F-033

# Web の半自動検索を利用した網羅的なテーマ関連知識習得支援方式

## Theme-oriented Comprehensive Knowledge Acquisition Using Semi-Auto Web Retrieval

服部 元† 武吉 朋也† 小野 智弘† 滝嶋 康弘†
Gen Hattori Tomoya Takeyoshi Chihiro Ono Yasuhiro Takishima

### 1. はじめに

近年，Web 上の情報が爆発的に増加しており，ユーザが必要とするテーマに関する知識を得るために，検索エンジンを利用した情報収集が必要不可欠となっている．代表的な知識を得る目的であれば，検索結果上位の Web ページを数件～十数件程度を閲覧することで満足できる．一方，例えば「おいしいカレーの作り方」のように，よりよい案を発見するために，材料や作業内容を比較する必要があるテーマの場合は，多くの Web ページを参照して網羅的に関連情報を収集する必要がある．しかしながら，検索結果の上位から順にユーザが諦めるまで閲覧する，あるいは検索キーワードを追加する等により検索結果を絞り込む等の，人手のみによる検索方法では情報の重複や抜けが生じる可能性が高い．特に数百万件もの検索結果が提示されるような場合には，網羅的な情報収集は困難である．既存のサービスとして，Wikipedia[1]等の辞書サイトや，はてなブックマーク[2]等のソーシャルブックマークサイトがあるが，ユーザが希望するテーマに合致しない場合には活用できない．

そこで本研究では，ユーザが網羅的に習得したいテーマに関連する情報を Web から短時間で収集することを支援する，テーマ関連知識習得支援方式を提案する．本方式はテーマに関連する Web ページを自動的に収集し，ユーザがまだ閲覧していない情報を多く含む Web ページを検索結果の上位に提示することで，短時間での網羅的な収集を可能にする．

以降，第 2 章では，本研究で想定する知識習得支援システムの概要と機能要件，および関連研究について述べる．第 3 章では，提案方式であるテーマ関連知識習得支援方式について述べ，第 4 章で評価実験とその結果について述べる．最後に第 5 章でまとめを述べる．

### 2. 知識習得支援システムの機能要件と関連研究

#### 2.1 システムの概要と機能要件

知識習得支援システムは，ユーザが設定したテーマに関連する情報の網羅的な収集を支援し，ユーザの知識習得に要する時間を削減することを目標とする．ここで「知識習得」とは，ユーザが設定したテーマに関連する情報を含む Web ページを収集・閲覧することと定義する．また「網羅的」とは，テーマに関連する全ての情報を内容の重複なく収集している状態と定義する．網羅的に情報収集するためには，テーマに関連する Web 情報全体を定義する必要があるが，これは困難である．本稿では，一定数のテーマに関連する情報を含む Web ページ群を起点として，関連する他の Web ページを繰り返したどりながら差分の情報を蓄積していくことで，テーマに関連する情報全体を収集することに近づくと仮定する．ここで，ユーザが設定するテーマは，例えば上述した「おいしいカレーの作り方」以外に，「お酒の飲み方」のように，一般的な知識やお酒の種類に応じた知識など多くの情報を幅広く集める必要があるテーマなどを想定する．

本研究で想定する知識習得支援システムを図 1 に示す．まず，ユーザがテーマ名を入力すると関連する Web ページの一覧が提示される．ユーザはそれらの Web ページを閲覧し，それらをテーマ名と共に習得済み情報として知識習得支援システムに登録する．次に知識習得支援システムは習得済み情報からテーマに関連する単語を推定し，Web 検索を行う．最後に，検索結果からテーマに無関係な情報と習得済みの情報と重複する情報を除外し，未習得情報としてユーザに提示する．表示および操作画面の例を図 2 に示す．テーマ名や網羅度(3.3 節(キ)で説明する)，検索結果などを表示する．

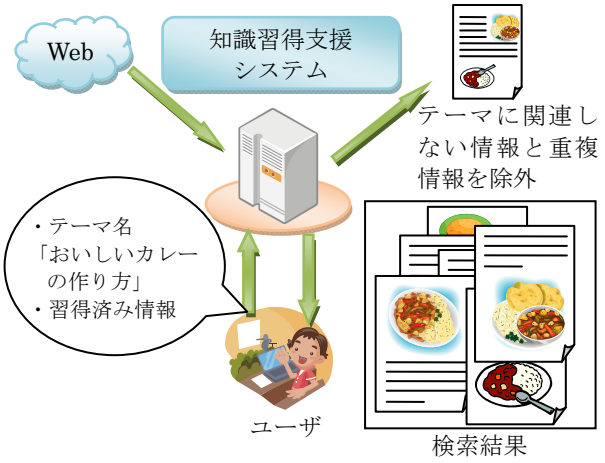


図 1 テーマ関連知識習得支援システムの利用シーン


図 2 検索結果の表示画面例

---

†（株）KDDI 研究所, KDDI R&D Laboratories Inc.

