Web 検索エンジンを用いた用語検索履歴からのシソーラス自動構築手法の評価と改良

安川 美智子† 山田 篤‡

†群馬大学工学部 〒376-8515 群馬県桐生市天神町 1-5-1
‡財団法人 京都高度技術研究所 〒600-8813 京都市下京区中堂寺南町 134 番地
E-mail: †michi@cs.gunma-u.ac.jp, ‡yamada@astem.or.jp

あらまし 過去に閲覧した Web ページを再閲覧する際に，目的とする Web ページを漏れなく検索できるキーワードを思いつくことは容易ではない．また，Web 検索を行う際に，検索対象について知っているが，名前は分からず，目的とする Web ページがうまく検索できないという問題がしばしば起こる．このような場合に，シソーラスを用いた検索語の選択支援が有効であると考えられる．我々は Web ページの検索支援に用いるシソーラスの自動構築手法をこれまでに提案してきた．本稿では，シソーラスを用いた検索質問拡張とシソーラスの評価方法について，検討し，構築したシソーラスの評価を行う．また，これまでに提案してきたシソーラス自動構築手法の分析を行い，シソーラスの精度を向上させるための改良手法の提案と，改良手法の評価も行う，

キーワード 情報検索，シソーラス，閲覧履歴，検索質問拡張

Evaluation and Improvement of the Automatic Thesaurus Construction Method from Terminology Search History using Web Search Engine

Michiko YASUKAWA† Atsushi YAMADA‡

†Faculty of Engineering, Gunma University, 1-5-1 Tenjin-cho, Kiryu, 376-8515, Japan
‡ASTEM RI, 134 Chudoji Minami-machi Shimogyo-ku Kyoto, 600-8813, Japan
E-mail: †michi@cs.gunma-u.ac.jp, ‡yamada@astem.or.jp

Abstract Query expansion using automatic constructed thesauri is worth investigating to reduce user's difficulties in selecting appropriate keywords for searching relevant web pages. Even when users know their objectives, they often fail in identifying or remembering proper names or terms that represent their objectives. We propose a method for automatically constructing a thesaurus from a set of search logs. After the user searches for initial terms using web search engines, our system counts frequencies of co-occurring terms appeared in his/her browsed pages, which are then used to calculate similarities of related terms. This paper evaluates experimentally the proposed method and improves it. Our query expansion method is also discussed.

Keyword Information Retrieval, Thesaurus, Web Browsing History, Query Expansion

1. はじめに

現在，Web 上の情報の検索や閲覧が日常的に行われるようになってきている．Web 上の情報を効果的，効率的に活用するためには，一度閲覧した Web ページを後で再度検索し，閲覧できることが望まれる．このような Web 上の情報の収集と閲覧を支援する個人用のアーカイブシステムが提案されている(文献[1]).

Web 上の情報や，Web アーカイブ中に保存された Web ページ（アーカイブデータ）を検索，閲覧する際の問題として，ユーザは自分の求めている情報を検索するための適切な検索キーワードを思いつくことが容易ではないという問題がある．たとえばユーザは，検索対象とする物や概念について知っているが，その対象物や概念の名前が分からないということがある．また，過去に検索したことがある Web ページの検索キーワードを思い出せないということもある．そのような場合に，Web ページの検索に役立つシソーラスがあれば，ユーザは，より検索を行いやすくなると考えられる．文献[3][4]では，Web ページの検索に役立つシソーラスを，ユーザの Web 検索・閲覧の履歴と閲覧した Web ページから自動構築する手法が提案されている．この提案手法は，ユーザが Web 検索エンジンを用いて用語検索を行った際の Web 検索・閲覧履歴に注目し，用語検索の文脈において関連の強い語を抽出する点に

特徴がある．文献[3][4]では，提案手法により構築されたシソーラスの例示と構築手法の特徴についての考察を行っているが，構築されたシソーラスの解析的な評価を行っていなかった．本稿では，シソーラスを用いた検索質問拡張の手法とシソーラスの評価方法を検討し，提案手法の分析と問題点を解決するための改良について述べる．また，改良手法により構築したシソーラスの評価を行う．

以下，2 章でシソーラス構築と検索質問拡張の手法について，3 章で提案手法の分析と改良，4 章で評価実験，５章でまとめと今後の課題を述べる．

2. シソーラス自動構築とシソーラスを用いた検索質問拡張

シソーラス自動構築手法の関連研究と，用語検索履歴からのシソーラス自動構築手法の概略を説明し，シソーラスを用いた検索質問拡張の手法について述べる．

2.1. シソーラス構築手法

シソーラスとは辞典のことであり，情報検索において検索キーワードを選択するために用いられる．シソーラスは以下の 2 つのタイプに分けられる(文献[7])．

- 手動で構築されるシソーラス
- 自動構築されるシソーラス

手動構築のシソーラスとしては WordNet(文献[16])，日本語語彙体系(文献[15])，分類語彙表(文献[18])などがある．また，文献[8]では，メタサーチ，ユーザインタフェース，シソーラスを使った Web 検索の手法が提案されており，ユーザが手動でシソーラスを作成する GUI が提供されている．

