特集論文

選択的メタサーチエンジンにおけるシソーラスを用いたサーチエンジン選択手法の提案

A Method for Search Engine Selection using Thesaurus for Selective Meta-Search Engine

後藤 将志 Shoji Goto
名古屋工業大学 知能情報システム学科
Departent of Intelligence and Computer Science, Nagoya Institute of Technology
shoji@ics.nitech.ac.jp, http://www-toralab.ics.nitech.ac.jp/~shoji/

大囿 忠親 Tadachika Ozono
（同　上）
ozono@ics.nitech.ac.jp, http://www-toralab.ics.nitech.ac.jp/~ozono/

新谷 虎松 Toramatsu Shintani
（同　上）
tora@ics.nitech.ac.jp, http://www-toralab.ics.nitech.ac.jp/professor/tora.html

keywords: search engine selection, meta-search engine, thesaurus, distributed information retrieval

Summary

In this paper, we propose a new method for selecting search engines on WWW for selective meta-search engine. In selective meta-search engine, a method is needed that would enable selecting appropriate search engines for users' queries. Most existing methods use statistical data such as document frequency. These methods may select inappropriate search engines if a query contains polysemous words. In this paper, we describe an search engine selection method based on thesaurus. In our method, a thesaurus is constructed from documents in a search engine and is used as a source description of the search engine. The form of a particular thesaurus depends on the documents used for its construction. Our method enables search engine selection by considering relationship between terms and overcomes the problems caused by polysemous words. Further, our method does not have a centralized broker maintaining data, such as document frequency for all search engines. As a result, it is easy to add a new search engine, and meta-search engines become more scalable with our method compared to other existing methods.

1. はじめに

近年，WWW で利用できる情報が膨大になるにつれて，ユーザにとって必要な情報を見つけることは困難になっている．WWW 上で必要な情報を探すとき，ユーザはサーチエンジンを利用することが多い．しかし，WWW の情報量の増加の速さに対し，現在の集中型の汎用サーチエンジンが追い付くことができず，各サーチエンジンが持つ WWW 上の情報のカバー率は年々低下し続けている [Lawrence 99]．この問題に対する一つのアイデアとして，既存の複数のサーチエンジンを利用するメタサーチエンジンが挙げられる．メタサーチエンジンの中でも特に，ユーザの検索要求に適したサーチエンジンを選択する，選択的メタサーチエンジンの研究が注目されている [山田 01]．検索要求に対し適切なサーチエンジンを選択するには，各サーチエンジンが持つ情報の特徴を獲得する必要がある．サーチエンジンが持つ情報の特徴を表現したものを Source Description と呼ぶ．これまでに提案されてきたサーチエンジンの選択手法では，語の出現頻度など，語単独で得られる情報に基づいて Source Description を構築する．しかし，このような手法では，多義語や低出現頻度の語が原因となり不適切なサーチエンジンの選択を行う可能性がある．

本論文では，選択的メタサーチエンジンにおいて，シソーラスを Source Description として利用するサーチエンジン選択手法を提案する．本提案手法では，語の出現頻度以外に，語同士の関連の強さを利用して選択の精度を向上させる．メタサーチエンジンでは，各サーチエンジンの検索質問の形式や，検索結果の提示方法の異種性を吸収するための仕組みを提供する Wrapper と呼ばれるモジュールが必要となる．本提案手法で Source Description として利用するシソーラスは，各サーチエンジン毎に独立に構築される．そのため，本提案手法に基づいて構築されたメタサーチエンジンでは，新しいサーチエンジンを利用するための追加作業が容易になり，スケーラビリティの向上を図ることができる．

本論文ではまず，既存のメタサーチエンジンや，その関連技術である分散情報検索と情報源選択手法について述べる．次に，提案手法で Source Description として用いるシソーラスの自動構築手法について述べる．その後，構築されたシソーラスを用いたサーチエンジン選択アルゴリズムについて説明する．そして，提案手法の有効性を確認するために行った評価実験と，それに対する考察を述べ，最後にまとめる．

2. メタサーチエンジンと分散情報検索

メタサーチエンジンでは，複数のサーチエンジンを利用することにより，広い範囲から収集された情報を利用できる．そのため，検索の再現率の向上が期待できる．しかし，検索質問に関係が薄いサーチエンジンを利用した場合，適合率が低下することがある．そのため，検索質問に応じて，サーチエンジンを選択的に利用する選択的サーチエンジンの研究が活発に行われている [Howe 97, Lawrence 98, Fan 99]．また，近年では分散情報検索に関する研究が盛んに行われている．分散情報検索とは，複数の情報源を利用した情報検索 [Callan 00] である．サーチエンジンを一つの情報源とみなすと，メタサーチエンジンは分散情報検索のアプリケーションの 1 つといえる．分散情報検索において，情報源が情報の種類に基づいて適切に分割されている状況では，ユーザの検索要求に対し，適切な情報源を選ぶことにより，集中型の検索システムより良い結果が得られると報告されている [Powell 00]．そのため，分散情報検索の研究ではさまざまな情報源選択手法が提案されている [Hawking 99].