このようなテーマ関連知識習得支援システムを実現するためには，システムが Web 上からテーマに関連する情報を網羅的に収集できなければならない．ただし，ユーザが設定可能なテーマは無限に存在し得るため，事前に関連情報を収集しておくことは困難である．また，習得済み情報は手作業で登録するため，高々10 件程度の少ない数であると考えられ，さらにいずれもテーマに対する関連度が高く，出現する単語に偏りがある．このため単語の出現頻度のような統計的情報だけでは十分な情報収集の精度が出ない可能性が高い．よって，知識習得支援システムの機能要件として，以下の 2 つが挙げられる．

機能要件 1：習得済み情報の登録数が少ない場合であっても高精度な情報収集ができること

機能要件 2：任意に指定されたテーマの関連情報を，短時間で重複なく収集すること

### 2.2 関連研究

以下に，機能要件 1 および機能要件 2 に関する関連研究についてそれぞれ述べる．

機能要件 1 に関し，ユーザが登録した正解文書数が少ない場合の検索精度向上手法として，トランスダクティブ学習を利用したクエリ拡張手法[8]がある．この手法は，適合文書が 1 つのみ与えられた場合に対し，みなしの非適合文書を利用する疑似フィードバックにより，関連文書を効率的に収集することを目的としている．目的の方向性は本提案と同様であるが，本提案では複数の適合文書が登録されることを許容しているため，疑似フィードバックは利用しないこととする．

機能要件 2 に関し，既存手法として，対話的文書検索に分類される技術がある．システムが検索結果を自動分類し，ユーザがその分類を選択することを繰り返すことで，ユーザが必要とする情報を徐々に絞り込む手法がある．例えば，文書集合から主要なキーワードを抽出し，関連するキーワード同士が近くに配置されるキーワードマップを利用する分類手法[3]や，出現単語が類似する文書をクラスタ化する分類手法[4][5]がある．また，ペンインタフェースと検索目的指定による検索操作支援手法[6]は，ペン型の入力インタフェースを利用してユーザが検索語を選択すると，システムがその周辺語に応じた検索意図を推定し，メニュー(地図が見たい，ニュースリリースが見たいなど) をユーザに提示して絞り込む．しかしながら，ユーザがテーマを任意に指定できないことや，あらかじめ決められたテーマに限られてしまう問題がある．評価対象の情報源(ここではテレビの字幕)で述べられている内容が，Web 上で公開されている情報のうち，どのくらいを網羅しているかを測定する手法がある[7]．TF-IDF 値の高い名詞とその共起語の組を作り，それを検索語として Web 検索した結果上位 100 件との差分を抽出している．しかしながら，抽出した差分が評価対象の内容にどの程度関連しているかを評価できないため，テーマを網羅的に収集する方式には応用できない．

## 3. テーマ関連知識習得支援方式の提案

2 つの機能要件を満たすテーマ関連知識習得支援方式を提案する．機能要件 1 については，単語間の意味の近さを利用した単語スコアリング手法に基づく単語空間構築を行う方針とする．具体的には，単語の出現頻度に加えてシソーラス辞書(言語工学研究所シソーラス辞書[11]，37 万語収録)を利用し，テーマに対する単語の関連度の高さを推定することで，テーマに関連する単語を収集する．機能要件 2 については，対話的文書検索を利用したテーマ関連単語の半自動拡張を行い，未習得の情報を優先的に提示することでユーザの知識習得を支援する方針とする．具体的には，ユーザがテーマに関連する情報を繰り返しシステムに入力することで単語空間を自動的に拡張し，関連情報を収集する．さらに，短時間での知識習得を可能とするため，習得済みの Web ページと重複する情報を含む Web ページを取り除いてユーザに提示する．

以下，3.1 節では，知識習得支援を行うシステムについて，機能構成と処理シーケンスについて述べる．3.2 節では，未習得情報を判定するためのテーマに関連する単語集合の構築方法について述べる．

