社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－146 (43)
2008／9／23

Wikipediaのリンク構造とカテゴリ構造を用いた検索語からの専門語の抽出

中谷 誠† AdamJatowt† 大島 裕明† 田中 克己†

† 京都大学情報学研究科社会情報学専攻 〒606–8501 京都市左京区吉田本町
E-mail: †{nakatani,adam,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 検索語が属する分野における専門語は，ユーザが検索結果に含まれるウェブページの内容を理解する上で重要な役割を持っている．専門語が多く含まれるウェブページは，非専門家ユーザにとっては理解しづらいが，一方で専門家ユーザにとっては読みやすく詳細な情報を得る上で有用である．本研究では，Wikipediaのリンク構造とカテゴリ構造を用いて，ユーザの入力した検索語からその語に関する専門語を抽出する手法について述べる．Wikipedia中で検索語が含まれている記事のカテゴリ情報を集約することによって検索語の属する専門領域を検出し，その領域の内外でリンクの出現頻度を分析することによって専門語を抽出する．本研究の提案手法は幅広い分野を網羅しており多言語対応しているWikipediaを用いているので，検索語の分野や言語に関係なく専門語を取得することができる．
キーワード Wikipedia，専門語抽出

Extraction of Technical Terms for Query Keywords by Link and Category Structure of Wikipedia

Makoto NAKATANI†, Adam JATOWT†, Hiroaki OHSHIMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University Yoshida–honmachi, Sakyo, Kyoto, 606–8501 Japan
E-mail: †{nakatani,adam,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract Technical terms for query keywords are important for users to grasp the meaning of the contents of Web pages included in search results. In particular, Web pages with many technical terms are difficult to be understood by non-expert users, yet, they are rather readable and useful for expert users who can acquire detailed information from them. This paper proposes a method for extracting technical terms for query keywords by using link and category structure of Wikipedia. We detect technical domains of query keywords by aggregating category information of Wikipedia articles and we extract technical terms by analyzing link frequency in inside and outside of the domains. Since our proposed method uses Wikipedia, thus it is domain-independent covering many different topics and it does not depend on particular language of query keywords.
Key words Wikipedia, Technical Term Extraction

1. はじめに

近年ウェブから情報を取得するために検索エンジンが広く用いられるようになってきている．ウェブページの爆発的な増加と検索技術の向上により，検索エンジンは検索語に適合した多くのウェブページをユーザに返してくれるようになってきた．しかし，既存のウェブ検索エンジンは同じ検索語が入力されたとき，どのユーザに対しても同じ検索結果を返すため，検索を行う際のユーザの動機を反映したものとはなっていない．中村らのアンケート調査[1]によると，ユーザが検索を行う動機のうち次の2点が大きな割合を占めている．

- 検索語について知らないため (46%)
- 検索語についてより深く知りたいため (36.8%)

検索語について知らないために検索を行うユーザは，できるだけ分かりやすく検索語についての説明や定義が記述されたウェブページを求めるだろう．また，検索語についてより深く知りたいために検索を行うユーザは検索語に関する知識をある程度持っていると考えられ，検索語についての詳細な情報が記述されているウェブページを求めるだろう．すなわち，この調査結果は同じ検索語が入力されたとしても，ユーザによってその

キーワードについての理解度が異なっており，ユーザの求める情報も異なっている可能性があることを示している．本研究では，検索結果中に含まれる専門語に着目する．専門語とは一般的には特定の学問領域や業界においてのみ使用される用語のことである．検索語についてよく知らないユーザはにとって，専門語が多く含まれるウェブページを理解することは困難であると考えられる．一方，検索語に関する知識をある程度持っているユーザにとっては専門語が使用されているウェブページのほうが読みやすく，より詳細な情報を得る上で有用である．これは一般的に専門語の持つ情報量が多いためである．このように，検索語についての理解度をもとにユーザにとって理解しやすく有用なウェブページを発見する上で，専門語は重要な要因の一つであるといえる．