これまでに提案されている情報源選択手法では，主に語の出現頻度のような，語を単独で扱った場合に得られる情報を利用して Source Description を構築する．例えば，SavvySearch [Howe 97] では，検索質問中の語と，その語に対する検索結果の数をサーチエンジンの Source Description として利用する．また，分散情報検索の情報源選択手法の 1 つである CORI [Callan 95] では，各情報源に含まれる文書集合中で，ある語がいくつの文書のに出現するかを表す document frequency と，いくつの情報源であの語が出現しているかを表す source frequency を利用している．適切なサーチエンジンを選択する際に，source frequency のような，全情報源に関する情報（以下，グローバル情報と呼ぶ）を利用するには，全情報源に含まれる語すべての情報を管理するデータベースが必要となる．そのため，従来手法では，Source Description 再構築時にグローバル情報の更新が必要となる．また，新たな情報源を利用するときもグローバル情報の変更が必要となり，追加作業が複雑になる．また，語の出現頻度に基づく手法では，基本的に高出現頻度の語で情報源を特徴づけする．しかし，低出現頻度の語がその情報源が持つ情報の特徴を表現することも多い．また，複数の意味を持つ多義語の意味は，用いられている場面の背景にあるドメインによって異なることが多い．しかし，語の出現頻度のみでは，語義を同定することは困難である．そのため，Source Description として語の出現頻度のみを利用した手法では，低出現頻度の語や多義語が原因で誤った情報源を選択する可能性がある．

本研究では語同士の関連の強さに着目した．そして，語の出現頻度に加え，語同士の関連の強さをサーチエンジンの特徴として利用するサーチエンジン選択手法を提案する．語同士の関連を表現した辞書をシソーラスと呼ぶ．本提案手法では，シソーラスにおける語同士の関係を用いることにより，低出現頻度の語にも特徴として利用することができる．また，語の意味を他の語との関係によって区別することにより，多義語の問題も解消できると考えられる．このような語同士の関係はベクトル空間モデルでは表現できない [Park 95].

また，分散情報検索において，検索の適合率を上げるには，各データベースに概念的に一致した情報が含まれている必要がある．従来のメタサーチエンジンでは，Altavista[*1] や Google[*2] のような汎用サーチエンジンを用いている．汎用サーチエンジンでは，情報の種類に関係なくロボットと呼ばれるソフトウェアによって Web ページが自動収集される．そのため，汎用サーチエンジンは含まれる内容に概念的な統一性がない．汎用サーチエンジンをメタサーチエンジンで利用した場合，検索対象となる情報は増加するため，再現率は向上する可能性がある．しかし，適合率の向上は期待できない．そこで，本研究では専門サーチエンジンに着目した．専門サーチエンジンとは検索対象を特定のドメインに限定したサーチエンジン [北村 98] であり，年々その数は増加している．専門サーチエンジンでは，含まれる内容に関して概念的な統一性がある．そのため，専門サーチエンジンを利用したメタサーチエンジンでは文献 [Powell 00] の報告により，集中型の従来のサーチエンジンより，精度の高い検索が期待できる．

3. シソーラスに基づくサーチエンジン選択

本手法では，Source Description としてシソーラスを利用する．語同士の関連の強さは，語が利用されている場面によって大きく異なる．同一のドメインでは，同一の語は同じ場面で利用されることが多いため，語同士の関係は類似していると考えられる．そのため，各サーチエンジンに含まれる情報から構築されたシソーラスにより，そのサーチエンジンのドメインを表現することができる．本提案手法では，サーチエンジンが検索結果として提示する URL の HTML 文書からシソーラスを自動構築する．専門サーチエンジンが検索結果として提示され

*1 http://www.altavista.com/
*2 http://www.google.com/

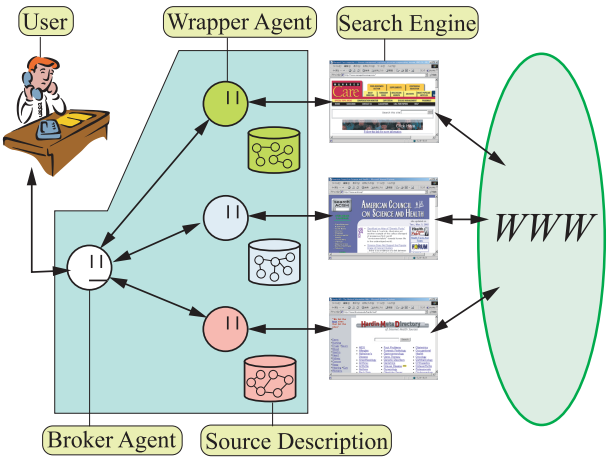

図 1 システム構成図

る URL の文書は，専門サーチエンジンが対象とするドメインに関連した文書である．そのため，それらの文書から自動構築されたシソーラスにはそのサーチエンジンの対象ドメインにおける語の関係が表現される．

本提案手法では，シソーラスを利用し，ユーザの検索質問とサーチエンジンが対象とするドメインの関連性を評価する．そして，関連の高いサーチエンジンのみを選択し，検索に利用することにより，検索の適合率を向上させる．本手法では，ユーザの検索要求に複数の語が含まれることを前提とする．そして，検索要求中の語は関連が強いという仮定に基づき，Source Description としてのシソーラスと検索要求の関連性を計算し，サーチエンジンの評価をする．この際，自動構築されたシソーラスに含まれる語は利用した情報によって偏る．本手法では，一般的な語彙不足を補うため，英語のシソーラスである WordNet [Miller 95] を利用する．

