授、野末俊比古助手。

参考文献

[1] 赤峰亨; 佐藤研治; 奥村明俊. 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. 情報処理学会第 52 回全国大会論文集, pp.4-201–4-202, 1996.

[2] Aizawa, A.; Kageura, K. "An approach to the automatic generation of multilingual keyword clusters". COMPUTERM'98, Canada, 1998.

[3] 福島ほか. "日本語情報検索システム評価用コレクション BMIR-J1", 自然言語シンポジウム, 大規模資源と自然言語処理, 電子情報処理学会, 言語理解とコミュニケーション研究会, 対話システム研究会.

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

[5] Kando, N. et.al. NTCIR: "NACSIS Test Collection Project". [Poster] the 20th Annual Collection of BCS–IRSG, France, 1998.

[6] Kando, N.; Aizawa, A. "Cross-Lingual Information Retrieval using Automaticaly Generated Multilingual Keyword Clusters". IRAL'98, Singapore, 1998.

[7] 松本裕治ほか. 日本語形態素解析システム『茶筌』version 1.5 使用説明書, 奈良先端科学技術大学院大学, http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html, 1997.

[8] Matsumura, A.; Takasu, A.; Adachi, J. "Information Retrieval using Structured Index for Japanese Text". IRAL'98, Singapore, 1998.

[9] NTT アドバンステクノロジ株式会社. シソーラス辞書展開キット デベロッパーズマニュアル 第 3 版, 74p., 1998.

[10] 太田千晶; 奥村学. EDR 電子化辞書を用いたクエリー拡張による検索支援. 言語処理学会第 3 回年次大会発表論文集, 京都大学, pp.373–376, 1997.

[11] 赤峰亨; 佐藤研治; 奥村明俊. 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. 情報処理学会第 52 回全国大会論文集, pp.4-201–4-202, 1996.

[12] 斉藤公一ほか. 概念に基づく検索要求文の拡張. 97–FI–47–10, pp.127–135, 1997.

[13] 佐藤研治; 赤峰亨; 奥村明俊. 単語共起によるクエリー展開を用いた大規模テキスト. 情報処理学会第 52 回全国大会論文集, pp.4-199–4-200, 1996.

[14] 下畑光夫; 坂本仁. 多様分類情報による検索語拡張, 96–FI–43–11, pp.135–140, 1996.

[15] Voorhees, E.M. "Query Expansion using Lexical-Semantic Relations". In Proc. of SIGIR'94, pp.62–69, 1994.