本研究ではオンライン百科事典である Wikipedia から検索語に関する専門語を抽出する手法を提案する．既存の専門語抽出手法においては，あらかじめ用意した専門分野コーパスを対象に行う手法が多く取られてきたが，専門分野ごとにコーパスを用意しなければならないことや新しい用語に対応できないことが問題点であった．一方，Wikipedia は幅広い分野の用語や新語をカバーしているため，Wikipedia をコーパスとすることで既存手法の問題点を解消できるものと考えられる．また，本研究では日本語のみを対象としているが，Wikipedia を用いた提案手法は対象とする言語に依存しないため他言語への拡張が容易であるという利点を持つ．

本研究の提案手法の概要を図 1 に示す．本研究では，入力として検索語 q と専門語の候補 t が与えられ，t が q についての専門語であるかを判別するシステムを提案する．まず，検索語の Wikipedia 記事およびその記事を参照している他の Wikipedia 記事のカテゴリ情報を集約することによって，検索語の含まれている領域を検出する．次に，検出された領域を元に Wikipedia から専門分野コーパスおよび一般コーパスを生成し，それぞれのコーパスにおける候補語の Wikipedia 記事へのリンク数を分析することによって専門語であるかの判別を行う．

以下，本論文では，第 2 章で Wikipedia の概要について示し，第 3 章で関連研究について述べる．第 4 章で検索語の属する専門領域を検出する手法と予備実験について，第 5 章で専門語の抽出手法について述べる．第 6 章では専門語抽出手法についての評価実験について，第 7 章でまとめと今後の課題について述べる．

2. Wikipedia の概要

Wikipedia は誰でも記事を編集することが可能であるフリーのオンライン百科事典である．Wikipedia の発表した統計(注1)によると，2008 年 8 月において約 50 万の記事が投稿されており，2000 万回以上のユーザによる編集が行われている．Nature 誌によると自然科学のトピックにおいて Wikipedia は Britannica 百科事典と同程度の精度を示しているという [2].

Wikipedia は既存の辞書データとは異なるいくつかの特徴を持っている．図 2 に Wikipedia の記事構造の概要を示す．Wikipedia では一つの用語についての説明が一つの記事となっており，記事中に出現する用語の中で説明を要すると思われるものについて，その用語の記事へのリンクが張られている．また，Wikipedia には記事間のリンク構造だけでなく概念木により構成されるカテゴリ構造があり，各記事は一つ以上のカテゴリに属している．

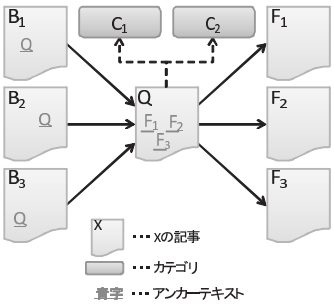

図 2　Wikipedia の記事構造

Wikipedia は大量の記事をダウンロードするためのクローリングが禁止されており，その代替として再配布・再利用のために全てのデータベース・データの提供が無償で行われている．本研究では，2008 年 6 月 24 日にダンプされた Wikipedia 日本語版のデータベースをダウンロードして利用している(注2)．

3. 関 連 研 究

3.1 Wikipedia マイニング

Wikipedia は記事同士がリンクでつながっているなど既存の辞書データとは異なる特徴を持っており，近年データマイニングや自然言語処理の対象として注目を集めている．Milne ら [3] は，農学分野のシソーラスである Agrovoc と比較して，Wikipedia には専門用語や上位・下位といった用語間の意味的関係が十分に含まれていることを示した．Wikipedia の構造を利用して語と語の関連性の強さを抽出しようといういくつかの試みがある [4] [5]．Strube ら [4] は，Wikipedia のカテゴリ木上での距離によって語と語の関連度を求める手法を提案している．中山ら [5] は Wikipedia の記事間のリンク構造を利用して大規模な連想シソーラスを構築する手法を提案している．また，機械学習によって Wikipedia から「WindowsXP は Microsoft の Product である」といったような語と語の関係そのものを抽出しようという Nguyen ら [6] の試みもある．