本提案手法では，Source Description が各サーチエンジン毎に独立に構築される．そのため，新たなサーチエンジンをメタサーチエンジンで利用するとき，他のサーチエンジンを変更すること無く追加することができ，スケーラビリティを向上させることができる．専門サーチエンジンをメタサーチエンジンで利用する場合，できる限り多くの専門サーチエンジンを利用することが望ましい．なぜなら，利用する専門サーチエンジンの数が増えれば，メタサーチエンジンではより幅広いドメインの情報を扱うことができ，メタサーチエンジンの利用場面が増加するためである．そのため，本提案手法の特性は，専門サーチエンジンを利用する上で重要になる．

3·1 システム構成

本提案手法に基づいて構築されるメタサーチエンジンの構成を 図 1 に示す．本システムでは，各機能をエージェント化することにより，システムの信頼性を向上させるとともに，スケーラビリティに優れたシステムとなる．本システムは大きく分けて 2 種類のエージェントが存在する．一つは Wrapper エージェント，もう一つが Broker エージェントである．本システムでは，1 つの Broker エージェントと複数の Wrapper エージェントが通信し，情報をやり取りすることにより，メタサーチエンジンとして動作する．

Wrapper エージェントは各サーチエンジンに一つずつ割り振られる．Wrapper エージェントの機能は大きく分けて二つある．一つめの機能は，検索におけるサーチエンジンの入力（検索質問）と出力（検索結果）の異種性の吸収である．一般に，検索質問や検索結果の提示形式はサーチエンジン毎に異なる．複数のサーチエンジンを利用するメタサーチエンジンでは，Wrapper と呼ばれるモジュールによってこの異種性を解消する．Wrapper エージェントはこのための機能を提供する．二つ目の機能は，本手法で Source Description として用いるシソーラスの構築と検索質問の評価である． Wrapper エージェントは Source Description としてシソーラスを定期的に作成する．検索質問を受け取ると，自身が構築したシソーラスを元に，サーチエンジンの評価値を計算する．これらを実現する具体的な手法は後に述べる．

Broker エージェントはユーザとのインタフェースになり，ユーザから検索質問を受け取り，検索結果を提示する．Broker エージェントは Wrapper エージェントの場所のみを把握しており，各 Wrapper エージェントが管理しているサーチエンジンの情報は持たない．ユーザから検索要求を受け取ると，Broker エージェントが把握しているすべての Wrapper エージェントに検索質問を送信する．そして，それぞれの Wrapper エージェントから返ってきた評価値を元に，適切な Wrapper エージェントを選択し，実際に検索を実行させる．そして，各 Wrapper エージェントから返ってきた結果を統合し，ユーザに検索結果として提示する．

本提案手法に基づいて実現されるメタサーチエンジンを利用したときの典型的な検索プロセスにおける，システムの流れを以下に示す．

(1) ユーザが検索質問を Broker エージェントに送信する．
(2) 検索質問を受け取った Broker エージェントは，すべての Wrapper エージェントに対し検索質問をブロードキャストする．
(3) 検索質問を受け取った Wrapper エージェントは，事前に構築したシソーラスを利用し，自分が担当するサーチエンジンと検索要求との関連度を計算する．そして，その値を Broker エージェントに送信する．
(4) Broker エージェントは，各 Wrapper エージェントから返ってきた値を元に上位 n 個のサーチエンジンを選択し，実際に検索を実行させる．
(5) 選ばれた Wrapper エージェントは実際にサーチエンジンで検索し，得られた結果を Broker エージェントに返す．

(6) 各 Wrapper エージェントから送られる検索結果を統合し，ユーザに提示する．

4. Source Description の構築

4·1 検索質問生成

本システムにおける Wrapper エージェントは，サーチエンジンが検索結果として提示した URL の文書からシソーラスを構築し，Source Description として利用する．検索結果を取得するには，検索質問を作成する必要がある．本節では，シソーラスを構築する前段階として必要となる検索質問の生成について述べる．

一般に，あるドメインには更に細分化されたサブドメインが存在する．専門サーチエンジンを用いて検索を行った場合でも，検索結果は検索質問により偏りが生じる．提案手法では，専門サーチエンジンが対象とするドメイン全般における語と語の関係を得ることが望まれる．そのため，偏りのない文書集合を取得し，シソーラスの構築に利用する必要がある．結果，サーチエンジンにとって特定性より網羅性が高い検索質問が必要となる．そこで，本論文では専門サーチエンジンが対象とするドメインにおいて出現頻度が高い語を検索質問として利用する手法を提案する．高出現頻度の語はサブドメインに関係なく，そのサーチエンジンが対象とするドメインのさまざまな Web ページに含まれている可能性が高い．一般に，出現頻度の高い語は特定性が低いため，特定の文書を見つけるには不適切である．しかし，多くの文書において数多く存在するため網羅性が高く，本手法には適していると考えられる．

本手法では，検索要求として利用する語を発見するために，サーチエンジンの検索インタフェース用の Web ページと，そこからリンクが張られている Web ページからキーワードを抽出する．ある語 t の重みは式 (1) の計算式で求める．

$$w_t = \sum_{d \in D} tf_{d_t} * \sum_{d \in D} f(t,d) \qquad (1)$$

ここで，d は 1 つの Web ページ，D は Web ページの集合を表す．tf_{d_t} は，Web ページ d で語 t が出現する数，$f(t,d)$ は語 t が Web ページ d に含まれている場合 1，含まれていない場合は 0 となる．式 (1) で得られる語の重みは，ある語が一つの Web ページに多く出現するほど，また，多くの Web ページに出現するほど語の網羅性が高いという仮定に基づいたものである．本手法では，語の重み w_t が大きな語を利用して検索を行い，検索結果を取得する．