### 3.1 知識習得支援システムの動作フロー

本研究で想定する知識習得支援システムの動作フローを図 3 に示す．ユーザは知識習得支援システムを以下の手順で利用することを想定する．

① ユーザがテーマ名を入力して登録する．
② ユーザが知識習得支援システムに学習データを登録する．ユーザは適当な検索語(例えばテーマ名)を入力して検索を行い，検索結果を閲覧してテーマに合致すると判断すれば「習得済み」のフラグを付与する．これを一定数以上繰り返す．
③ システムがテーマ名および習得済み情報を利用して，関連情報が得られるような検索クエリを生成する．
④ システムが③で生成した検索クエリを利用して Web 検索を行う．
⑤ システムが検索結果を受信する．
⑥ システムが検索結果を精査し，ユーザに提示する優先度を判定して未習得情報を含む Web ページのリストを生成する．
⑦ ユーザが上記リストを受信する．
⑧ ユーザが結果を閲覧し，内容がテーマに合致する Web ページを選択し，Web ページを習得済み情報としてシステムに登録する．以降，②に戻り繰り返す．

Web ページにはテーマが複数存在することがある[9]．そのため本システムでは，ユーザがシステムに習得情報を教示する方法として，Web ページ全体を教示する手段だけではなく，その一部分の文章のみを教示する手段を提供した．なお，複数のテーマを自動抽出する技術は，既存技術[9][10]等の手法があり，本件に応用するとユーザの手間を削減できる可能性はあるが，本稿では扱わない．

図 4 に，知識習得支援システムの機能構成を示す．また，以下に各機能の概要を述べる．

テーマ情報 DB：テーマに関する情報を蓄積する．具体的には，テーマ名，テーマ全体の単語空間，習得済みの単語空間を蓄積する．単語空間とは，テーマに関連する単語とそのテーマに対する関連度の高さを表すスコア(以下，単語スコアと呼ぶ，3.3 節で説明する)の組を，複数の単語について蓄積したものとする．

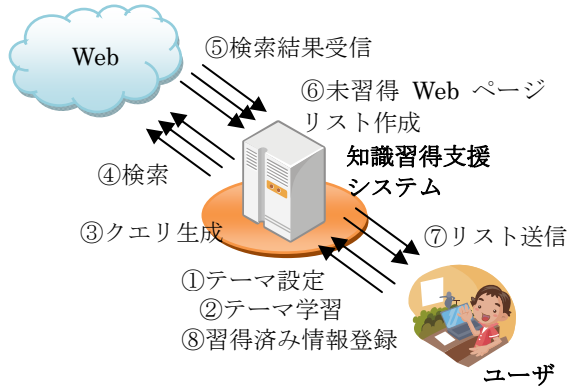


図 3 想定する知識習得支援システムの動作概要

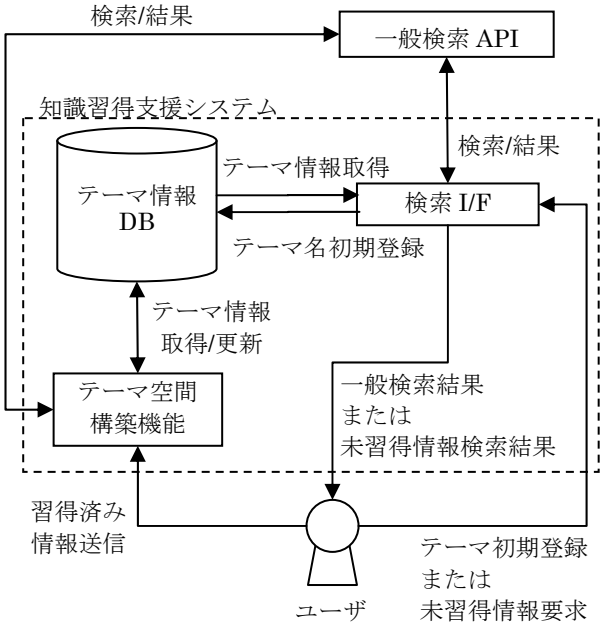


図 4 知識習得支援システムの機能構成

検索 I/F：初期状態では，テーマ名の登録，テーマ情報 DB 内の検索結果およびテーマ名に基づく一般検索 API を利用した検索結果を表示するインタフェースである．テーマ登録済みの状態では，テーマ情報 DB から未習得情報を抽出し，ユーザに提供する．

テーマ空間構築機能：ユーザから習得済み情報を受信し，一般検索 API を利用しながらテーマ情報 DB を更新する．習得済み情報は Web ページ単位で受信し，URL と HTML テキスト，または，ユーザが選択した HTML 中の一部のテキストからなる．

一般検索 API：一般に公開されている Web 検索の API (Application Program Interface)を指す．具体的には，Google AJAX Search API[12]や，Yahoo! 検索 Web API[13]等がある．