また，Wikipedia から得られた知識を利用したいくつかのアプリケーションが提案されている．Milne ら [7] によって提案された新しい検索エンジン *Koru* は，Wikipedia から得られた関連語によってクエリ拡張を行うものである．Mihalcea ら [8] によって提案された *Wikify*! は，Wikipedia から得られたキーワードを用いて通常のウェブページに含まれるキーワードに

(注1)：http://ja.wikipedia.org/wiki/特別:Statistics

(注2)：http://download.wikimedia.org/jawiki/

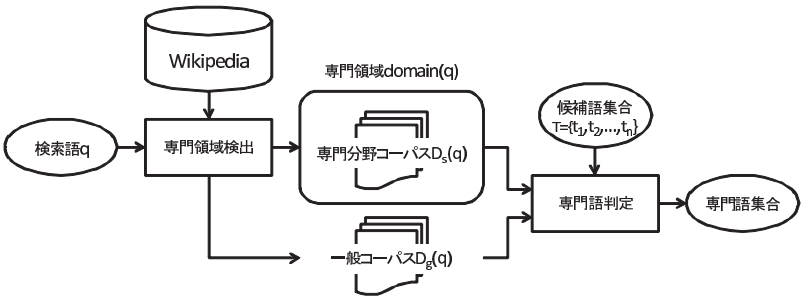

図 1　提案手法の概要

Wikipedia 記事へのリンクを付与するシステムである．

3.2　専門語抽出

専門語を抽出するためには専門家による人手が必要であるため，新しく大規模な専門用語辞書を構築することは非常にコストが大きい．そのような背景をもとに，主に自然言語処理の分野で自動的に用語抽出を行う手法が提案されてきた．

合原ら [9] は医療生物分野のコーパスから機械学習を用いて用語を抽出する手法を提案している．中川ら [10] はコーパスから連接名詞を抽出し HITS アルゴリズムのアナロジーを用いて連接名詞のスコアリングを行う FLR 法を提案している．これらのアプローチは専門分野コーパス内での単語間の表層的な連接関係をもとに抽出を行っているため，抽出された用語がその分野でのみ使用される専門語であるかは判別できない．

用語の分野判定を行う手法としては Chung [11] や木田ら [12] の研究がある．Chung は，人手で作成した専門分野コーパスと一般コーパスの間での用語の出現頻度の比を用いて用語の分野判定を行う手法を提案している．木田らは，既知の専門用語をもとにウェブ検索エンジンを利用して専門分野コーパスを作成し，それを対象として専門用語の判別を行う手法を提案している．本研究における提案手法は，Wikipedia のデータを用いることによって，あらかじめ専門分野コーパスや専門用語セットを用意する必要がないという点でこれらのアプローチとは異なる．

4.　検索語の属する専門領域の検出

専門語とは特定の領域においてのみ用いられる用語のことであるため，専門語を抽出するためにはまず検索語の属する専門領域を検出する必要がある．検索語へのリンクを持っている Wikipedia 記事のカテゴリ情報を集約することによって，検索語の属する専門領域を検出する．

4.1　提 案 手 法

検索語を q，Wikipedia の全記事集合を $E = \{e_1, e_2, ..., e_M\}$，全カテゴリ集合を $C = \{c_1, c_2, ..., c_N\}$ とする．まず，全記事集合 E の部分集合である $E_q = \{e_q\} \cup \{e_i | e_i \rightarrow e_q\}$ を取得する．集合 E_q は検索語 q の Wikipedia 記事 e_q および記事 e_q へのリンクを持っている他の Wikipedia 記事を含んでいる．

次に，集合 E_q 中の各要素 e_i について次式で表わされるベクトル $CL(e_i)$ を求める．

$$CL(e_i) = (\delta_{i,1}, \delta_{i,2}, ..., \delta_{i,N}) \tag{1}$$

$$\delta_{i,j} = \begin{cases} 1/G(N_{c_j}) & (e_i \text{が } c_j \text{に属するとき}) \\ 0 & (e_i \text{が } c_j \text{に属さないとき}) \end{cases} \tag{2}$$