4·2 シソーラスの構築

検索結果から獲得した Web ページに含まれる語を元にシソーラスを構築する．シソーラスの自動構築において，語の同義関係や上位 / 下位関係を自動的に判別することは大変困難である．そのため，自動構築されたシソーラスは，単純に語の関連を表現するにとどまっている．このため，自動構築されたシソーラスは，木構造ではなく，より複雑なグラフ構造を持つ．現在提案されている主なシソーラスの自動構築手法は共起関係に基づいている．これらの手法では，同一の場所に現れる語は関係が強いと仮定し，シソーラスを構築する．共起関係に基づくシソーラス構築では，共起した語の類似尺度によって，さまざまな手法が提案されている [Kim 98]．現在提案されている手法は大きく分けて二つに分類される．一つは語の出現頻度に重きをおいたもので，これには Jaccard 類似度や Dice 類似度などが含まれる．これら手法では，低い出現頻度の語に関して信頼に欠けるという data sparseness 問題が起こると指摘されている [Park 95]．もう一つが語同士の関係に重きをおいた手法である．本論文で提案するサーチエンジン選択手法では，サーチエンジンの評価尺度として，語の出現頻度と語同士の関連を利用する．そのため，今回は語同士の類似尺度として，語の関係に重きをおいた手法である 2 つの語の出現についての条件付き確率の平均値を利用する [Park 95]．この類似尺度は式 (2) で定義される．ただし，$P(x|y)$ は，語 y の出現ときに，語 x が出現する条件付き確率である．

$$\mathrm{similarity}(x,y) = \frac{P(x|y) + P(y|x)}{2} \qquad (2)$$

5. サーチエンジン評価アルゴリズム

本提案手法では，前節で述べたシソーラスを各サーチエンジン毎に作成する．そして，ユーザから与えられた検索質問を元に，各サーチエンジンについて検索要求との関連度を計算し，関連度の高い n 個のサーチエンジンを選択する．以下では検索要求とサーチエンジンの関連度を計算するためのアルゴリズムについて述べる．

本手法では，ユーザの検索要求に複数の語が含まれることを前提とする．そして，検索要求中の語は，同じドメインでよく用いられ，お互いに関連が強いという仮定を利用する．関連の強い語は，グラフ構造をしたシソーラス中では隣接しているか，近い距離で結ばれている．本アルゴリズムでは，検索要求中の二つの語のシソーラスにおける距離を利用して語同士の関係の強さを計算する．本提案手法では，検索質問中の語のすべての組み合わせで語同士の関係の強さを計算し，その平均値を検索質問とサーチエンジンの関連度とする．平均値が高いほどユーザの検索質問とサーチエンジンの関連が強いといえる．検索質問中の語が Source Description 中に無い場合，サーチエンジンの評価ができない．そこで，一般的な語彙に関するシソーラスである WordNet からその語の類似語を抜き出し，その語を代わりに利用し，語同士の関連の強さを計算する．図 2 に本アルゴリズムを示す．

Algorithm

$q=\{t_1,t_2,\cdots,t_i\}$: 検索質問
$T_A=(V_A,E_A)$: 自動構築されたシソーラス
$T_W=(V_W,E_W)$: WordNet
$V=\{T_1,T_2,\cdots,T_k\}$: 頂点集合
$E=\{e_1,e_2,\cdots,e_l\}$: 辺集合
$e_j=(T_x,T_y,s_j), 1\le j\le l, 1\le x,y\le k$: 辺

$Terms\leftarrow\phi$, $eval\leftarrow 0$,
α and β are given($0<\alpha<1, 0<\beta<1$)

function EvalSource(q:set):double
  foreach term t_i **in** query q **do**
    MapTerm($t_i, tf_{t_i}, 1.0$)
  end
  foreach $\langle t_1, tf_{t_1}, w_{t_1}\rangle, \langle t_2, tf_{t_2}, w_{t_2}\rangle \in Terms$ **and** $t_1\neq t_2$ **do**
    $eval\leftarrow eval+\mathrm{relation}(t_1,t_2)$
  end
  return $\dfrac{eval}{|Terms|C_2}$
end

procedure MapTerm(t:string, tf_t:int, w:double)
  if $w<\beta$ **then return**
  if a term t is in the source description **then do**
    $Terms\leftarrow Terms\cup\langle t, tf_t, w\rangle$
  else if a term t is in WordNet **then do**
    $Neighbors_t\leftarrow\{t_i|\,\mathrm{distance}_w(t,t_i)=1\}$
    foreach t_i **in** $Neighbors_t$ **do**
      MapTerm($t_i, tf_{t_i}, w*\alpha$)
    end
  end if
end

