Web 閲覧履歴に基づくシソーラス自動構築

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 Web 上の情報を効果的，効率的に閲覧し，収集するためには，一度閲覧した Web ページを保存（アーカイブ）しておき，必要に応じて後で検索し，再閲覧できることが求められる．しかし，アーカイブされた，閲覧済みの Web ページを，後で再閲覧したいときに，目的とする Web ページを正確に検索できるキーワードを思いつくことは容易ではない．そこで，本論文では，ユーザが閲覧した Web ページ中のテキストデータをもとにシソーラスを自動構築し，構築したシソーラスを用いた検索質問拡張を行うことで，閲覧済み Web ページの検索を支援することを提案する．Web 検索エンジンを用いた用語検索の履歴をもとに，提案手法により自動構築したシソーラスの例についても報告する．

キーワード シソーラス，閲覧履歴，個人用アーカイブシステム

Automatic Thesaurus Construction Using Web Browsing History

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 For effective and efficient browsing and collection of information on the web, it is needed that web pages which were browsed before are archived and easy to be re-browsed. However, precise words, which can search archived pages in need of a user, are not easy to hit upon the user's idea. In this paper, we propose a method of automatic thesaurus construction using web pages browsed by user. Such a thesaurus is helpful for query expansion in searching archived web pages. We also illustrate constructed thesauri with an example.

Keyword Thesaurus, Web Browsing History, Personal Archiving System

1. はじめに

近年，あらゆる情報が Web 上で提供されるようになってきており，Web 上で提供される情報の効率的な検索,収集,閲覧に対する要求はますます高まっている．著者らは，これまでに，Web 上の情報の収集と閲覧を支援する個人用アーカイブシステム[1]を提案してきた．個人用アーカイブシステムは，WWW キャッシュの原理に基づくアーカイブ用のプロキシ（アーカイブプロキシ）を用いて，ユーザが閲覧した Web ページの複製を蓄積するシステムである．アーカイブデータとして蓄積されている，既に閲覧済みの Web ページ（以下，閲覧済み Web ページと呼ぶ）を，ユーザが効率よく再閲覧できるようにするためには，Web ページのカテゴライズやフィルタリング，検索などの Web ページに対するアクセス手段を提供することが必要となる．

我々はこれまでに，ユーザが一度閲覧した Web ページを後で効率よく再閲覧できるようにすることを目的として Web 検索エンジンに対する検索語の類似度を用いた Web ページの関連付け手法[2]を提案してきた．Web ページのカテゴライズ，フィルタリング，検索を含む，より高度な Web ページの閲覧支援を可能とするためには，キーワードの関連語リスト，すなわち，広義のシソーラスが有用であると考えられる．そこで，本論文ではユーザの閲覧済み Web ページをもとにしたシソーラス自動構築（図 1）を提案する．

一般に，ユーザが Web ページの検索・閲覧を行う理由や目的，ユーザの閲覧済み Web ページの内容はさまざまである．本論文では，閲覧済み Web ページの中でも特にユーザが興味を持ったある特定の主題についての Web ページを用いてシソーラスを自動構築し，閲覧済み Web ページを検索する際の検索質問拡張などに，自動構築したシソーラスを用いる，ということに焦点を当てる．