N_{c_j} はカテゴリ c_j に含まれる記事数であり，$G(N_{c_j})$ は $CL(e_i)$ の要素の中で含まれる記事数が多いものが大きな値を持ちやすくならないように正規化を行うための関数である．さらに，得られた各ベクトルを全て足し合わせたベクトル $\Gamma(q)$ を計算する．

$$\Gamma(q) = \frac{1}{|E_q|} \sum_{e_i \in E_q} CL(e_i) \tag{3}$$

$\Gamma(q) = (\gamma_1, \gamma_2, ..., \gamma_N)$ の要素 γ_j の値は，検索語 q がカテゴリ c_j に含まれる記事でどれだけ多くから参照されているかを表している．最後に，$\Gamma(q)$ の要素の中で値の大きい上位 K 個の要素に対応するカテゴリを，検索語 q の属する領域 $domain(q)$ として選択する．ここでは，K をウィンドウサイズと呼ぶことにする．検索語の属する領域として複数のカテゴリを割り当てるのは，Wikipedia のカテゴリ構造が細分化され過ぎている場合があるためである．

4.2　予 備 実 験

前節で述べた検索語の属する領域の検出手法について，正規化係数とウィンドウサイズを決定するための予備実験を行った．Wikipedia のカテゴリ「情報学」およびその下位カテゴリに属する記事の見出し語から，情報学に関する用語 50 語を選択し実験のための検索語セットとした．正規化係数 $G(N)$ として 1，$\log N$，$\sqrt{N}$，N の 4 種類を用いた．各検索語 q について正規化係数 $G(N)$ ごとに求めたベクトル $\Gamma(q)$ の要素を値の高い順にソートし，上位 10 件のカテゴリについてそれぞれ検索語 q の属する領域として適切であるかを評価した．予備実験において，提案手法によって検出しようとしている検索語の属する領域を表すものとして不適切であると考えられるいくつかのカテゴリは無視している．例として，「曖昧さ回避」や「編集保護中」といった記事の状態を表す意味合いの強いものや「20XX 年の〜」といった時間を表すものが挙げられる．

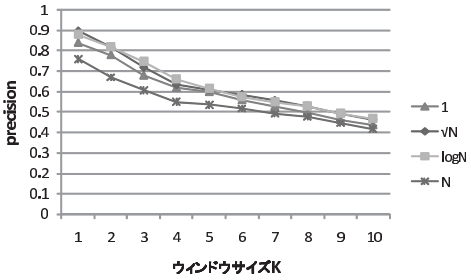

図 3　領域検出の精度

4.3　予備実験の結果および考察

評価実験の結果を図 3 に示す．横軸はウィンドウサイズの大きさを表し，縦軸は該当のウィンドウサイズを選択したときの検出精度の平均を表す．検索語セットに含まれる各語に対して検出精度は，

$$precision = \frac{domain(q) \text{ 中で適切と判断されるカテゴリの数}}{\text{ウィンドウサイズの大きさ}} \qquad (4)$$

により計算しており，全ての検索語に対して計算を行ったのち各ウィンドウサイズについて精度の平均を求めた．ウィンドウサイズを 1 としたとき，いずれの手法についてもその精度は 70〜90% となり，ウィンドウサイズを 3 としたときでも正規化係数として $G(N) = \log N$ または $G(N) = \sqrt{N}$ を選択した場合は 70% 程度の精度を保つことが分かった．正規化係数としていずれを選択しても，ウィンドウサイズを大きくするほど検索語の属する領域として不適切となる割合が高くなる．また，正規化を行わない場合 ($G(N) = 1$) に比べ，正規化係数として $G(N) = \log N$ または $G(N) = \sqrt{N}$ を選択すると全体として精度がやや上がることが分かったが，大きな改善といえるほどのものではなかった．これは，正規化を行わない場合に，単に含まれる記事数が多いカテゴリほど高い値を示してしまう可能性があり正規化を行うことでそれを防ぐことができたが，そのようなカテゴリは多くなく精度の改善にあまり影響を及ぼさなかったものと考えられる．一方，正規化係数として $G(N) = N$ を選択すると，正規化を行わない場合よりも精度が悪化してしまった．これは正規化係数の影響が強く出てしまい，含まれる記事数の少ないカテゴリが高い値を示しやすくなってしまったためであると考えられる．