手動構築のシソーラスには，構築や維持管理のためのコストが高いという問題がある．このため，手動ではなく，計算機によってコーパスベースのシソーラスを自動構築する手法が提案されている．シソーラス自動構築手法は，以下の 2 つのタイプに分けられる．

- 語の共起関係を使う手法
- 語の格関係を使う手法

文献[5]では，語の共起関係の潜在的意味分析により，シソーラスを構築して，構築したシソーラスを用いた検索性能を向上させる結果を得ている．

語の格関係を使う手法としては，たとえば，「主語-動詞」「動詞-目的語」「形容詞-名詞」の関係に注目し，以下のものは類似すると考えて，同義語・類義語の判断をする手法が提案されている（文献[17]）．

格関係を使う手法は，「同じ意味で違う表記（同義語・類義語）」の語のペアを見つけ出すことができるという利点がある．しかし，シソーラス構築元となるテキストデータが文法的な誤りを含むものである場合には，この手法はうまくいかない．

従来の研究が，主に論文や新聞などのテキストデータだけを対象としてきたのに対して，近年，Web 情報からシソーラスを自動構築する研究が行われている．Web 情報からのシソーラス自動構築は，Web ページ中のテキストデータだけでなく，Web ページ間のリンク構造や，アンカー文字列など，Web ならではの特徴を考慮して，関連語の抽出が行われる．

文献[6]で提案されているシソーラス自動構築手法は，Web のリンク分析の手法に基づくシソーラス自動構築手法であり，Web サイトのリンク構造から，表面上は明らかになっていない潜在的なコンテンツ構造を抽出し，シソーラスを構築する手法を提案している．文献[13]の提案手法は，Web ページや電子メールに含まれる語の共起関係からのシソーラス構築を行い，個々のユーザが閲覧した電子メールや Web ページを蓄積しておくパーソナルリポジトリの協調検索でのシソーラス利用を提案している．文献[2][3]で提案されているシソーラス自動構築手法は，Web 検索エンジンを用いた用語検索の履歴と閲覧済み Web ページをシソーラス構築元データとして，閲覧済み Web ページに含まれる語の共起関係をもとにシソーラスを構築するものである．提案手法は，ユーザが Web 検索エンジンを用いて用語検索を行う際の検索・閲覧履歴と，ユーザにより閲覧される用語説明の Web ページの関連を利用している点に特徴がある．用語検索履歴からのシソーラス自動構築手法の概略を次に述べる．

2.2. 用語検索履歴からのシソーラス構築手法

ユーザは，検索したい Web ページを特定可能なキーワード（「主キーワード」と呼ぶ）が分からない，あるいは，覚えていない場合がしばしばある．我々の提案するシソーラス自動構築手法は，主キーワードに関連のある副次的なキーワード（「副キーワード」と呼ぶ）をユーザが思い出すことができれば，シソーラスを用いた検索質問拡張により，ユーザの目的とする Web ページがうまく検索できるようにするものである．

たとえば，「エンダイブ」という用語について知りたいユーザが，検索エンジンで検索した際の用語検索履歴からシソーラスの自動構築を行う例は図 1 のようになる．このユーザが，閲覧した「エンダイブ」に関するページを後で再度閲覧する際に，「エンダイブ」という語を忘れてしまった場合など，ユーザの検索したいページを特徴付ける重要な検索キーワードが不明な