図 2 サーチエンジン評価アルゴリズム

ユーザの検索要求を q とすると，q は語の集合であり，$q=\{t_1,t_2,\cdots,t_i\}$ と表現することができる．また，Source Description として用いられるシソーラス T_A は，語を頂点に，語と語の関係を辺とすることにより，グラフで表現できる．同様に WordNet T_W もまたグラフで表現できる．あるシソーラスを T とすれば，T はグラフ中の頂点の集合 V と，頂点間を結ぶ辺の集合 E で表現でき，$T=\langle V,E\rangle$ と書くことができる．シソーラスにおいて，V はシソーラス中の語の集合と考えることができ，$V=\{T_1,T_2,\cdots T_n\}$ と表現することができる．ただし，T_i は，シソーラス中に含まれる語を示す．また，ノード間を結ぶ辺，つまり語と語の関係をを e_i とすれば，$E=\{e_1,e_2,\cdots e_m\}$ と表現できる．共起関係に基づいて自動構築されたシソーラスには，ノード間の関係，つまり語と語の関係に重みが存在する．そこで，辺 $e_i=\langle T_i,T_j,s_i\rangle$ と表現する．T_i,T_j はシソーラス中の語を示し，s_i は T_i と T_j の類似度で，式 (2) で計算できる．ただし，WordNet では語同士の類似度は記されていないため，$s=1$（一定値）となる．

ここで，本アルゴリズムの手続きを説明する．

【step1】 検索要求中のすべての語について，シソーラス T_A 中にその語が含まれるか調べる．含まれていれば，その語の重要度に関する重みを 1 として，シソーラス中の位置を保持する．

【step2】 語が T_A 中に無い場合，WordNet T_W を利用して，その語の類義語を利用するため，WordNet 中にその語が含まれるか調べる．

【step3】 含まれていた場合，WordNet 中で，その語の名詞の類義語（synset）を抽出する．複数の synset が見つかった場合は，上位に挙げられている語を利用する．もし名詞に見つからなかった場合は動詞の synset を抽出する．

【step4】 WordNet から抜き出された語について，T_A とのマッピングから始めて，**【step1】** 上と同様の手続きを行う．ただし，重要度に関する重みは，元となった語の重みの $\alpha(\alpha<1)$ 倍する．これは，WordNet 上での検索質問中の語との距離が遠くなればなるほど，元の語との関連が薄くなるためである．

【step5】 T_A にマッピングされた語間の距離を利用した，語と語の関連度を求める．ここでは，ノード間の距離が語の関係に反比例すると考え，距離が近ければ近いほど，語の関係が深いと考えた．語 t_1 と t_2 の関連度は式 (3) のように定義する．

$$\mathrm{relation}(t_1,t_2)=\frac{w_{t_1}*tf_{t_1}*w_{t_2}*tf_{t_2}}{\mathrm{distance}(t_1,t_2)}*\mathrm{similarity}(t_1,t_2) \quad (3)$$

ここで，w_1 と w_2 は語 t_1 と t_2 それぞれの重みを示す．$\mathrm{similarity}(t_1,t_2)$ は，シソーラスにおける語 t_1 と t_2 の関連度である．$\mathrm{distance}(t_1,t_2)$ は語 t_1 と t_2 にそれぞれ対応するノード間の距離である．グラフにおいて，ノード間の path は一つとは限らないため，ここでは最短 path を語 t_1 と t_2 の距離とした．また，t_1 と t_2 がシソーラス中で隣接してない場合，この二つの語の $similarity$ は，最短 path における辺の $similarity$ の積とする．ただし，グラフが非連結であった場合，path が 存在しない場合がある．このとき，similarity と $distance$ はそれぞれ式 (4) , 式 (5) とする．ただし，$|T_A|$ はシソーラス $|T_A|$ に含まれる語の数とする．

$$\mathrm{similarity}(t_1,t_2)=\frac{1}{|T_A|} \quad (4)$$

$$\mathrm{distance}(t_1,t_2)=|T_A| \quad (5)$$

【step6】 検索質問中の語のすべての組合わせに対し，**【step4】** の計算を行い，それらの平均値を求める．この平均値を与えられた検索質問とサーチエンジンの関連度とする．

6. 評 価 実 験

6·1 実 験 方 法

本実験では，10 ドメイン（ Art, Business, Computer, Food, Music, Movie, Health, Legal, Sports, Science ）からそれぞれ 3 サイトずつ，計 30 サイトの英語の専門サーチエンジンを実験対象にした．これらの専門検索エンジンは，専門サーチエンジンをドメイン毎に分類して紹介しているサイトである InvisibleWeb[*3] から選択した．例えば，Health ドメインの専門検索エンジンとして，Hardin MD[*4], ACSH[*5], ManagedCare[*6]を利用した．

実験ではまず，本論文で提案している手法により作成した検索要求を専門サーチエンジンに送信し，HTML 文書の集合を得る．その際，HTML 文書内のタグや JavaScript のコードなどの，文書の内容に直接関係のない記述を除去する．また，文書中の語に対し，不要語の除去と接辞処理を行う．不要語リストには英語文書の検索システム SMART で利用されていた不要語 571 語を，接辞処理に Porter の接辞処理 [Porter 80] を利用した．そして，編集された文書集合を利用し，各手法における Source Description を作成する．