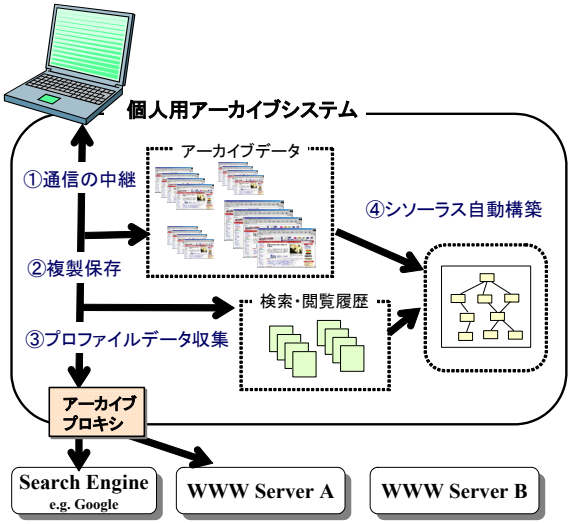

図 1 Web 閲覧履歴に基づくシソーラス自動構築

ユーザが閲覧済み Web ページを再閲覧したい，という要求は頻繁に発生するため，通常，Web ブラウザには，閲覧済み Web ページの履歴をリスト表示させる機能や，ブラウザのキャッシュに対して検索を行う機能が備えられている（たとえば，Internet Explorer の場合，閲覧済み Web ページを検索するには，[履歴] ボタンをクリックして表示される[履歴バー]の[検索]メニューをクリックする).

閲覧済み Web ページの数が少数で，閲覧時からの時間経過がわずかであれば，履歴の URL リストを一つずつ調べることや，全文検索や grep などを使って目的とする Web ページを見つけ出すことは容易である．しかし，Web ページ閲覧時からある程度の時間が経過し，他の多数の Web ページを閲覧した後では，閲覧したい Web ページの URL が履歴のリストに埋もれてしまい，また，目的の Web ページを特徴付ける，最も重要なキーワード（以下，「主キーワード」と呼ぶ）が思い出せないという事態も発生する．

そのような場合に，閲覧済み Web ページをもとに自動構築したシソーラスがあれば，主キーワードが思い出せない場合でも，閲覧済み Web ページに含まれる副次的なキーワード（「副キーワード」と呼ぶ）を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，目的とする Web ページに素早くアクセスできると考えられる．また，そのようなシソーラスは，閲覧済み Web ページを閲覧する際だけでなく，閲覧済み Web ページと類似の Web ページを新たに検索しようとする際にも役立つと考えられる．

本論文で想定しているシナリオは以下のようなものである．

【１】 ユーザが Web ページを見ていて，あるいは，オフラインの情報（テレビ，ラジオ，本や雑誌，広告など）から，気になる用語（主キーワード，Wp とする）を発見する．気になる用語とは，よく意味が分からない言葉や，もっと詳しく知りたい言葉などである．

【２】 その用語（主キーワード Wp） に関する Web ページを，Web 検索エンジン（たとえば Google 等）で検索する．

【３】 検索結果から，目的の Web ページ（主キーワード Wp に関するページ）を閲覧する．

（時間が経過する．他の Web ページを多数閲覧する．）

【４】 ブラウザの履歴検索を使って以前見た主キーワード Wp についての Web ページを検索しようとするが，主キーワード Wp を忘れてしまったので，代わりに，他に思いついたキーワード（副キーワード，Wq, Wr, Ws とする）を入力してみるが，うまく Wp に関するページが見つけられない．

【５】 ユーザの Web 閲覧履歴に基づき自動構築したシソーラスを使って，Wq, Wr, Ws の関連語 Wx, Wy, Wz を使った検索質問拡張を行い，検索結果の精度・再現率を向上させる．

なお，本論文で扱う範囲は，検索質問拡張に応用するためのシソーラス自動構築までとし，実際に検索質問拡張に応用することについては，議論しない．

以下，2 章で既存のシソーラス自動構築手法について述べ，3 章で Web 閲覧履歴に基づくシソーラス自動構築について述べる．また，3 章で述べた方法により自動構築したシソーラスの例を 4 章で説明し，最後に 5 章でまとめと今後の課題について述べる．

2. 関連研究

一般に，専門用語を集めて整理したものを専門用語集と呼ぶが，このうち特に，情報検索システムでの利用を想定し，専門用語をある特定の形式に整理したものをシソーラスと呼ぶ[3]．情報検索の分野では，クエリ処理を行う際に，クエリが短かすぎて検索漏れが発生するのを防ぐために，シソーラスを用いて検索質問拡張が行われる．たとえば，「 restaurants AND

(Mideastern OR vegetarian) AND inexpensive 」のようなクエリの場合，関連のある語のリスト，すなわちシソーラスを使って，Mideastern（中東）を特定の国名（たとえば，イラン，イラク，アフガニスタン）で拡張し，また，インドは中東とは見なされないが，vegetarian を拡張して，India をクエリに加えることで，インド料理も検索されるようにする，というテクニックが用いられる[4].