予備実験により，一定以上の精度を保ちつつ複数のカテゴリを含むように検索語の属する領域を選択するためには，正規化係数として $G(N) = \log N$ または $G(N) = \sqrt{N}$ を選択し，ウィンドウサイズを 3 程度に設定するとよいことが分かった．$G(N)$ を $\log n$，ウィンドウサイズを 3 としたときの実行結果を表 1 に示す．

5.　専門語判定

ある語 t が，前節で述べた手法により得られた検索語 q の属する領域における専門語であるかを判定するための手法について述べる．専門語抽出に関する既存の研究にならい，本研究においても専門用語コーパスと一般コーパスを生成し，両コーパス間での候補語 t の出現頻度の比によって専門語であるかどうかの判定を行う．ここでは，領域 $domain(q)$ に含まれるカテゴリに属する全 Wikipedia 記事を専門分野コーパス $D_s(q)$ として扱う．ただし，領域内の複数のカテゴリに属する記事は重複して考えないものとする．このようにして得られた専門分野コーパス以外の Wikipedia 記事を一般コーパス $D_g(q)$ として扱う．一般コーパスであまり使用されず，専門分野コーパスで頻出するような語が専門語である．以下，語 t が領域 $domain(q)$ においてどの程度専門的であるかを求めるための 2 つスコアリング手法について述べる．

表 1　検索語の領域検出例 (ウィンドウサイズ：3)

q	$domain(q)$
SQL インジェクション	セキュリティ技術，SQL，問い合わせ言語
関係データベース	データベース，データモデリング，問い合わせ言語
形態素解析	自然言語処理，全文検索，日本語入力システム
ダイクストラ法	検索アルゴリズム，データ構造，グラフ理論
ファジィ集合論	人工知能，論理学，意味論
ポリモーフィズム	オブジェクト指向，プログラミング， プログラミングパラダイム
パーセプトロン	分類，機械学習，統計学

5.1　tf.idf　法

$tf.idf$ 法は情報検索において文書の特徴ベクトルを求めるための代表的なスコアリング手法である [13]．文書 d 中での語 t の出現頻度を表す $tf(t,d)$ と，コーパス中で語 t が出現する文書数の逆数を表す $df(t)$ を乗じることによって，文書 d に特徴的に現れる語ほど高いスコアを与えることができる．ここでは，$tf.idf$ 法における tf のかわりに専門分野コーパス $D_s(q)$ 内で語 t へのリンクを持っている記事の総数 $lf(t, D_s(q))$ を，idf のかわりに Wikipedia 全体の中で語 t へのリンクを持っている記事の総数 $lf(t, D_W)$ の逆数を用いた次の式によりスコアリングを行う．

$$tf.idf(t,q) = lf(t, D_s(q)) \cdot \log \frac{|D_W|}{lf(t, D_W)} \qquad (5)$$

ただし、$|D_W|$ は Wikipedia の総記事数を表す．

5.2　χ^2 独立性検定

χ^2 独立性検定は 2 つの変数に対する 2 つの変数が互いに独立であるかを検定する統計的手法である．ここでは，語 t の記事へのリンクを持った記事が，専門分野コーパス $D_s(q)$ 内にどの程度集中して現われているかを求める．χ^2 独立性検定を行うために必要な情報を表 2 に示す．専門度を表すスコアとしては χ^2 値を用いる．

$$\chi^2(t,q) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(x_{ij} - e_{ij})^2}{e_{ij}} \qquad (6)$$

x_{ij} は表 2 の各要素の値を表し，e_{ij} はその期待値を表す．ただし，chi^2 値は語 t の記事へのリンクが一般コーパスに偏って出現している場合にも高い値を示してしまう．そのため，$D_s(q)$ 内の記事が語 t の記事へのリンクを持つ確率が $D_g(q)$ 内の記事が t の記事へのリンクを持つ確率よりも大きいという条件を課す．