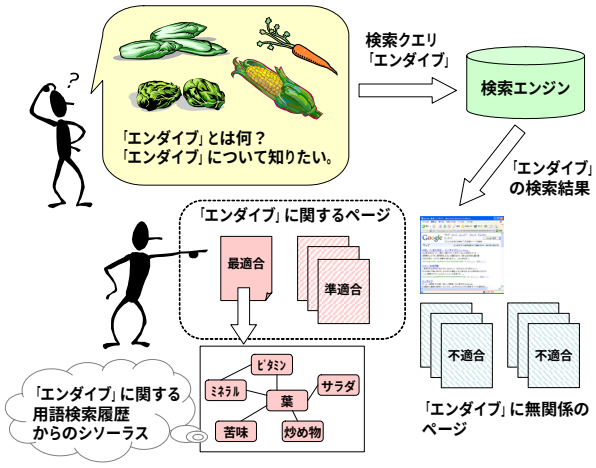

図 1 用語検索履歴からのシソーラス自動構築

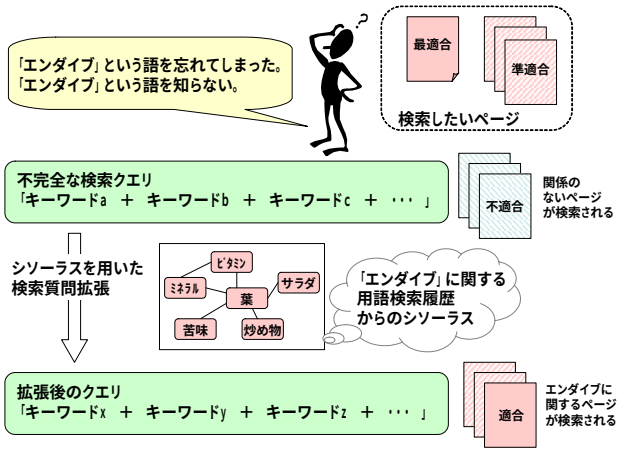

図 2 シソーラスを用いた検索質問拡張

場合がある．

重要な検索キーワードが不明なためにユーザが提示する検索クエリが不完全である場合に，シソーラスを用いた検索質問拡張を行うことで，ユーザが目的の Web ページを検索しやすくなると考えられる（図 2）．

提案手法では，以下の 3 つのタイプのシソーラス構築を行う．

（1）語の直接共起に基づくシソーラス

シソーラス構築元の Web ページ（図 1 の最適合のページ）の中の同一文中での語の共起頻度から，相互情報量(Mutual Information)を求め，互いに共起関係にある語を抽出する．直接共起の関係にある語 a と語 b の相互情報量 $MI(a,b)$ を次のように定義する．

$$MI(a,b) = \log 2 \frac{N \times freq(a,b)}{freq(a) \times freq(b)}$$

ここで， N は Web ページ中の語の頻度総数であり，$freq(a)$，$freq(b)$ はそれぞれ語 a，語 b が出現する頻度，$freq(a,b)$ は語 a と語 b が共出現する頻度である．

相互情報量 $MI(a,b)$ の値が閾値以上の語のペア語 a と語 b を抽出し，抽出した語のクラスタを直接共起に基づくシソーラスとする．

（2）語の間接共起に基づくシソーラス

シソーラス構築元の Web ページ中の同一文中では直接共起しないが，他の語（媒介語と呼ぶ）を介した間接的な共起関係にある語を抽出する．間接共起の関係にある語 b と語 c，および媒介語 a の相互情報量を直接共起で定義した相互情報量を用いて，次のように定義する．

$$MI(a,b) = \log 2 \frac{N \times freq(a,b)}{freq(a) \times freq(b)}$$

$$MI(a,c) = \log 2 \frac{N \times freq(a,c)}{freq(a) \times freq(c)}$$

相互情報量 $MI(a,b)$ と $MI(a,c)$ の値が閾値以上の語のペア語 b と語 c を抽出し，抽出した語のクラスタを間接共起に基づくシソーラスとする．

（3）潜在的意味分析に基づくシソーラス

シソーラス構築元の Web ページ中の同一文中での語の共起頻度を行列の要素とする共起行列 A の特異値分解により得られる特徴行列から語の類似度を計算し，類似度の高い語を抽出する．共起行列 A の特異値分解 $[U \quad S \quad V] = svd(A)$ により得られる特徴ベクトル U の語 a と語 b に対応する行ベクトルの余弦(Cosine)から，語 a と語 b の類似度 $sim(a,b)$ を次のように定義する．

$$sim(a,b) = \frac{\sum_{i=1}^{Tp}(u_i^a \times u_i^b)}{\sqrt{\sum_{i=1}^{Tp}(u_i^a \times u_i^a)} \times \sqrt{\sum_{i=1}^{Tp}(u_i^b \times u_i^b)}}$$

ここで， Tp は Web ページ中の語の異なり数， u_i^a は，特徴ベクトル U の語 a に対する行の， u_i^b は語 b に対応する行のそれぞれ i 番目の要素である．

類似度 $sim(a,b)$ の値が閾値以上の語のペア語 a と語 b を抽出し，抽出した語のクラスタを潜在的意味分析に基づくシソーラスとする．

2.3. 検索質問拡張手法

シソーラスを用いた検索質問拡張はブーリアンモデルによる情報検索システムに適していない．ブーリアンモデルによる検索では，索引語（検索対象ページの内容を特徴付ける語；index term）が含まれるかどうかだけが考慮されるので，語の出現頻度は無視されてしまうためである．ブーリアンモデルによる検索で検索質問拡張を行うと，検索性能を低下させてしまう．

このため，シソーラスを用いた検索質問拡張による性能向上を評価する際には，重み付きの情報検索システムが用いられる．たとえば，文献[6]では全文検索システム“ the Okapi system ”が，文献[7]では SMART version 11.0 が用いられている．本稿では，Web ページ中のテキストから抽出される全ての語を索引語として，以下のようなベクトル空間モデルに基づく重み付き検索で，検索質問拡張を行う．