### 3.2 単語空間構築手法

テーマ空間構築機能における，単語空間の構築方式について詳細を述べる．ユーザがテーマを登録し，一定数以上の習得済み情報を登録すると，テーマ空間構築機能が，習得済み単語空間を構築する．習得済み単語空間とは，習得済み Web ページに含まれる単語と単語スコア(後述)のリストである．次に，テーマ空間構築機能が習得済み単語空間の単語を利用して Web ページを検索し，習得済み単語空間に対する類似度が一定値以上の Web ページを収集して，習得済み単語空間を含むテーマ全体の単語空間を構築する．以降，習得済み情報を受信する毎に，2 つの単語空間を自動拡張する．2 つの単語空間と評価対象文書の関係の概念図を図 5 に示す．A～E は 3 つの円で分割された領域を指す．

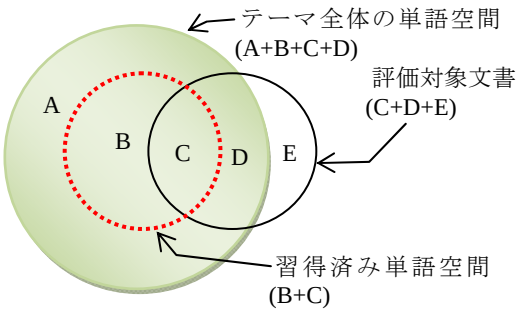


図 5 単語空間概念図

以下に，ステップに分けて処理の詳細を述べる．

**Step1.** テーマキーワードの抽出

テーマ空間構築機能が，習得済み情報として受信した Web ページの HTML を取得してテキストの抽出を行い，形態素解析により単語を取り出す．次に，各単語の IDF 値(Inverse Document Frequency)を取得する．ここで IDF 値は，1994 年から 2006 年までの毎日新聞オンラインニュース 13 年分の 135 万記事を利用してあらかじめ作成したものを利用した．

**Step2.** 単語スコアの算出

習得済み Web ページの単語一覧を取り出して単語ごとに出現頻度(以下，TF 値)を算出する．さらにシソーラス辞書を利用した 2 単語間のリンク数を計測する．ここでリンクとは，シソーラス辞書において上位概念または下位概念として登録されている単語同士の場合に 1 と数える指標とする．すべての単語について他の単語とのリンクの有無を判定し，その合計値をリンク数とする．総リンク数はすべての単語のリンク数の合計値とする．**Step1.** で算出した IDF 値，TF 値，リンク数，総リンク数を利用した，単語 $w$ のスコア $Score(w)$の算出式を式(1)に示す．算出結果はテーマ情報 DB に保存する．

$$Score(w) = \frac{\text{TF*IDF*リンク数}}{\text{TF 合計*総リンク数}} \qquad (1)$$

**Step3.** 自動拡張用検索キーワードの抽出

テーマ空間を拡張するための自動拡張用検索キーワードを決定する．ここで，自動拡張用検索キーワードは，テーマ名に加えて，固有名詞，動詞，サ変名詞を単語スコアが高い順にそれぞれ一定数ずつ選択する．固有名詞は，一般的に内容を特定して絞りやすくする語であり，本件においても有効であると考えられる．動詞およびサ変名詞については，例えば「ジュースの作り方」や「ジ

ャムの作り方」のように，材料名が同じでも調理方法(絞る，潰す，煮る等)が異なることで，テーマが変わる場合を識別するために導入する．

**Step4.** テーマ関連文書候補の抽出

選択した拡張用検索キーワードについて，Yahoo!API を利用して AND 検索を行い，検索結果の上位最大 100 件程度をテーマ関連文書候補とする．

**Step5.** テーマ関連文書候補のクリーニング

既にテーマ関連文書としてテーマ情報 DB に登録されている文書は外す．ここで，(外された文書数)÷(テーマ関連文書候補数) が一定値を超える場合は，テーマを拡張する効果が無いと判断し，自動拡張処理を終了する．それ以外の場合，テーマ関連文書候補のテキストに対し，形態素解析を行って候補単語を取り出し，新聞コーパスを使用して，各単語の IDF 値を取得する．式(2)の条件を満たした場合にその文書を「テーマ全体」を構成する自動検索したテーマ関連文書として判定し，テーマ情報 DB に登録する．同時に，テーマ全体の単語空間の更新も行うが，自動拡張により新たに追加された単語については，ユーザの確認がないことから，単語スコアを $r$ 倍 (0<r<1)する．

$$\frac{\text{習得済み単語空間中の候補単語のスコア合計値}}{\text{対象文書の全単語数}} > s \qquad (2)$$

**Step6.** 未習得情報要求結果の並び変え

ユーザが未習得情報要求を行うと，検索 I/F がテーマ情報 DB から **Step5.**で登録したテーマ関連文書を提示する．このとき，テーマ関連文書を優先度順に並べる必要があるため，あらかじめ，並び変えるための指標を算出する．ここでは，関連度および未習得度を指標とし，テーマ全体と習得済みの単語空間を利用して算出する．式(3)および式(4)に，図 5 に基づく関連度および未習得度の算出式を示す．本稿では，未習得度，関連度の順でソートした．なお，$S$(“領域名”)は，図 5 の各領域に含まれる単語の単語スコア合計値を表す．