表 2　2 × 2 分表

$D_s(q)$ 内で語 t の記事へのリンクを持つ記事数	$D_s(q)$ 内で語 t の記事へのリンクを持たない記事数
$D_g(q)$ 内で語 t の記事へのリンクを持つ記事数	$D_g(q)$ 内で語 t の記事へのリンクを持たない記事数

6. 評価実験

6.1 評価実験の概要

専門語抽出手法についての評価実験の概要について述べる．第 5 章で述べた専門語判定手法の入力は，検索語 q についての領域 $domain(q)$ と候補語 t である．領域 $domain(q)$ を検出する際には，4.3 で述べた予備実験についての考察を踏まえ，正規化係数 $G(N)$ として $\log N$ を，ウィンドウサイズとして $K = 3$ を採用した．

次に，専門語の候補の取得方法について述べる．ここでは検索語 q を入力としたウェブ検索結果中から専門語の候補を抽出する．まず，Yahoo!のウェブ検索 Web サービス API(注3)を利用して，検索語 q を入力として 100 件の検索結果を取得する．次に，各検索結果中のタイトルおよびスニペットから Wikipedia の見出し語となっている語を専門語の候補として抽出する．ただし,「2008 年」といった年月を表す語のように，専門語として不適切であると考えられる見出し語はフィルタリングをかけて取り除いている．このようにして得られた候補語集合 T の各要素 t について，第 5 章で述べた 2 種類のスコアリング手法を適用する．

評価実験のための検索語としては，予備実験に用いたものと同じ情報学に関する 50 個の用語を用いた．

6.2 実験結果および考察

スコアリング手法の精度を示す実験結果を図 4 に示す．横軸はスコアリングされた候補語を高い順にソートしたときの順位を表し，縦軸はその順位に現れた語が検索語の領域における専門語であるかについての平均適合率である．具体的には，各順位での平均適合率はその順位で得られた語のうち検索語の属する領域における専門語であると判断できる語の含まれる割合によって求めている．上位 10 件程度までは χ^2 値をスコアリング手法として用いたほうが全体的に高い適合率を示している．それ以下の語については，両者はほぼ同程度である．これは χ^2 値を用いたときに，専門語として適切な語がより高いスコアを示しているということを意味している．

提案手法の実行例を表 3 に示す.「SQL インジェクション」という検索語により得られたウェブ検索結果上位 100 件から,「SQL」や「脆弱性」といった 15 個の候補語が抽出された.「SQL インジェクション」の属する領域を検出し，各候補語について $tf.idf$ と χ^2 値によってスコアリングを行った．$tf.idf$ では「コンピュータ」や「インターネット」といった一般的な語が高いスコアを示し，逆に「改竄」や「脆弱性」といった情報セキュリティに関する専門的な用語が下位にランクされた．χ^2 値によるスコアリングでは，上位に情報セキュリティに関する専門語が並び，一般的な語は下位にランキングされた．

(注3)：http://developer.yahoo.co.jp/search/web/V1/webSearch.html

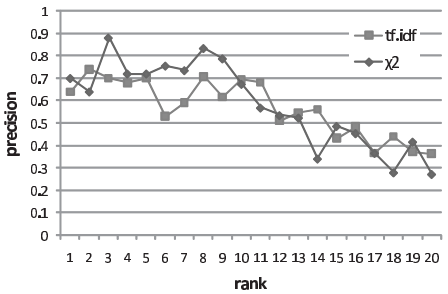

図 4　ウェブ検索結果からの専門語抽出の精度

本研究では，専門語のスコアリング手法について述べたが，上位何個または上位何%の語を専門語として抽出するかについては言及していない．どのくらいの個数の専門語が抽出できればよいかという議論は，アプリケーションに依存した話であるためである．また，ある語が専門語であるかという判断は検索語をどのような面から捉えるかに依存するという問題もある．例えば,「SQL インジェクション」という検索語からは，表 1 に示すように {“セキュリティ技術”，“SQL”，“問い合わせ言語”} という領域が検出されるが,「SQL インジェクション」を「セキュリティ技術」という観点から見るか,「SQL」「問い合わせ言語」という観点から見るかによって，専門語として判断される語は変化するだろう．一つの解決策としては，検出された領域に含まれるカテゴリ集合に対して Wikipedia のカテゴリ木の情報などを利用してクラスタリングを行い，各クラスタに本研究の提案手法を適用することによって観点ごとに相対的に専門語を求めることが考えられる．