（1）ベクトル空間モデルに基づく重み付き検索

検索クエリ q（query；検索質問）と Web ページ d（document；検索文書）の類似度は，索引語の重みを要素とするベクトルの余弦(Cosine)により求める．

検索クエリ q と Web ページ d の類似度 $sim(q,d)$ を次のように定義する．

$$sim(q,d) = \frac{\sum_{i=1}^{Ta}(w_i^q \times w_i^d)}{\sqrt{\sum_{i=1}^{Ta}(w_i^q \times w_i^q)} \times \sqrt{\sum_{i=1}^{Ta}(w_i^d \times w_i^d)}}$$

ここで，Ta は検索対象の索引語の異なり数，w_i^q は検索クエリ中の，w_i^d は Web ページの，それぞれ i 番目の索引語 t_i に対する重みである．

索引語 t_i の重み w_i は，TF/IDF 値（term frequency / inverse document frequency）の尺度を用いて，次のように定義する．

$$w_i = freq(t_i) \times \left(1 + \log\left(\frac{n}{df(t_i)}\right)\right)$$

ここで，$freq(t_i)$ は索引語 t_i の出現頻度，$df(t_i)$ は索引語 t_i の出現する Web ページの数，n は検索対象となる Web ページの総数である．

（2）検索質問拡張

$\mathbf{q} = (w_1^q, w_2^q, \ldots, w_T^q)$ をクエリ q の重みベクトル，$\mathbf{q}_e = (w_1^{qe}, w_2^{qe}, \ldots, w_T^{qe})$ を拡張後クエリ q_e の重みベクトル，$\mathbf{a} = (a_1, a_2, \ldots, a_T)$ を質問拡張により拡張後クエリに付加される重みベクトルとして，検索質問拡張を次のように定義する．

$$\mathbf{q}_e = \mathbf{q} + \mathbf{a}$$

ユーザが検索したい Web ページを d とすると，d と検索クエリ q の類似度が最小で，d と拡張後の検索クエリ q_e の類似度が最大のとき，すなわち，

$$sim(q,d) = 0$$

$$sim(q_e,d) = 1$$

のとき，検索質問拡張の効果が最大となる．
また，

$$sim(q_e,d) = 0$$

のときは，拡張後の検索質問は，検索対象ページの特徴を全く表現できていないことになり，

$$sim(q\ ,d) = 1.0$$

のときは，検索質問拡張を行う前のクエリが既に対象ページを適切に表現できていることになるため，検索質問拡張をする必要がないということになる．

数値を使った簡単な具体例を以下に示す．検索したい Web ページ d と拡張前のクエリ q の値が，

$d = (1 \quad 0.2 \quad 0.1 \quad 0\ )$
$q = (0 \quad 0.2 \quad 0.1 \quad 0\ )$

のときに，拡張後のクエリが，たとえば，

$$q_e = (1 \quad 0.2 \quad 0.1 \quad 0\ )$$

の方が

$$q_e = (0 \quad 0.1 \quad 0.2 \quad 1\ )$$

よりも Web ページ d との類似度が高い質問拡張を行えている．

3. 提案手法の分析と改良

前節で述べたシソーラス構築手法（2.2 節）と検索質問拡張の手法（2.3 節）でどのような検索性能向上が得られるかを，以下のような評価データと，評価方法により分析する．

（1）評価データ

評価用のデータ（Web ページ群）を，以下の手順により人手で収集した．

Web ページ収集の手順

検索対象となる用語を検索キーワードとして入力して，Web 検索エンジンを用いた用語の検索を行い，検索結果のトップから順番に Web ページを閲覧し，適合な Web ページを保存し，適合な Web ページが 10 件得られた時点で，検索を終了する．

表 1 評価データ収集の用語検索で用いた用語

カテゴリ	用語
野菜	コールラビ，パクチョイ， パースニップ，トマピー
薬の成分	アロエ，ウイキョウ

表 2 評価データの統計量

	ページ中の語		文の数	一文中の語数
	出現数	異なり数		
最小	126	86	6	1
最大	1490	832	192	117
平均	391	234	41	10

実際の Web 検索の場面では，検索結果として得られる Web ページが適合であるか，あるいは有用であるかどうかの基準は，人によって，状況によって様々であると考えられるが，ここでは，用語に関連のある説明文を含まないページ（たとえば，検索用語と同じ名前をもつパソコンショップや歌手グループのホームページ）やテキストデータの分量が少なすぎるページ（写真やフラッシュなどの画像が主体でテキストデータが 5 行に満たないページ）を除外し，他は評価用の適合な Web ページとして扱うこととした．