(a) 関連度=$S$(C)/$S$(C+D+E) (3)
評価対象文書に含まれる単語のうち，習得済み範囲に含まれる単語の割合として算出する．ここでは，ユーザの意思が明確に示されている C の領域のみを関連度算出の対象とする．

(b) 未習得度=$S$(D)/$S$(C+D) (4)
評価対象文書に含まれ，かつテーマに関連する単語のうち，習得済み範囲に含まれない単語の重み付き割合として算出する．

**Step7.** 網羅度の算出

テーマ全体に対する習得した情報の割合の指標として，網羅度を算出する．テーマ全体の単語空間に対する，習得済みの単語空間の割合で表す．図 5 に基づき，算出式を式(5)に示す．この値は，図 2 で示した画面例において，右上に 100%を上限値としたグラフで表示する．

網羅度=$S$(B+C)/$S$(A+B+C+D) (5)

## 4. 評価実験

機能要件 1 および機能要件 2 に対し，提案方式の有効性を確認するための評価実験を行う．機能要件 1 である「習得済み情報の登録数が少ない場合であっても高精度な情報収集ができること」について，一定の検索結果件数内で，提案方式が一般の検索エンジンでは得られない情報をどれだけ多く得られているか，という指標で評価実験(以降，実験 1 と呼ぶ)を行った．同時に，一般の検索エンジンが提案方式の結果にはない情報をどれだけ得られたかについても評価を行い，比較する．また，機能要件 2 である「テーマの関連情報を，短時間で重複なく収集すること」について，提案方式が一般の検索エンジンよりも未習得情報をどれだけ検索できたか，という指標で評価実験(以降，実験 2 と呼ぶ)を行った．実験 2 についても実験 1 と同様に，同時にその逆についても評価して比較した．なお，本評価実験では，提案方式のパラメタの最適値を導出するための予備実験を最初に行い，実験 1 および実験 2 では，その結果を利用した．

以下，4.1 節で具体的な評価方法について述べ，4.2 節で実験結果を示す．4.3 節で考察を述べる．

### 4.1 実験方法

#### 4.1.1 共通実験条件

まず，3 つの評価実験に共通の実験条件を述べる．評価実験用のテーマとして，「テーマ 1: おいしいカレーの作り方」，「テーマ 2: 体にやさしいお酒の飲み方」，「テーマ 3: 情報大航海」，「テーマ 4: ワークシェアリングとは」の 4 つを選択した．これらは，1 つのサイトを情報源とするだけでは，十分な情報が得られない難易度が高い 2 テーマ(テーマ 1, 2)と，一般の検索エンジンでも容易に検索可能な多義性のない単語の意味を調べることが目的の 2 テーマ(テーマ 3, 4)の基準で選択した．各テーマ名を提案方式の評価用システムに入力してテーマ登録を行った．

次に，一般の検索エンジンを利用してテーマ名で検索を行った．被験者が検索結果 1 位から順に閲覧し，テーマに則した内容であれば，評価用システムに「習得済み」として登録した．習得済みが 10 件になるまで学習を繰り返し，10 件を登録し終えた時点でテーマ空間構築機能を起動して，テーマ全体の単語空間と習得済みの単語空間を構築した．最後に，未習得情報を含む Web ページを検索し，学習中に閲覧した Web ページ数との和が 50 ページになる数だけ，検索結果上位から抽出した．比較対象は，一般の検索エンジンである Yahoo!検索 Web API とし，テーマ名を検索キーワードとして検索した結果の上位 50 件を抽出した．

被験者(1 名)に，提案方式と比較対象のそれぞれについて，50 件すべての検索結果の Web ページを閲覧してもらい，内容の重複に関する判断をしてもらった．この判断は情報を含むかどうかという客観的な判断でほぼ行えることから，被験者は 1 名でも大きな問題とはならないと判断した．上位 50 件までの評価とした妥当性については，検索エンジンに関する調査[14]によると，検索結果は 5 ページ目までしか見ないユーザが 93.6%と報告していることから，本実験ではそれに合わせている．検索結果の画面に 1 ページあたり 10 件の URL が表示されているとすると，5 ページで 50 件程度となる．また，繰り返しの学習効果については学習 1 回あたりの未習得情報の検索性能の積み重ねで