シソーラスを用いた検索質問拡張を行っているものとしては，[5][6][7]が提案されている．シソーラスは意味的に関連のある語をもとに人手で作成される他，ドキュメント集合の中の語の共起に基づいて，ほとんど，あるいは，全く人間の関与なしに自動生成することも可能である[4]．ドキュメント集合を用いたシソーラスの自動構築の研究としては，[8][9][10][11][12]などがある．また[13][14]や[15]では，テキストデータからのキーワード抽出を提案している．

テキストデータから語の共起情報を抽出する方法として，相互情報量を用いる手法[16], Kullback-Leibler divergence や Jensen-Shannon divergence を用いる手法[17]があり，これらの手法は，[13]などの研究で用いられている．また，LSA(潜在的意味分析)を用いてシソーラスを自動構築する手法[18]が提案されており，[19]では，この手法を採用している．

3. 提案手法

上で述べたシソーラス自動構築の確立された基本的手法([16], [18])を参考に，以下の 3 つの手法により閲覧済み Web ページからのシソーラス自動構築を試みる．シソーラス自動構築の基本的手法は，言語処理や情報検索の分野で確立されたものであるが，これを閲覧済み Web ページを用いたシソーラス自動構築に応用するには，データスパースネスの問題に加えて，Web ページが学術論文や新聞記事などと比べて不均質であることを考慮する必要がある．

（１）語の直接共起を用いる手法

同一文中での語の共起頻度から，相互情報量を求め，互いに共起関係にある語を抽出する[16]．語 T_i と語 T_j の相互情報量 $M(T_i,T_j)$ は，以下のように計算する．

$$M(T_i,T_j)=\log\frac{N\cdot freq(T_i,T_j)}{freq(T_i)\cdot freq(T_j)}$$

ただし，N は Web ページ中の語の異なり語数であり，$freq(T_i,T_j)$ は語 T_i と語 T_j の共起頻度，$freq(T_i)$ と $freq(T_j)$ は，それぞれ語 T_i と語 T_j の出現頻度である．

相互情報量が閾値 P を越えるものを関連語として抽出する．P は，$P=k\times\max(M(T_i,T_j))$ で計算する．k は定数である．また，共起頻度にも閾値を設定し，$freq(T_i,T_j)$ が有効共起回数 C に満たないものは，無効と考え，除外することとした．これは，特に Web ページでは相互情報量が高くても共起頻度が低いものは信頼性が低いと考えられるためである．

（２）語の間接共起を用いる手法

同一文では共起しないが，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関連語として抽出する．間接共起は，上記の（１）の直接共起と同様の計算式により相互情報量を計算し，$M(T_i,T_p)$ と $M(T_j,T_p)$ がともに閾値 P を越えるものを関連語とする．ここで，T_p は語 T_i と語 T_j を結び付けている語である．間接共起の場合も，共起回数が有効共起回数 C に満たないものは除外する．また媒介語の個数が閾値 E に満たないものも除外する．E は定数である．

（３）潜在的意味分析（LSA）を用いる手法

同一文中での語の共起頻度から，共起行列を作成し，共起行列を特異値分解（SVD）し[18]，得られた特徴ベクトルの語 T_i に対応する行と，語 T_j に対応する行の類似度（尺度として余弦(cosine)を使用する）を計算し，類似度 $\cos(V_i,V_j)$ が閾値 Q を越えるものを関連語として抽出する．Q は，$Q=s\times\max(\cos(V_i,V_j))$ で計算する．s は定数である．

後述のシソーラス自動構築の例では，定数値として $C=2$，$E=2$，$k=0.8$，$s=0.3$ を用いた．

Web ページに対する前処理として語の分割処理には，Microsoft Windows2000 以降の基本サービスとなっている Indexing Service[20]の Japanese Word Breaker を使用した．Word Breaker は語の分割処理の後，予め定義されている言語毎のストップワードの除去も自動的に行う．