検索エンジンには Google を，検索対象の用語には「現代用語の基礎知識 (CD-ROM 版)」のカテゴリ 2 種類（『スーパーでみかける「新顔野菜」話題学』および『薬の成分話題学《生薬系》』）の中の見出し語 6 個を用いた（表 1）．

収集した評価データ Web ページ 60 件に含まれる文の数と語の数の統計量（60 件のページの最小，最大，平均値）は表 2 のようになっている．

（2）構築されるシソーラスの評価方法

本研究の目的は，「ユーザが索引語（Web ページ中の語）のうち，主キーワード（最も特定性の高い索引語）は分からないが，副キーワード（主キーワード以外の特定性の高い索引語）は分かる，というときに，検索質問拡張により，検索がうまくいくようにして，ユーザを支援する」ことである．

「検索がうまくいく」「うまくいかない」というのは，ユーザが検索に何を要求しているかにより異なる．ユーザの要求は大きく以下の 2 種類に分けられる．

- 最も適合な Web ページを 1 件だけ閲覧したい
- 適合な Web ページを全て閲覧したい

そこで上記のユーザ要求に応えられるかを検証するために，以下の数量により，検索の性能を測定する．

- 最適合な Web ページの順位
- 検索結果上位 n 件中の適合 Web ページの数

評価用データの個々の Web ページの索引語の重みベクトルから，「主キーワードは分からないが，副キーワードは分かる」というユーザの検索クエリを，擬似的に生成し，拡張前クエリとして用いて評価を行うことを考えた．評価の手順は以下のようになる．

まず，評価データの Web ページから，2.2 節で述べた 3 種類のシソーラスを構築する．また，評価データの Web ページの索引語の重みベクトルを，擬似的なクエリと見立てて，これを，最適合 Web ページとの類似度が最大であることから，ここでは**理想値クエリ**と呼ぶ．さらに，理想値クエリの主キーワードの索引語重みをゼロに設定したものを擬似的な**拡張前クエリ**として，拡張前クエリに対して，構築したシソーラスを用いた検索質問拡張を行い，**拡張後クエリ**を生成する．そして，拡張前クエリを用いた検索と拡張後クエリを用いた検索を行い，検索結果を比較する．

（3）予備実験と考察

上に述べた評価データと評価方法により，シソーラスの評価を行う．具体的には以下の手順で予備実験を行った．まず，評価データの Web ページ 60 件をシソーラス構築の元データとして，2.2 節で述べたシソーラス自動構築手法（文献[3][4]で提案した手法）により，60 件のシソーラスを構築した．また，評価データの Web ページ 60 件から，理想値クエリと拡張前クエリをそれぞれ 60 個作成し，さらに，拡張前クエリに対して，シソーラスを用いた拡張を行った拡張後クエリ 60 個を生成した．そして，拡張前クエリと拡張後クエリの両方で，評価データの Web ページ 60 件の検索を行い，拡張前クエリでの検索結果と拡張後クエリでの検索結果とを比較した．閾値等の定数の値は文献[3][4]と同様に，有効共起回数は 2 回以上，間接共起の媒介語の数は 2 個以上，各ページにおける有効な相互情報量は，ページ中の相互情報量の最大値の 0.8 倍）とした．

クエリに追加される重みベクトル $\mathbf{a}=(a_1,a_2,\ldots,a_T)$ の値には，シソーラスの中で定義されている，関連語の類似度を，類似度の最大値で正規化した値（値の範囲は 0 以上，1 以下）を用いた．

拡張前クエリと拡張後クエリで，評価データの Web ページ 60 件の検索を行って，検索結果として，最適合な Web ページの順位と検索結果上位 10 件に含まれる適合 Web ページの数を調べた．最適合な Web ページ

の順位を，表 3 に示す．表 3 において，「理想値」は「理想値クエリ」（最適合との類似度が最大で，主キーワードが欠落していないクエリ）に対応し，「拡張前」は「拡張前クエリ」（理想値クエリの主キーワードの重みをゼロに設定して作成したクエリ）に対応し，「拡張後（直接）」「拡張後（間接）」「拡張後（潜在）」は，それぞれ，直接，間接，潜在のシソーラスで拡張を行った拡張後クエリに対応している．表 3 に示すように最適合ページの順位は，ほぼ全ての場合で 1 位であった．これは，擬似的なクエリが，元の Web ページから生成されているためであると考えられる．

次に，検索結果上位 10 件に含まれる適合文書の平均数と割合をそれぞれ，表 4 と図 3 に示す．表 4 と図 3 において，「適合」は，検索結果上位 10 件中に含まれる，クエリに適合な Web ページの数である．具体的には，たとえば，用語「アロエ」の Web ページから作成された擬似クエリに対して検索された，用語「アロエ」に関する Web ページの数である．

表 3　最適合ページの順位の分布

	順位 1 位	順位 2 位	順位 3 位以上
理想値	59	1	0
拡張前	59	1	0
拡張後（直接）	59	1	0
拡張後（間接）	59	1	0
拡張後（潜在）	58	2	0