本実験では検索質問を以下のように生成する．まず，同一ドメインに属する専門サーチエンジンから収集した文書集合からランダムに 2 つ文書を選択する．そして，2 つの文書内で出現頻度が上位 100 語の中からランダムに 2 から 4 語を抽出し，そのドメインに関連した検索質問として利用する．本実験では，それぞれの手法によってランキングされたサーチエンジンの上位 3 つのうち，検索質問と同一ドメインの専門サーチエンジンの割合を適合率として評価する．そして，Source Description を構築するときに利用する文書数と適合率の関係を調べる．また，ランキング結果の上位 n 個の内，検索質問と同一ドメインのサーチエンジンの数で再現率を評価する．この際，n の値と再現率の関係を調べる．本実験では，各ドメイン毎にサーチエンジンを 3 種類ずつ用意しているため，理想的には，上位 3 個がすべて検索質問と同一のドメインであることが望ましい．本実験では，500 種類の検索質問に対し，サーチエンジン選択の実験を行い，適合率と再現率の平均を計算する．本実験では，本提案手法と CORI [Callan 95] ，そして CVV [Yuwono 97] との比較実験を行った．CORI と CVV は，語の出現頻度に基づいた手法であり，本論文で主張する問題が存在すると考えられる．

6·2 構築されたシソーラス

本節では，4·1 節 で述べた検索質問生成を利用して収集した文書から構築されたシソーラスを示す．検索質問生成において，Health ドメインでは検索質問として “medicine” や “disease” が生成された．また，Legal ドメインでは “law” や “court” が得られた．生成された語は概ね，その分野では多くの場面で用いられていると思われる語が得られた．本提案手法はデータ収集のための検索質問生成のためには有効に働いているといえる．

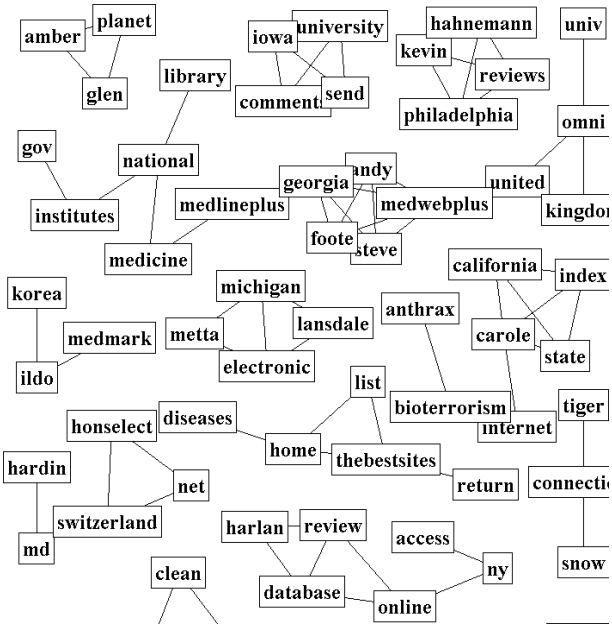

図 3 構築されたシソーラスの例

ここで，Health ドメインに属する専門サーチエンジン Hardin MD において得られた文書集合から構築されたシソーラスの一部を 図 3 に示す．構築されたシソーラス全体を示すことは困難なので，ここでは語同士の関連度が 0.6 以上のもののみを記した．図 3 では，四角で記された情報が文書中の語を表現し，関連のある語はリンクで結ばれている．ただし，語はすべて小文字で表記され，リンクで結ばれている語同士の距離が小さいほど関係が強いことを表している．構築されたシソーラスでは，人名や機関名などの固有名詞や 2 つ以上の語で一つのものを表現する複合名詞には高い関連が示されている．その他にも医療関係では関連の強い語が記されている．特に anthrax（炭疽病）と bioterrorism（バイオテロ）の関連が強いと記されていることは興味深い．ここで，図に記された語は，出現頻度によってフィルタがかけられていないことに注意したい．ここでは純粋な語同士の関連のみでフィルタがかけられているのだが，シソーラスを見れば，医療系の情報源であることが推測できる．ここからも語同士の関連がサーチエンジンが対象とするドメインの特定に有用であることが分かる．

6·3 シソーラスに基づく選択手法の有効性

シソーラス構築時に利用した文書数と適合率の関係を図 4 に示す．グラフの横軸は Source Description を作成す

*3 http://www.invisibleweb.com/
*4 http://www.lib.uiowa.edu/hardin/md/
*5 http://www.acsh.org/
*6 http://www.managedcaremag.com/