間接共起や潜在的意味分析を行う上で，それ自体は

意味のない語であっても，関連語を抽出する上で役立つ場合もあることから，無闇に語を除去することは望ましくないが，Word Breaker のストップワードとして定義されていない一部の指示代名詞（「これら」「それら」「こんな」「そんな」等）と，半角英数字 1 文字，ひらがな 1 文字の語，及び，「copyrights」「All」「Rights」「reserved」等の Web ページの著作権表示に使われる語は，シソーラス自動構築の精度を低下させるため，除外する．

潜在的意味分析で用いる特異値分解のためのアルゴリズムと実装は種々提供されているが，後述のシソーラス自動構築の例では，S-Plus[21]を使用した．

4. Web 閲覧履歴に基づくシソーラス自動構築の例

上に述べた手法により自動構築したシソーラスの例を表 1～表 10 に示す．検索閲覧履歴としては，Web 検索エンジンを用いた用語検索の履歴を用いた．検索対象の用語として「現代用語の基礎知識 1991-2003 年版」から選んだ用語を Google で検索し，検索結果から Web ページを 1 つ選んでアーカイブデータとして保存し，これを閲覧済み Web ページとして用いることとした．

直接	{音，警告}{WIRED，日本，最新，NEWS，最，先，...}{歩行，子ども，検知，NHTSA，研究，安全，...}{従来，動体，動，ワイス}{消耗，品}{透視，超音波，バンパー，プラスチック，製，...}{利用，改良}{物体，運転}{型，オプション}{DVD，ソフト} ...
間接	{WIRED，ウェブセキュリティ，ウェブページ，...}{従来，タイミング，赤外線，内部，透視，...}{研究，会，機関，交通，高速，...}{超音波，動，バックアップ，フォード，搭載，...}{NTT，Digital，Translations，other，portions，...} ...
潜在	{彼ら，容易}{以下，300 ドル，販売}{1960 年代，存在}{companies，affiliated}{以上，20000 点}{常に，ぬぐ，泥}{背後，横，障害，物}{映画，放送，情報}{作品，主演，検索，監督}{消耗，品，サーチ，プリンタ}{従来，動体}{反射，対象，利用} ...

表 1 用語「**車載レーダー**〔現代工学用語〕」の用語検索履歴からのシソーラス[1]

直接	{暑さ，蒸し暑，感じ，不快，指数，...}{高，夏場，低，夏}{維持，水，加湿，器，使用，...}{数値，訴え，示}{相当，感覚，私たち，例え}{変化，放射}{指数，不快，アメリカ}{乾燥，皮膚，生息，問題，環境，...}{リクガメ，飼育，皮膚，生息，発生}
間接	{温度，体感，気候，甲羅，左右され，...}{アルミニウム，カルキ，空中，残余，供給，...}{感じ，夏}{アメリカ，humidity，temperature，気象，局，...}{蒸し暑，植物，心理，新陳代謝，性別，...}{生息，低}
潜在	{空間，居住，系，局，気象，...}{インフォメーション，ウイルス，供給，期間，空中，...}{表現，例えば，深，生活}{供給，局}{相応，遅れ，微妙，数時間，大幅}{水源，方法，注意，効果，最も}{速度，風速，量}{見，不足，チェック，カサカサ，四肢} ...