（件）

表 4　検索結果上位 10 件の適合数の平均

	適合	カテゴリ適合	不適合
理想値	6.20	3.38	0.42
拡張前	5.07	4.15	0.78
拡張後（直接）	5.13	3.85	1.02
拡張後（間接）	6.20	3.38	0.42
拡張後（潜在）	4.43	3.42	2.15

（件）

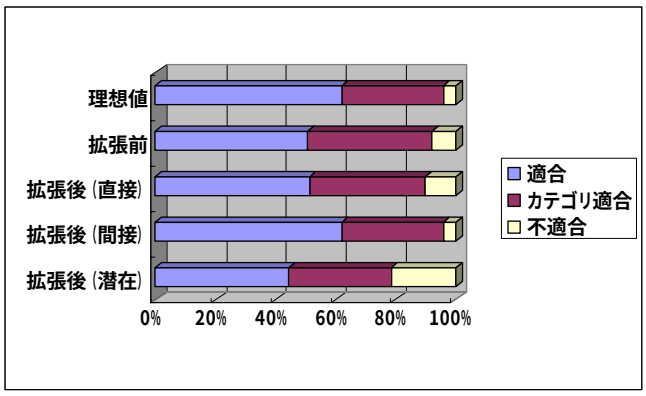

図 3　検索結果上位 10 件の適合数の割合

表 5　検索結果上位 10 件の適合ページの一致

	直接のみ	間接のみ	潜在のみ	直接と間接	間接と潜在	直接と潜在	直・間・潜	直接	間接	潜在
平均	0.5	1.0	0.4	1.1	0.5	0.1	3.4	5.1	5.9	4.4

（件）

「カテゴリ適合」は用語に対しては適合ではないが，用語と同じカテゴリの Web ページが検索された場合である．（たとえば，「アロエ」で同じ「薬」のカテゴリである「ウイキョウ」が検索された場合）

表 4 と図 3 の結果から，シソーラスを用いた検索質問拡張を行うことで，主キーワードが復元されるわけではないが，副キーワードに適切な重み調整がなされたことで，適合ページに類似のクエリを生成することに，ある程度は成功していると言える．特に，間接共起は，主キーワードが欠落していない場合に等しい検索性能が得られている．

次に，3 種類の性質の異なるシソーラスで拡張を行った場合に，検索結果上位 10 件に含まれる適合ページに，どれだけ一致・差異があるかを表 5 に示す．表 5 において「直・間・潜」は，直接と間接と潜在のシソーラスで共通して上位 10 件に含まれていた適合ページの数（平均）を表している．表 5 の結果より，用いるシソーラスによって，拡張後クエリの検索結果に差異が出ることが分かる．個々のシソーラスの性質をうまく利用することにより検索精度の向上が期待できるが，そのためには，個々のシソーラスの精度を高める必要がある．そこで，次にシソーラスの精度を高める改良手法について考える．

（4）シソーラス自動構築手法の改良

Web ページ中に含まれる文の数が少なく，語の出現頻度が低い Web ページでは，相互情報量による共起関係の抽出が適さない．Web ページ中に含まれる文の数が少ない場合は，語の出現回数が有効共起回数に満たない場合や，相互情報量に基づいて共起関係をとらえると閾値に満たない場合があるため，うまく語の関連をとらえることができないからである．

逆に，全体的な文数，語数が多く，語の出現頻度も高いページでは，頻度の低い語をシソーラスに含めるとシソーラスの精度が悪くなると考えられる．

以上のことから，シソーラス構築元の Web ページのサイズの多様さに合わせて，シソーラス構築を行うことで，より精度の高いシソーラス構築が行えると考えられる．シソーラス自動構築手法の改良として，以下の対策を行うことを考えた．

サイズの小さい Web ページに対する対策

直接,間接の共起関係の抽出に,ダイス係数を用いる.語 a と語 b のダイス係数を以下のように定義する.

$$D(a,b) = \frac{2 \times freq(a,b)}{freq(a) + freq(b)}$$

ここで, $freq(a)$, $freq(b)$ はそれぞれ語 a, 語 b が出現する頻度, $freq(a,b)$ は語 a と語 b が共出現する頻度である.

サイズの大きい Web ページに対する対策

共起関係の抽出に関与する語を,頻度の高い語に限定する.

また,表 2 の統計量で示したように,極端に文中の語数が多いものがある.Web ページからの文・語の抽出には,Windows2000 以降の Indexing Service に含まれる Japanese Word Breaker を使用しているが,元の Web ページの HTML 記述によっては,Word Breaker が文の切れ目の抽出に失敗するために,このような極端に長い文が含まれてしまう.一文中の共起関係の有効な範囲を平均的な語数である 10 語程度に予め限定しておくことで,このようなエラーにも対処できるようになると考えられる.