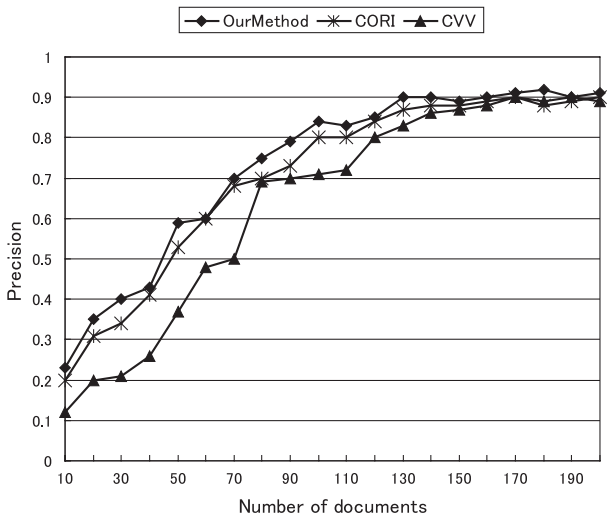

図 4 適合率に関する実験結果

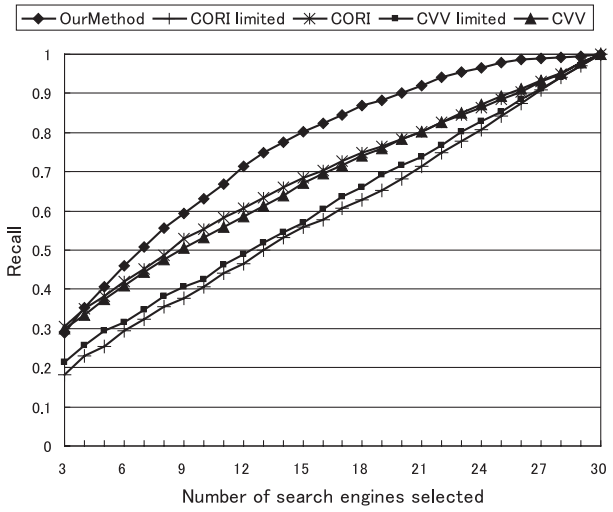

図 5 再現率に関する実験結果

るために利用した文書数，縦軸が選択の適合率である．従来手法（ CORI, CVV ）に比べ，提案手法（ Our Method ）には適合率の向上が見られる．特に Source Description を作成するために利用した文書数が低いときにその差が大きいことがわかる．しかし，Source Description の構築に利用する文書数が増加すると適合率に差が見られなくなった．このことから，十分に多い情報を利用した場合，語の出現頻度から得られた情報で語の関連と同等の表現ができるといえる．しかし，文書数を多くすると，Source Description の構築に多くの時間や，計算機リソースが必要となる．そのため，少量の情報からでも適切な Source Description を構築する必要がある．少ない情報からは語の出現頻度のみでは選択を誤る可能性がある．本提案手法では，語の出現頻度に加え，語同士の関連を利用することにより，より多くの情報が利用できる．そのため，本提案手法は，少ない文章から Source Description を作成したいような状況で有効な手法であるといえる．

表 1 WordNet の有無による再現率の違い

n	3	4	5	6
WordNet 有り	0.299	0.357	0.411	0.469
WordNet 無し	0.290	0.354	0.405	0.459

ランキングの結果の上位 n に対する再現率の結果を 図 5 に示す．グラフの横軸は n の値，縦軸は検索要求に対する選択の再現率の値である．実験では，計算機リソースの関係上，シソーラス中の語数は 1,000 に制限した．そこで，比較手法である CORI と CVV について，同様に語数を 1,000 に制限した場合（ CORI limited, CVV limited ）と，無制限に語を利用した場合（ CORI, CVV ）とで，提案手法（ Our Method ）との選択精度の比較を行った．まず，語数を制限した比較手法と本提案手法とでは，選択の精度に関して大きな差が得られた．このことから，語の出現頻度以外にも，語同士の関係という情報がサーチエンジンの選択にとって有効であるといえる．次に，語数を制限した本提案手法と，語数を制限しない比較手法とで，選択の精度を比較した場合，選択するサーチエンジン数が 3 から 5 ぐらいまではほぼ同等の精度であるが，選択数を増加させると選択の精度に大きな開きが現れた．従来手法でサーチエンジンの選択に失敗するのは，語の出現頻度に特徴が現れないときである．しかし本実験から，提案手法では，そのようなサーチエンジンにおいても，語同士の関係を用いることにより，より高精度な選択が実現できていることがわかる．サーチエンジンの選択において，サーチエンジンの数が増加すると，適切なサーチエンジンの選択を実現するために管理すべき情報も増加する．そのため，出現するすべての語の出現頻度を管理することは困難であり，利用する語を制限する必要がある．本提案手法では，同一語数を利用した既存の手法だけでなく，無制限に利用した手法とも同程度かそれ以上の選択精度が得られることがわかった．

6·4 WordNet による語彙補完の有効性

提案手法では，語彙の不足を解消するために，WordNet を利用することを提案している．そこで，WordNet を利用した場合と，利用しない場合で，提案手法における再現率の評価を行った．その結果を 表 1 に示す．本評価では，ランキングされた結果の上位 n 個を選んだときの再現率を比較した．表 1 では n の値が 3 から 6 の間の結果を示した．実験では WordNet を利用したとき，利用しない場合と比べて再現率がわずかに向上しただけであった．これは，あるドメインの検索要求に含まれる語彙に関して，WordNet が類義語を見つけられないため，WordNet がほとんど利用されていないことが原因となっていた．文献 [Manala 99] で指摘されているように，WordNet は汎用的なシソーラスであり，ドメイン固有の語などは含まれていないことが多い．結果として，専門サーチエンジ

ンを利用したメタサーチエンジンにおいて，WordNet を利用した語彙不足の解消は困難であるといえる．

7. お わ り に

本論文では，WWW のメタサーチエンジンにおける，シソーラスを利用したサーチエンジン選択アルゴリズムを提案した．本手法では，サーチエンジンが検索結果として提示した URL の HTML 文書からシソーラスを自動構築し，シソーラスと検索要求の関連性を評価する．語同士の関係が記述されているシソーラスを利用することにより，語の意味内容を考慮したサーチエンジンの選択が実現でき，従来の語の出現頻度を利用した手法に比べ，低出現頻度の語や多義語によって生じる問題に対処できる．本手法では，検索質問とサーチエンジンの関連を評価するために，シソーラスがグラフ構造をしていることから，グラフ中の語と語の距離を利用した．提案手法の有効性を確認するために行った実験から，提案手法が従来手法に比べ優れた選択精度が得られることを示した．