表 2 用語「**体感温度**〔気象用語〕」の用語検索履歴からのシソーラス[2]

直接	{写真，スナップ，撮，エポック，Epoch，...}{Ascending，Ascension，Node}{双曲線，曲線，物，放，ケプラー，...}{32，刻，間，フェーズ，休止}{大気，らせん，状，降下，残留，...}{Inclination，Orbital}{変化，抗力}{補正，摂動}{力，働，重力} ...
間接	{残留，N1，抵抗，引き起こ，降下，...}{バーン，Attitude，座標，構成，系}{エポック，T0，Time，写真，スナップ，...}{突き出，完全に，1 本，2 個所，Nodes}{参考，資料，Translation，2 次，第 1，...}{32，1.5，刻，間，12 時間，240}...
潜在	{要素，軌道，計算，番号，面，...}{指定，指}{RAAN，RA}{速度，速}{中心，地球}{天文，天，家}{遠，地点，近，近づ，180 度}{角，角度，傾斜，傾}{交点，昇，交点，衛星，赤道}{高度，高，10}{経，赤，経度，緯度}{運動，平均}{正確，正}...

表 3 用語「**軌道要素**〔宇宙開発用語〕」の用語検索履歴からのシソーラス[3]

[1] *http://www.hotwired.co.jp/news/news/technology/story/2957.html 車載レーダーで事故防止* を閲覧した

[2] *http://www001.upp.so-net.ne.jp/tortoise/wagayanokamechan_045.htm 体感温度と不快指数* を閲覧した

[3] *http://www.jamsat.or.jp/keps/kepmodel.html*

直接	{以来, 開発}{進行, 行}{結晶, 低, LDPE, 度, 引っ張り}{圧, Ziegler, Natta, 系, 触媒}{反応, 必要}{重合, 重}{分子, 性}{ポリマ, ポリマー}{化学, 見}{製造, 方法, 圧}
間接	{圧, 異なり常, HPDE, Ziegler, Natta, ...}{結晶, 密度, 違い, 分類, 引っ張り, ...}{度, 枝, 弱, 徴, 長, ...}{LDPE, 規則正, 強, 区, 不透明, ...}{以来, 進行, 行, 開発}
潜在	{2000554, 200053691}{進行, 行}{Station, WhatsNew}{それぞれ, 特徴}{常, 用い}{25, by, ブレビコミン}{HPDE, であ, 異なり常}{下, 法, 10～20MPa}{ひと, 知, lett}{枝, 弱}

表 4 用語「**メタロセン触媒**〔新素材用語〕」
の用語検索履歴からのシソーラス[4]

直接	{産卵, アカウミガメ}{美, 風紋, 広が, 風, 田島}{埋立, 1972 年, 埋立て, 最近, 出, ...}
間接	{田島, 最近, 市街, 線, 大量, ...}{埋立, 捨て場, 通報, 異物, 過去, ...}{出来事, 役所, いい加減, 市民, 出現, ...}{約, 1 部, 波, 30 年, 65000 ㎡, ...}
潜在	{全国, 変わ}{当時, 離れ}{次第, 天候}{廃棄, 埋め立て}{処分場, 民間, 大騒ぎ}{2003.12.05, 地球, ゴミ}{堤防, 改修, 影響}{10 年間, 前, 燃え, およそ, いい加減, ...}{砂浜, 市民, 80m, 削}{実施, 打ち上げ, 覆, 防止}{流れ, 露出, 島, 層状, 中田} ...

表 5 用語「**海岸浸食**〔海洋開発用語〕」
の用語検索履歴からのシソーラス[5]

直接	{活動, 火山, 動, 地下, 観測}{山, 富士, 震源}{観測, 地下, 付近, 発生, 震源, ...}{数, 10}
間接	{発生, 付近, 平成, 12 年, 12 月, ...}{観測, 地下, 変化, 活発, 地殻}{活動, マグマ, 火口, 起き, しばしば, ...}
潜在	{地震, 周波, 低, 発生, 富士, ...}

表 6 用語「**低周波地震**〔地震・火山用語〕」
の用語検索履歴からのシソーラス[6]