あるため，本実験では一定数の習得済み情報を 1 度登録して得られた未習得情報の検索結果に対する評価を行った．なお，3.3 節で記載した変数 r, s の値は実験的に決定し，r=0.5, s=0.3 と設定した．

以下，テーマ空間構築機能の拡張用検索キーワードの設定値を決めるための予備実験について述べる．また，予備実験で得られた値を用いて実験 1 および実験 2 を実施する．

#### 4.1.2 予備実験

3.2 節の(エ)で述べた拡張用検索キーワードの品詞毎の採録数のパラメタについて，4 つのテーマに対して最適となる値を実験的に決定する．固有名詞，動詞，サ変名詞それぞれについて，0 個～5 個に昇順に変化させた場合の実験 1 および実験 2 の評価値を参照する．0 個の場合の評価結果よりも評価値が上で，かつ，上昇し続けるまでの固有名詞，動詞，サ変名詞それぞれの個数を計測する．0 個の場合の評価結果よりも下回る場合には「効果なし」と判定する．

#### 4.1.3 実験 1

提案方式が，一般の検索エンジンでは得られない情報をどれだけ集めることができたかを評価する．各テーマについて，提案方式において検索した未習得情報を含む Web ページの内容の全部または一部が，一般の検索エンジンの結果において得られた Web ページのいずれかに記載されているかどうかを判定し，記載されていない情報を含む Web ページ数をカウントした．比較対象として，その逆，つまり，一般の検索エンジンの検索結果の Web ページの内容の全部または一部が，提案方式の検索結果に記載されていない情報を含むかどうかを同様に判定し，その Web ページ数をカウントした．

#### 4.1.4 実験 2

提案方式が，習得済みの Web ページの内容に含まれない未習得情報をどれだけ検索できたかを評価する．各テーマについて，提案方式で習得済みとした 10 件の内容にない情報を含む Web ページが，提案方式で検索した未習得情報を含む Web ページの検索結果に何件あるかを判定してカウントした．一般の検索エンジンの結果を比較対象とし，同様に，提案方式で習得済みとした 10 件の内容にない情報を含む Web ページ数を判定してカウントした．

### 4.2 実験結果

#### 4.2.1 予備実験の結果

結果を表 1 に示す．表より，効果なしの項目を除くと，最低値は固有名詞を 3 個まで，動詞を 1 個まで，サ変名詞を 2 個までとなり，テーマ名も検索語となることから，合計で 7 個となる．ただし，検索語の数が多くなると検索結果の数が必要以上に減少する可能性がある．よって調整を行い，予備実験の結論としては，固有名詞とサ変名詞から 1 個ずつ減らし，固有名詞 2 個，動詞 1 個，サ変名詞 1 個とし，これを設定 A とした．また，テーマ 3, 4 において，動詞とサ変名詞の効果がない場合があったため，それらを固有名詞に割り当て，これを設定 B とした．これらにテーマ名を加えた 5 つの検索語を設定し，実験 1 および実験 2 を行うこととした．設定 A と設定 B を表 2 にまとめた．

表 1 最適な各品詞の個数

| | テーマ名 | 固有名詞 | 動詞 | サ変名詞 |
|---|---|---|---|---|
| 1 | おいしいカレーの作り方 | 5 個まで | 2 個まで | 2 個まで |
| 2 | 体にやさしいお酒の飲み方 | 4 個まで | 1 個まで | 3 個まで |
| 3 | 情報大航海 | 3 個まで | 効果なし | 2 個まで |
| 4 | ワークシェアリングとは | 5 個まで | 1 個まで | 効果なし |
| 最小値 | | 3 個 | 1 個 | 2 個 |

表 2 拡張用検索キーワードに利用する単語の個数

| | テーマ名 | 固有名詞 | 動詞 | サ変名詞 |
|---|---|---|---|---|
| 設定 A | 1 | 2 | 1 | 1 |
| 設定 B | 1 | 4 | 0 | 0 |

#### 4.2.2 実験 1 の結果

設定 A の場合の実験結果を表 3 に示す．カッコ内の数値は，「(比較対象にはない情報を含む Web ページの数/10 件の習得以降の Web ページ数)」を表し，その比の値を左に記した．また，一般の検索の結果の値に対する提案方式の結果の値の割合を「倍率」として記載した．表より，特にテーマ 1, 2,において倍率が約 3 倍という高い値が得られた．提案方式の方が一般の検索結果にはない情報をより広い範囲で検索できていたといえる．一方，テーマ 4 は 0.78 であり，提案方式が下回った．

