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

Automatic Thesaurus Construction from History of Terminology Search using Web Search Engine

安川 美智子† 山田 篤‡
Michiko YASUKAWA Atsushi YAMADA

Web 上で提供される情報の効率的な検索，閲覧，収集に対する要求が高まっている．ユーザが，効率よく Web 上の情報を検索し，閲覧・収集できるようにするためには，シソーラスを用いた検索質問拡張が有効であると考えられるが，このような検索質問拡張を実現するためには，検索質問拡張に利用可能なシソーラスを用意しておくことが必要となる．従来提案されているシソーラス自動構築手法は，純粋なテキストデータをもとにしているものがほとんどであり，Web 情報検索の特徴を十分に考慮しているとは言えない．そこで，本論文では，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閲覧履歴，及び，閲覧済み Web ページからシソーラスを自動構築することを提案する．提案手法を用いて自動構築したシソーラスの例についても報告する．

For effectiveness and efficiency of search, browse and collection of information on the web, query expansion using thesauri is worth investigating. Although useful and excellent thesauri for such use are required to be pre-constructed, traditional ways of automatic thesauri construction have been based on pure text and have not deliberated on characteristics of web search. In this paper, we propose a method of automatic thesaurus construction from a user's browsing history and browsed web pages of terminology search using web search engine. We also illustrate constructed thesauri by our proposed method with an example.

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．

我々は，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を提案してきた[1]．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．

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．

そこで本論文では，検索質問拡張などの検索・閲覧支援に応用するためのシソーラスを自動構築する手法を提案する．以下，2 章で関連研究について述べ，3 章で提案手法である，Web 検索エンジンを用いた用語検索履歴に基づくシソーラス自動構築について述べる．また，自動構築したシソーラスの例を 4 章で示し，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

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．

情報検索の分野では，検索の精度と再現率を向上させる目的で，シソーラスを用いた検索質問拡張が行われる．シソーラスは人手で構築すると手間が大きく，保守や維持に時間と費用がかかるという問題がある．このため，コンピュータを用いたシソーラス自動構築の手法が提案されている．

純粋なテキストデータからのシソーラス自動構築に関する研究としては，[2][3][4]などが提案されている． 従来のシソーラス自動構築に関する研究では，純粋なテキストデータを元データとしているものがほとんどであるが，最近では，純粋なテキストデータだけでなく，Web のリンク構造を利用したシソーラス自動構築手法も提案されている[5]．本論文で提案するのは，個々のユーザが Web サーチエンジンを利用して用語検索を行った際の検索・閲覧履歴をもとに，シソーラスを自動構築する手法である．

3. 提案手法

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り，このような雑多な Web 閲覧履歴を元データとしてシソーラスを構築すると精度の良いシソーラスが得られない．そこで本論文では，ユーザが Web 検索エンジンを用いて，用語検索

† 正会員 群馬大学工学部情報工学科
michi@cs.gunma-u.ac.jp
‡ 財団法人京都高度技術研究所
yamada@astem.or.jp

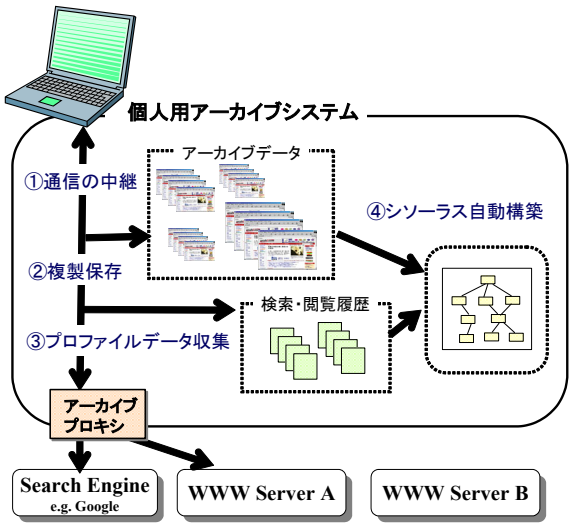

図 1 用語検索履歴に基づくシソーラス自動構築
Fig.1 Thesaurus Construction from History of Terminology Search

を行った際の Web 検索・閲覧履歴，及び，閲覧済み Web ページからシソーラスの自動構築を行う手法を提案する（図 1）．

ユーザが行う，Web 検索エンジンを用いた用語検索は以下のように特徴付けられる．

① ユーザは 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，そのユーザにとって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 ユーザは Web 検索エンジン（たとえば Google 等）を用いて，その用語（主キーワード Wp） に関する Web ページを検索する．
③ 検索エンジンが返した検索結果 Web ページから，ユーザは，主キーワード Wp に関連すると思われる Web ページを 1 つまたは複数，選択して，閲覧する．