直接	{院, 泰, 門, 梨, 東, ...}{奨, 忠, 洞}{人, 訪問, 外国, 観光, 客}{利用, 水, バルネオセラピー}{宿泊, 交通, コミ}{美容, 多, 店, 韓国, 観光, ...}{ショッピング, 施設, 多}
間接	{ショッピングタウン, タワー, バス, ミリオーレ, 大型, ...}{10 回, バス, 1 日, 運行, 往復, ...}{憩い, 多, 観光, 客, 最近}{水, 神経, 神経痛, 専門, Hydrotherapy, ...}{宿泊, 施設, 文化, 旅行, クーポンブック}{忠, 南山, ふもと, ジャンチュンドン, ソウル, ...}...
潜在	{記者, 任}{読者, 評価, ソウル}{方法, 力}{06:50, 21:10}{Interactive, ZIO, Inc}{比較的, 皮膚, 訪れ}{素敵, 聞, 心}{バス, 1 日, デラックス}{便利, 法, 林, 名前}{何, Balneo, Therapy, 自慢, フィットネスクラブ}{最近, 多様, 趣向, 憩い, 場所} ...

表 7 用語「**バルネオセラピー**〔美容一般用語〕」
の用語検索履歴からのシソーラス[7]

直接	{発芽, ジャガイモ, 防止}{出, 装置, 物質, 高, 物}{栄養, 毒性}{利用, 利}{照射, 線, 考え, 保存, 用}
間接	{照射, 長期, 目的, 軍需, 歴史, ...}{出, エネルギー, 殺虫, 止め, コバルト, ...}{装置, 電子, 高め, 作物, エネルギー, ...}{防止, 商品, タマネギ, 組織, 部分, ...}{発芽, 影響, 細胞, 利用, 最初, ...}
潜在	{線, 照射, 放射}{多, 方面, 力}{短, 延長, 非常, 破壊, 多}{生殖, 組織, 他, おこな, 総合}{価値, 部分, 落と, 歴史, 方法, ...}{長期, 目的, 歴史, 軍需}{不安, 方法, もの, 知, 大規模, ...}{発生, 飛び出, 能力, 粒子, 強}{変化, 器官} ...

表 8 用語「**放射線食品**〔食生活用語〕」
の用語検索履歴からのシソーラス[8]

直接	{初, 影響, 語}{テレビ, 母親}
間接	{初, 言葉, 意味, 異変, 最初, ...}
潜在	{語, 初, 影響, 乳幼児, 最近, ...}

表 9 用語「**初語異変**〔育児用語〕」
の用語検索履歴からのシソーラス[9]

軌道要素の解説 を閲覧した
[4] *http://www.chem-station.com/yukitopics/prieti.htm ポリエチレンの作り方* を閲覧した
[5] *http://tyousakai.hp.infoseek.co.jp/hamamatu.htm 海岸浸食で「ゴミ」が出てきた！* を閲覧した
[6] *http://www1.pref.yamanashi.jp/bousai/MtFuji/teishuuha.htm 低周波地震* を閲覧した
[7] *http://www.koreanavi.com/travel/hotel/20000222-3.html KoreaNavi - 旅行 - 宿泊・交通* を閲覧した
[8] *http://eiyougaku.hp.infoseek.co.jp/housya.htm 放射線照射食品とは* を閲覧した
[9] *http://www.tnt-net.co.jp/kosodate/009/3/009.html 初語異変* を閲覧した

直接	{助成，診断，耐震，木造}{安全，性}
間接	{助成，対象，限度，制度，100000 円，...}
潜在	{住宅，安全，性，助成，木造，...}{強，個人}{進め，づくり，設け，生活}

表 10 用語「**耐震診断助成制度**〔住生活用語〕」の用語検索履歴からのシソーラス[10]

考察

表 1～表 10 のシソーラス自動構築の例から，直接共起は，Web ページの文章から直接共起を抽出するため，それぞれの Web ページの主題と関連のあるシソーラス自動構築が行えていると言える．

間接共起は，直接共起からだけでは得られない，より広義の関連語の抽出が行えている．潜在的意味索引は，特に語の延べ総数が少ない Web ページでは，潜在的な意味をとらえることが難しく，人間が直感的に関連をとらえにくい語も含まれている．

Web 検索エンジンを用いた用語検索で，検索結果として得られる Web ページは，以下のような特徴がある．このため，一般的な Web ページをテキストデータとして用いる場合と比較して，比較的精度の良いシソーラス自動構築が行える．