4. 評価実験

シソーラスの自動構築に用いる語を出現頻度の高い語上位 20 件に,また,一文中の語の距離を 10 語に,それぞれ限定して,共起計算に,ダイス係数を使ってシソーラス自動構築と検索質問拡張を行った.

最適合ページの順位を表 5 に,検索結果上位 10 件の適合数の平均を表 6 に,検索結果上位 10 件の適合数の割合を図 4 に示す.また,改良による検索精度の改善結果を図 5 と表 7 に,拡張語の例を表 8 に示す.

考察

改良前(表 4,図 3)と改良後(表 6,図 4)の結果から,「拡張後(間接)」の検索性能は,「拡張後(直接)」や「拡張後(潜在)」での検索性能と比較して,あまり改善されていない.直接共起と間接共起は,「同一文で共起するもの」と「同一文では共起しないもの」という排他的な関係にある.Web ページのサイズが大きい場合の対策として,共起関係の抽出に関与する語を高頻度語に限定したことで,雑音になる情報が削減され,より Web ページの主題に近い語の共起関係が語の直接共起として抽出されることとなったが,逆に雑音が減ることで,間接共起を抽出するための媒介語が減少することとなり,間接共起の共起語の抽出が行われにくくなったと考えられる.

表 5　最適合ページの順位の分布(改良後)

	順位 1 位	順位 2 位	順位 3 位以上
理想値	59	1	0
拡張前	59	1	0
拡張後(直接)	59	1	0
拡張後(間接)	59	1	0
拡張後(潜在)	59	1	0

(件)

表 6　検索結果上位 10 件の適合数の平均(改良後)

	適合	カテゴリ適合	不適合
理想値	6.20	3.38	0.42
拡張前	5.07	4.15	0.78
拡張後(直接)	5.38	3.82	0.80
拡張後(間接)	6.28	3.32	0.40
拡張後(潜在)	5.07	3.82	1.12

(件)

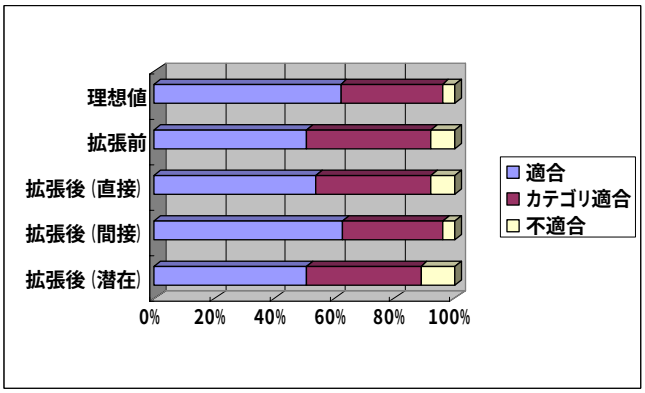

図 4　検索結果上位 10 件の適合数の割合(改良後)

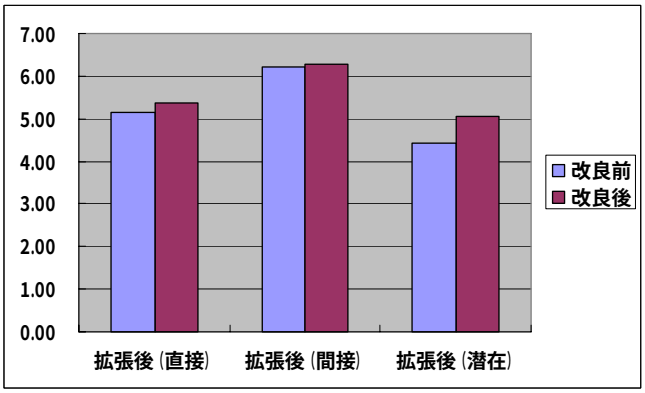

図 5　検索結果上位 10 件の適合数(平均)の比較

表 7　改良手法による検索精度の悪化と改善

	検索精度悪化	変化なし	検索精度改善
拡張後(直接)	10	20	30
拡張後(間接)	1	53	6
拡張後(潜在)	16	21	23

(件)

表 8　改良により精度が改善された拡張語の例

	用語「パースニップ」の直接シソーラスで関連度の高い語
改良前	塩，コショウ，調理，済み
改良後	オーブン，入れる，美味，鶏

表 9　改良により精度が悪化した拡張語の例

	用語「パースニップ」の間接シソーラスで関連度の高い語
改良前	リゾット，適量，小さじ
改良後	塩，使う，ドレッシング，大きい，材料

図 5，表 7 から改良手法によって，個々のシソーラスの精度が改善していることが分かる．また，表 4 と表 6 から「拡張後（直接）」,「拡張後（間接）」,「拡張後（潜在）」の上位 10 件中の適合数の平均を求めると改良前は 5.4 件，改良後は 5.6 件となっており，改良手法により，3 種類のシソーラスの平均的な性能も向上していると言える．