提案するシソーラス自動構築手法では，上記の③で，ユーザが閲覧した 1 つまたは複数の検索結果 Web ページから構成される，用語説明の Web ページ集合からテキストデータを抽出し，シソーラス自動構築の元データとして利用する．

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[6]，Kullback-Leibler divergence や Jensen-Shannon divergence を用いる手法[7]があり，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[8]が提案されている．提案するシソーラス自動構築手法では，相互情報量と LSA(潜在的意味分析)を用いて，以下のような 3 つの手法により，テキストデータからの関連語抽出を行う．

（１）語の直接共起を用いる手法

用語説明の単一 Web ページ中に含まれる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．語 Ti と語 Tj の相互情報量 $M(Ti,Tj)$ は，以下のように計算する．

$$M(Ti,Tj)=\log\frac{N\cdot freq(Ti,Tj)}{freq(Ti)\cdot freq(Tj)}$$

ただし，N は Web ページ中の語の異なり語数であり，$freq(Ti,Tj)$ は語 Ti と語 Tj の共起頻度，$freq(Ti)$ と $freq(Tj)$ は，それぞれ語 Ti と語 Tj 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，$P=k\times\max(M(Ti,Tj))$ 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，$freq(Ti,Tj)$ 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，$M(Ti,Tp)$ と $M(Tj,Tp)$ がともに閾値 P を越えるものを関連語とする．ここで，Tp は語 Ti と語 Tj 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，得られた特徴ベクトルの語 Ti に対応する行と，語 Tj 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 $R=\cos(Vi,Vj)$ が閾値 Q を越えるものを関連語として抽出する． Q は，$Q=s\times\max(R)$ で計算する． s は定数である．

上記の(1)～(3)により関連語を抽出した後，関連度（直接共起と間接共起の場合は相互情報量の値，潜在的意味分析を用いる場合は特徴ベクトルの余弦から得られる類似度の値）の高い順に，以下のような手順で関連語のクラスタリングを行う．

クラスタリングの手順

i. 既存のクラスタの中に共通の語を含むクラスタがなければ新規クラスタを生成する．
ii. 既存のクラスタの中に共通の語が含まれるクラスタがあれば，そのクラスタに語をマージする．
iii. 既存のクラスタに共通の語を含むクラスタが複数存在する場合は，関連度の高い語を含むクラスタに語をマージする．

用語説明の Web ページ集合が，1 つの Web ページではなく，複数の Web ページから構成される場合（すなわち，ユーザが検索結果から複数の Web ページを選択し，閲覧した場合）は，個々の Web ページについて関連語の抽出とクラスタリングを行い，クラスタリングの結果をマージする．

4. 用語検索履歴からのシソーラス構築の例

提案手法を用いて，以下のような条件で，シソーラスの自動構築を実験的に行った． Web 検索エンジンで検索する検索対象用語として「現代用語の基礎知識 1991-2003 年版」から，ランダムに選択した用語を使用した．

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
Table 1 Thesaurus Constructed from Terminology Search History of "automotive-radar"

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 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
Table 2 Thesaurus Constructed from Terminology Search History of "sensible-temperature"

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 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
Table 3 Thesaurus Constructed from Terminology Search History of "orbital-element"

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 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
Table 4 Thesaurus Constructed from Terminology Search History of "metallocene"

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 5 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
Table 5 Thesaurus Constructed from Terminology Search History of "low-frequency-earthquake"

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, ...}

選択された用語（たとえば「車載レーダー」など）を Web 検索エンジンの検索キーワードとして入力し，検索を行う．Web 検索エンジンから返される検索結果から，検索結果のタイトル下のテキスト（検索結果 Web ページ中のキーワードが一致した部分を抜粋したテキスト）をもとに，用語に関連する 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．Web 検索エンジンには Google（http://www.google.co.jp/）を使用した．

Web ページから抽出したテキストデータ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[9]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，除外した．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，本論文におけるシソーラス自動構築の例では，S-Plus[10]を使用した．3 章で述べた関連度計算における定数値は，$C = 2$，$E = 2$，$k = 0.8$，$s = 0.3$ とした．

上記により自動構築したシソーラスの例を表 1～表 5 に示す．表 1～表 5 は，一つの表が一つの用語検索履歴から構築されるシソーラスに対応しており，直接，間接，潜在はそれ