提案手法に基づいて構築されるシステムでは，選択に利用する Source Description がサーチエンジン毎に独立に構築することができるため，本システムにおける Wrapper エージェントはサーチエンジン毎に完全に独立して構築できる．これにより新しいサーチエンジンを追加するには，そのサーチエンジンの Wrapper の機能を備えたエージェントを追加するだけでよく，スケーラビリティに優れたシステムになる．

今後の課題として，HTML 文書に特化したシソーラス自動構築の新しい手法の考案が考えられる．HTML 文書はタグ付けされており，半構造化されている．このような文書からのシソーラス自動構築は，平文を対象としたときよりも高い精度で作ることができると考えられる．また，現在のシソーラスは共起するすべての語の組み合わせを考慮している．そのため，シソーラス構築に必要な時間やメモリ等の計算機リソースが大きくなってしまう．そこで，シソーラス構築の高速化や最適化が必要になると考えられる．

◇ 参 考 文 献 ◇

[Callan 95] Callan, J. P., Lu, Z., and Croft, W. B.: Searching Distributed Collections With Inference Networks, in *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21–28 (1995).

[Callan 00] Callan, J.: *Advances in Information Retrieval*, chapter 5: Distributed information retrieval, pp. 127–150, Kluwer Academic Publishers (2000).

[Fan 99] Fan, Y. and Gauch, S.: Adaptive Agents for Information Gathering from Multiple, Distributed Information Sources, in *Proc. of 1999 AAAI Symposium on Intelligent Agents in Cyberspace* (1999).

[Hawking 99] Hawking, D. and Thistlewaite, P.: Methods for Information Server Selection, *ACM Transactions on Information Systems*, Vol. 17, No. 1, pp. 40–76 (1999).

[Howe 97] Howe, A. E. and Dreilinger, D.: SAVVYSEARCH: A Metasearch Engine That Learns Which Search Engines to Query, *AI Magazine*, Vol. 18, No. 2, pp. 19–25 (1997).

[Kim 98] Kim, M.-C. and Choi, K.-S.: A Comparison of Collocation-Based Similarity Measures in Query Expansion, *Information Processing and Management*, Vol. 35, No. 1, pp. 19–30 (1998).

[Lawrence 98] Lawrence, S. and Giles, C. L.: Inquirus, the NECI Meta Search Engine, in *Proc. of the 7th International World Wide Web Conference*, pp. 95–105 (1998).

[Lawrence 99] Lawrence, S. and Giles, C. L.: Accessibility of Information on the Web, *Nature*, Vol. 400, pp. 107–109 (1999).

[Manala 99] Manala, R., Tokunaga, T., and Tanaka, H.: Combining General Hand-Made and Automatically Constructed Thesauri for Information Retrieval, in *Proc. of the 16th International Joint Conference on Artificial Intelligence*, pp. 920–925 (1999).

[Miller 95] Miller, G. A.: WordNet: A Lexical Database for English, *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41 (1995).

[Park 95] Park, Y. C., Han, Y. S., and Choi, K.-S.: Automatic thesaurus construction using Bayesian networks, in *Proc. of the International Conference on Information and knowledge Management*, pp. 212–217 (1995).

[Porter 80] Porter, M. F.: An Algorithm for Suffix Stripping, *Automated Library and Information Systems*, Vol. 14, No. 3, pp. 130–137 (1980).

[Powell 00] Powell, A. L., French, J. C., Callan, J., Connell, M., and Viles, C. L.: The Impact of Database Selection on Distributed Searching, in *Proc. of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–239 (2000).

[Yuwono 97] Yuwono, B. and Lee, D. L.: Server Ranking for Distributed Text Retrieval Systems on the Internet, in *Proc. of the 5th International Conference on Database Systems for Advanced Applications*, pp. 41–49 (1997).

[山田 01] 山田, 村田, 北村：知的 Web 情報システム, 人工知能学会誌, Vol. 16, No. 4, pp. 495–502 (2001).

[北村 98] 北村泰彦：インターネット上での知的情報統合, in *Advanced Database Symposium '98*, pp. 167–174 (1998).

〔担当委員：武田英明〕

2001 年 9 月 17 日 受理

著 者 紹 介

後藤 将志（学生会員）

2000 年名古屋工業大学工学部知能情報システム学科卒業．2002 年同大学院博士前期課程修了．WWW 上の情報検索，情報エージェント，知識共有支援に興味を持つ．電子情報通信学会，情報処理学会，計測自動制御学会各学生会員．

大囿 忠親（正会員）

1995 年名古屋工業大学工学部知能情報システム学科卒業．2000 年同大学大学院工学研究科電気情報工学専攻博士後期課程終了．同年より同大学工学部知能情報システム学科助手．現在に至る．博士（工学）．情報エージェントの研究に従事．AAAI，ACM，情報処理学会，日本ソフトウェア科学会各正会員．

新谷 虎松（正会員）

1982 年東京理科大学大学院修士課程修了．同年富士通（株）国際情報社会科学研究所入所．知識システム，論理プログラミングなどの研究に従事．1993 年名古屋工業大学知能情報システム学科助教授．1999 年同大学同学科教授．1999 年から 2000 年にかけて米国カーネギーメロン大学ロボティクス研究所客員研究員．現在に至る．博士（工学）．意思決定支援システム，マルチエージェントシステム，電子商取引支援システム，WWW 情報検索の研究に従事．AAAI，電子情報通信学会，情報処理学会，日本ソフトウェア科学会各会員．