次に，拡張された検索語の考察を行う．クエリ拡張に失敗している場合はページの主題に対してより特定性の低い語が強調され，拡張に成功している場合は，より特定性の高い語が強調されていると考えられる．

たとえば,「パースニップ」という名前の野菜は,「加熱すると臭みが抜け,甘味が増す」「長時間煮込んでも煮崩れしないので，煮込み料理に向く」という特徴があり，適合ページには，これらの概念に関連する記述が含まれている．表 8 の改良後，表 9 の改良前の検索精度が高いのは,表 8 の改良前より改良後の方が,また,表 9 の改良後より改良前の方が,「パースニップ」という用語に関して，より特定性が高い語がクエリ拡張に用いられているためであると考えられる．

5. おわりに

本稿では，検索質問拡張の手法と自動構築されたシソーラスの評価方法を検討し，用語検索閲覧履歴からのシソーラス自動構築手法の改良を行った．また，改良手法に対する評価も行った．用語検索閲覧履歴からのシソーラス自動構築において,個々の Web ページのサイズのばらつきを考慮することで，構築されるシソーラスの精度向上が期待できる．本稿では，ユーザの用語検索履歴からのシソーラス自動構築を行い，ユーザの閲覧済み Web ページの再検索における検索質問拡張を想定して，自動構築したシソーラスの評価を行った．検索対象を再閲覧の Web ページだけでなく，ユーザにとって未知の Web ページの検索する際の検索質問拡張を検討していくことが今後の課題である．

文　献

[1] 安川美智子，山田篤，星野寛，大瀬戸豪志，上林彌彦,"Web コンテンツの収集と再利用を支援する個人用アーカイブシステム"，情処研報 No. 2002-DBS129-18. 2003 年.

[2] 安川美智子，山田篤，星野寛，大瀬戸豪志，上林彌彦,"Web 検索エンジンに対する検索語の類似度に基づく関連文書の検索", FIT2002, D-8, pp.15-16, 2002 年.

[3] 安川美智子，山田 篤，"Web 検索エンジンを用いた用語検索履歴からのシソーラス自動構築"，日本データベース学会 Letters Vol.3, No.1, pp.105-108, 2004 年.

[4] 安川 美智子，山田 篤，"Web 閲覧履歴に基づくシソーラス自動構築", DEWS2004, I-2:Web 検索(2), 2004 年.

[5] Hinrich Schutze, Jan O. Pedersen, "A Cooccurrence -Based Thesaurus and Two Applications to Information Retrieval", Inf. Process. Manage. 33(3): 307-318 , 1997.

[6] Zheng Chen, Shengping Liu, Liu Wenyin, Geguang Pu, Wei-Ying Ma, "Building a web thesaurus from web link structure", SIGIR 2003, pp.48-55, 2003.

[7] Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka, "Combining multiple evidence from different types of thesaurus for query expansion", SIGIR, 1999.

[8] Pavel Braslavski, Gleb Alshanski, Anton Shishkin, "ProThes: thesaurus-based meta-search engine for a specific application domain", WWW2004, pp.222-223, 2004.

[9] 相澤 彰子，影浦 峡，"著者キーワード中での共起に基づく専門用語間の関連度計算法"，信学会論，Vol.J83-D1 No.11 pp.1154-1162 2000 年.

[10] 松尾豊，石塚満，"語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム"，人工知能学会論文誌, Vol.17, No.3, pp.217-223, 2003 年.

[11] 松尾 豊，福田 隼人，石塚 満，"ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援"，人工知能学会論文誌，Vol.18, No.4, pp.203-211, 2003 年.

[12] 大澤 幸生，ネルス E. ベンソン，谷内田 正彦，"KeyGraph：語の共起グラフの分割・統合によるキーワード抽出"，信学会論文，Vol.J82-D1 No.2 pp.391-400 1999 年.

[13] Sen Yoshida, Takashi Yukawa, Kazuhiro Kuwabara, "Constructing and Examining Personalized Cooccurrence-based Thesauri on Web Pages", WWW2003, 2003 年.

[14] Roget, P. M., Kirkpatrick, B., "Roget's thesaurus, New ed.", Penguin, 1998.

[15] 池原悟，"分類語彙表 1.意味体系"，岩波書店，1997 年.

[16] WordNet, "a lexical database for the English language", http://www.cogsci.princeton.edu/~wn/

[17] Y. Jing and W. Bruce Croft, "An Association Thesaurus for Information Retrieval", RIAO94, pp.146-160, 1994.

[18] 国立国語研究所編，"分類語彙表 増補改訂版"，大日本図書, 2004 年.