ぞれ，3 章の「(1)語の直接共起を用いる手法」，「(2)語の間接共起を用いる手法」，「(3)潜在的意味分析」を用いる手法である．{ }でくくられている語のリストが関連語のクラスタであり，関連度の高いものから順に語を記述している．クラスタ内の語数，クラスタ数が多数あるものについては関連度の高いもののみ記述し，関連度の低いものを省略している（「...」と表記）．

考察

表 1～表 5 のシソーラスの例は，実験的に自動構築したものであり，シソーラス構築の元データとなる Web ページ集合は単一 Web ページのみから構成されている．しかし，実際に個々のユーザの用語検索履歴を用いてシソーラス自動構築を行う場合には，Web ページ集合は，それぞれのユーザの興味や背景知識によって，選択され，閲覧されるため，ユーザ毎に異なってくるものと考えられる．

表 1～表 5 より，同じ元データからのシソーラス構築であっても，シソーラス構築手法で用いる語の共起関係の捉え方が異なると，生成される語のクラスタは異なる，ということが分かる．直接共起は，Web ページから直接的な共起関係を取り出すため，Web ページの主題と関連のあるシソーラス構築が行えていると言える．また間接共起，潜在的意味分析は，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われていると言える．

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．

- 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．

このため，Web 検索・履歴全般を対象とするのではなく，Web 検索エンジンを用いた用語検索の閲覧履歴，及び，閲覧 Web ページのみに元データを限定してシソーラス自動構築を行うことで比較的精度の良いシソーラスが得られると考えられる．ただし，中には，元データとして適切とはいえない場合もあり，表 5 の例では，元データとなる 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．

提案手法により構築されるシソーラスのクラスタに含まれる語は，元データとなる Web ページに依存し，クラスタの性質は，語の共起関係の捉え方（すなわち，関連語抽出の手法）に依存する．また，クラスタの集合であるシソーラスは，個々のユーザがどのような Web ページを選択し，閲覧するかに依存する．シソーラスの精度や有効性は，シソーラスを用いた検索質問拡張を行うことで検索のパフォーマンスがどの程度向上するかにより，検証される．Mandala らは，性質の異なる複数のシソーラスを組み合わせることで，情報検索のパフォーマンスを向上させる手法を提案している[11]．我々の提案手法により構築される，直接，間接，潜在の 3 つの異なるタイプのシソーラスを組み合わせて，検索質問拡張を行うことにより， Web 情報検索のパフォーマンスを向上させていくことが今後の課題である．

5. まとめ

本論文では，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検索・閲覧履歴，及び，閲覧 Web ページから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の自動構築を行った．提案手法により，個々のユーザの用語検索履歴から，個々のユーザ毎のシソーラス自動構築を行うことができる．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．

[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] 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.
[2] Gerda Ruge: Automatic Detection of Thesaurus Relations for Information Retrieval Applications, LNCS-1337, 1997.
[3] Erich Schweighofer, Werner Winiwarter: Refining The Selectivity Of Thesauri By Means Of Statistical Analysis, Terminology and Knowledge Engineering, 1993.
[4] Hinrich Schutze: Automatic Word Sense Discrimination, Computational Linguistics, 1998.
[5] Zheng Chen, Shengping Liu, Liu Wenyin, Geguang Pu, Wei-Ying Ma: Building a web thesaurus from web link structure. SIGIR, 2003.
[6] Kenneth Ward Church and Patrick Hanks: Word association norms, mutual information and lexicography, Association of Computational Linguistics, pp.76-82, 1989.
[7] Ido Dagan, Lillian Lee, and Fernando Pereira: Similarity-Based Models of Word Cooccurrence Probabilities, Machine Learning 34(1-3), pp 43--69, 1999.
[8] Hinrich Schutze: Dimensions of Meaning. Proceedings of Supercomputing, pp.787-796, 1992.
[9] Indexing Service Version 3.0 http://msdn.microsoft.com/library/en-us/dnanchor/html/indexserv.asp
[10] S-plus, http://www.msi.co.jp/splus/
[11] Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka: Query expansion using heterogeneous thesauri. Inf. Process. Manage. 36(3): 361-378, 2000

安川 美智子 Michiko YASUKAWA

群馬大学工学部情報工学科助手．2003 京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．電子的著作権管理システム，Web 情報検索に関する研究に従事．ACM，IEEE-CS，情報処理学会，日本データベース学会各会員．

山田 篤 Atsushi YAMADA

財団法人京都高度技術研究所情報メディア研究室長．1991 京都大学大学院博士後期課程研究指導認定退学．1998 より，京都大学大学院情報学研究科客員助教授．博士（工学）．
言語処理系の研究に従事．情報処理学会，人工知能学会，認知科学会，言語処理学会，ソフトウェア科学会各会員．