次に設定 B の場合の実験結果を表 4 に示す．テーマ 4 の提案方式の倍率が 0.78 から 1.04 に改善されて提案方式の方が上回る結果となった．逆に，テーマ 1 については 3.06 から 1.26 に大きく減少し，固有名詞よりも動作に関する単語が重要なテーマであることが分かる．

表 3 他方にない情報を含む Web ページの割合(設定 A)

| | テーマ名 | 提案方式 | 一般の検索 | 倍率 |
|---|---|---|---|---|
| 1 | おいしいカレーの作り方 | 0.49(18/37) | 0.16(6/37) | 3.06 |
| 2 | 体にやさしいお酒の飲み方 | 0.5(15/30) | 0.17(5/30) | 2.94 |
| 3 | 情報大航海 | 0.25(7/28) | 0.14(4/28) | 1.79 |
| 4 | ワークシェアリングとは | 0.53(21/40) | 0.68(27/40) | 0.78▽ |

※カッコ内の数値は，「(比較対象にはない情報を含む Web ページの数/10 件の習得以降の Web ページ数)」を表す
※▽は提案方式が一般の検索よりも低いデータを表す

表 4 動詞とサ変名詞を取り除いた場合の
他方にない情報を含む Web ページの割合(設定 B)

| | テーマ名 | 提案方式 | 一般の検索 | 倍率 |
|---|---|---|---|---|
| 1 | おいしいカレーの作り方 | 0.39(14/36) | 0.31(11/36) | 1.26 |
| 2 | 体にやさしいお酒の飲み方 | 0.86(25/29) | 0.28(8/29) | 3.07 |
| 3 | 情報大航海 | 0.27(8/30) | 0.1(3/30) | 2.70 |
| 4 | ワークシェアリングとは | 0.55(21/38) | 0.53(20/38) | 1.04 |

※カッコ内の数値は，「(比較対象にはない情報を含む Web ページの数/10 件の習得以降の Web ページ数)」を表す

#### 4.2.3 実験 2 の結果

設定 A の場合の実験結果を表 5 に示す．カッコ内の数値は，(上位 10 件にはない情報を含む Web ページの数/10 件の習得以降の Web ページ数)を表し，その比の値を左に記

載した．実験 1 と同様に，倍率を算出して比較した．表 5 より，テーマ 1, 2, 4 において倍率が 1.00 を超えており，提案方式の方が一般の検索結果よりもより多くの未習得情報を検索できていた．一方，テーマ 3 においては 1.00 となり，一般の検索と結果は変わらなかった．

設定 B の場合の実験結果を表 6 に示す．表より，テーマ 3 について，提案方式の倍率が 1.00 から 2.35 に大幅に改善され，テーマ 3 は動作に関する単語は無関係であるといえる．また，テーマ 1 については提案方式の倍率が 1.46 から 0.64 に下がっており，実験 1 と同様に，テーマ 1 は動作に関する単語が重要であることが分かる．

表 5 習得した 10 件にはない情報を含む Web ページの割合 (設定 A)

| | テーマ名 | 提案方式 | 一般の検索 | 倍率 |
|---|---|---|---|---|
| 1 | おいしいカレーの作り方 | 0.35(13/37) | 0.24(9/37) | 1.46 |
| 2 | 体にやさしいお酒の飲み方 | 0.23(7/30) | 0.17(5/30) | 1.35 |
| 3 | 情報大航海 | 0.25(7/28) | 0.25(7/28) | 1.00 |
| 4 | ワークシェアリングとは | 0.75(30/40) | 0.68(27/40) | 1.10 |

※カッコ内の数値は，「(比較対象にはない情報を含む Web ページの数/10 件の習得以降の Web ページ数)」を表す

表 6 動詞とサ変名詞を取り除いた場合の，習得した 10 件にはない情報を含む Web ページの割合(設定 B)

| | テーマ名 | 提案方式 | 一般の検索 | 倍率 |
|---|---|---|---|---|
| 1 | おいしいカレーの作り方 | 0.14(5/36) | 0.22(8/36) | 0.64▽ |
| 2 | 体にやさしいお酒の飲み方 | 0.21(6/29) | 0.17(5/29) | 1.24 |
| 3 | 情報大航海 | 0.4(12/30) | 0.17(5/30) | 2.35 |
| 4 | ワークシェアリングとは | 0.79(30/38) | 0.66(25/38) | 1.20 |

※カッコ内の数値は，「(比較対象にはない情報を含む Web ページの数/10 件の習得以降の Web ページ数)」を表す
※▽は提案方式が一般の検索よりも低いデータを表す