- 説明する，あるいは，論じるという形式の文が多く含まれた，論文や新聞記事に似たテキストデータであり，日本語も正確で，テキストからの語の抽出の際の失敗が少ない．
- Web ページ中に含まれる文の数が多いのに対して，Web ページ中で述べられているトピックの数は絞り込まれており，トピックを説明する上でキーワードになる語が効果的に使用されている場合が多い．

ただし，中には，シソーラスを構築するためのテキストデータとして適切とはいえないものもあり，表 9，表 10 の例では，シソーラス構築に用いた Web ページが上記の特徴を満たさないものであったため，関連語の抽出が適切に行われているとは言えない結果となっている．

5. まとめ

本論文では，ユーザの閲覧済み Web ページ中のテキストデータを用いて，シソーラスを自動構築する手法を提案した．また，提案手法を用いて実際にシソーラス自動構築を行った．その結果，閲覧済み Web ページから，ある程度，人間の直感にも合致する関連語のリストを抽出できることが分かった．自動構築したシソーラスを用いた検索質問拡張などのアプリケーションを検討することが今後の課題である．

[10] *http://www.city.tama.tokyo.jp/life/jutaku/mokuzo.htm 木造住宅耐震診断助成制度* を閲覧した

謝辞

本研究の一部は，(財)群馬大学科学技術振興会の支援を受けて行われた．

文　献

[1] 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，安川美智子，山田篤，星野寛，大瀬戸豪志，上林彌彦，情処研報 No. 2002-DBS129-18. 2003 年.

[2] 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，安川美智子，山田篤，星野寛，大瀬戸豪志，上林彌彦，FIT2002，D-8，pp.15-16，2002 年.

[3] 岩波講座マルチメディア情報学　情報の組織化，長尾他，岩波書店，2000

[4] Information Storage and Retrieval, Robert R. Korfhage, John Wiley & Sons Inc, 1997

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

[6] Query expansion using lexical-semantic relations, Ellen M. Voorhees, SIGIR, 1994

[7] Query expansion using domain-adapted, weighted thesaurus in an extended Boolean model, Oh-Woog Kwon, Myoung-Cheol Kim, Key-Sun Choi, CIKM, 1994

[8] Automatic Detection of Thesaurus Relations for Information Retrieval Applications, Gerda Ruge, LNCS-1337, 1997

[9] Refining The Selectivity Of Thesauri By Means Of Statistical Analysis, Erich Schweighofer, Werner Winiwarter, Terminology and Knowledge Engineering, 1993

[10] Automatic Word Sense Discrimination, Hinrich Schutze, Computational Linguistics, 1998

[11] Information Retrieval Based on Word Senses, Hinrich Schutze and Jan O. Pedersen, SDAIR, 1995

[12] 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，相澤 彰子，影浦 峡，信学会論，Vol.J83-D1 No.11 pp.1154-1162 2000 年 11 月

[13] "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"，松尾豊，石塚満，人工知能学会論文誌，Vol.17，No.3，pp.217-223

[14] 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，松尾 豊，福田 隼人，石塚 満，人工知能学会論文誌，Vol.18，No.4，pp.203-211

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

[16] Word association norms, mutual information and

lexicography, Kenneth Ward Church and Patrick Hanks. , Association of Computational Linguistics, pp.76-82, 1989.

[17] Similarity-Based Models of Word Cooccurrence Probabilities, Ido Dagan, Lillian Lee, and Fernando Pereira. Machine Learning 34(1-3), pp 43--69, 1999.

[18] Hinrich Schutze: Dimensions of Meaning. SC 1992: pp.787-796

[19] Constructing and Examining Personalized Cooccurrence-based Thesauri on Web Pages, Sen Yoshida, Takashi Yukawa, Kazuhiro Kuwabara, WWW2003

[20] Indexing Service Version 3.0
http://msdn.microsoft.com/library/en-us/dnanchor/html/indexserv.asp

[21] S-plus, http://www.msi.co.jp/splus/