### 4.3 考察

テーマ 1 およびテーマ 2 については，実験 1 および実験 2 の結果のいずれも提案方式は一般の検索よりも高い値が得られており，機能要件 1 および機能要件 2 に対する提案方式の有効性を確認できた．提案方式は，未習得情報をより多く収集し，一定の検索結果数でテーマに関連する情報を密度濃く収集できているといえる．また，実験 1 の表 3 において最大で一般の検索と比較して 3.07 倍の効果が出ており，特に機能要件 1 に対する効果が高いことが確認できた．

拡張用検索キーワードの選択方法については，すべてのテーマについて一律の単語選択方法を適用すると，適切とならないテーマがあることが分かった．テーマ 1 と 2 については，提案方式が適するテーマであり，材料とそれに深く関連する調理方法や飲み方などの動作に関する単語が重要となっていた．一方，テーマ 3 と 4 については，動作に関する単語がほとんど関連しない単純なテーマであり，動詞やサ変名詞を検索キーワードとすることで，必要以上に情報を絞り込んでしまったと考えられる．よって，ユーザが入力したテーマ名からいずれのタイプかを推定し，拡張用検索キーワードの選択方法を適応的に変えることで，提案方式の汎用性を上げることができると考えられる．

## 5. おわりに

本稿では，Web 検索を利用して，ユーザが指定したテーマについて Web から知識を短時間で網羅的に習得する場合において，半自動的に関連情報を収集することを可能にする，知識習得支援システムの実現方法について検討し，テーマ関連知識習得支援方式を提案した．提案方式は，システムがテーマに関する単語空間を対話的文書検索により半自動的に拡張し，少ない習得文書数であっても未習得の情報を優先的に提示できること，および，情報の重複が減るように提示できることを特徴とする．評価実験において 4 つのテーマを選択して一般の検索エンジンの検索結果と比較した結果，特に，動作に関する単語により情報の細分化が行われるような比較的領域が広いテーマの場合に，提案方式は一般の検索の約 3 倍の広範囲な情報が収集できることを示した．また，テーマの単語空間を適切に拡張するための検索キーワードの生成方法において，テーマの種別に応じた適応的な単語選択を行うことで，精度向上の見通しを得た．

今後は，実用化に向けたフィールド実験を行い，様々なテーマを対象とした提案方式の有効性を検証する．

### 参考文献

[1] Wikipedia, http://ja.wikipedia.org/.
[2] はてなブックマーク, http://b.hatena.ne.jp/.
[3] 高間, 廣田, “免疫ネットワーク・メタファに基づく Web 情報可視化手法”, 日本ファジィ学会誌, Vol. 14, No. 5, pp. 472-481 (Oct. 2002)
[4] Douglass R. Cutting, David R. Karger, Jan O. Pedersen and John W. Turkey, “Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections”, Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 318-329 (June 1992).
[5] 小林, 佐藤, 三末, 田中, “Web 検索結果の概観提示による情報収集支援インタフェース”, 第 19 回人工知能学会全国大会, No. 3C3-03 (June 2005).
[6] 石谷, 鈴木, 布目, “連鎖検索と近傍検索に基づく Web コンテンツへの効率的なアクセス方法”, 第 6 回 Web インテリジェンスとインタラクション研究会(SIG-WI2), pp. 31-36 (July 2006).
[7] 甲谷, 湯本, 小山, 田島, 田中, “TV ニュース映像の話題の網羅性・一般性・受容度の可視化による視聴支援”, 日本データベース学会 Letters, Vol. 6, No. 1, pp. 61-64 (June 2007).
[8] 岡部, 山田, “トランスダクティブ学習による最小文書判定からのクエリ拡張”, 人工知能学会論文誌, Vol. 21, No. 4, pp. 398-405 (July 2006).
[9] 上田, 斉藤, “多重トピックテキストの確率モデルーパラメトリック混合モデルー”, 電子情報通信学会論文誌, Vol. J87-D-II, No.3, pp. 872-883 (Mar. 2004).
[10] 高木, 藤井, 石川, “検索質問の主題分析に基づく類似文書検索と特許検索への応用”, 情報処理学会論文誌, Vol. 46, No. 4, pp. 1074-1081 (Apr. 2005).
[11] 言語工学研究所, http://www.gengokk.co.jp/.
[12] Google AJAX Search API, http://code.google.com/intl/ja/apis/ajaxsearch/.
[13] Yahoo!検索 Web API, http://developer.yahoo.co.jp/webapi/search/.
[14] Web マーケティングガイド, 第 4 回検索エンジンに関する調査, http://www.e-research.biz/profile/prosem/000164.html.