日本語版 Wikipedia を用いて提案手法を一通り実行するのに所要する時間は，検索結果中から抽出される候補語の個数に依存するが，一つの検索語につき数秒程度であった．

7. まとめと今後の課題

本研究では，Wikipedia のリンク構造とカテゴリ構造を利用して，ユーザの入力した検索語からその語に関する専門語を抽出する手法を提案した．まず，検索語の Wikipedia 記事およびそのバックリンク先の記事のカテゴリ情報を集約することによって検索語の属する専門領域を検出する手法について述べた．次に，ある語が検索語の属する領域における専門語であるかを判定するためのスコアリング手法について述べた．評価実験においては，検索語を実際に検索エンジンに入力したときに得られる検索結果中に含まれる専門語を抽出することを行い，スコアリング手法が妥当であることが示された．

今後の展開として，本研究での提案手法によって抽出された専門語をベースとしてウェブページの「理解しやすさ」を求めることを考えている．非専門家にとって専門語の少ないページ，あるいは専門語が含まれていたとしてもその定義について述べ

表 3　ウェブ検索結果からの専門語抽出例（q=“SQL インジェクション”）

t	$tf.idf(t,q)$	t	$\chi^2(t,q)$
SQL	79.72	SQL	10777.5
コンピュータセキュリティ	64.64	コンピュータセキュリティ	6341.3
コンピュータ	61.56	ストアドプロシージャ	5984.0
データベース	32.76	脆弱性	2722.6
パスワード	31.97	改竄	2122.3
インターネット	30.99	パスワード	2006.2
リスク	27.87	セキュリティホール	1979.1
セキュリティホール	26.35	リスク	1349.4
ストアドプロシージャ	26.26	データベース	783.5
改竄	23.57	コンピュータ	431.8
ウェブサイト	22.37	情報	223.2
脆弱性	20.97	ウェブサイト	156.4
情報	19.87	インターネット	155.8
企業	10.93	日本	7.066
日本	7.68	企業	2.485

られているページは理解しやすいと考えられる．一方，専門家にとっては専門語を使わずに長々と説明されるウェブページよりも，専門語を使用して簡潔に述べられたページのほうが理解しやすいといえる．ウェブページ中に含まれる専門語に加え，画像の多さや既存の Readability テストなどを踏まえた上で，ユーザの検索語に対する理解度に合ったウェブページの発見を支援するシステムを考えている．また，本研究での提案手法は Wikipedia の記事構造のみを用いており，対象としている言語に全く依存しない．そのため他言語への拡張を容易に行うことが可能であり，理解しやすい非母国語のウェブページの発見や専門語の対訳といった応用も考えられる．

謝辞　本研究の一部は，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」，および，京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」，および，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，課題番号 18049041），ならびに，計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」（研究代表者：安達淳，Y00-01，課題番号：18049073）および，文部科学省科学研究費補助金若手研究 (B)「情報検索とウェブアーカイブにおけるマイニング」（研究代表者：Adam Jatowt，課題番号：18700111）によるものです．ここに記して謝意を表します．

文　献

[1] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama and K. Tanaka: “Trustworthiness analysis of web search results”, Proceedings of the 11th ECDL (2007).

[2] J. Giles: “Internet encyclopedia go head to head”, Nature, **438**, (2005).

[3] D. Milne, O. Medelyan and I. H. Witten: “Mining domain-specific thesauri from wikipedia: A case study”, International Conference on Web Intelligence (2006).

[4] M. Strube and S. P. Ponzetto: “Wikirelate! computing semantic relatedness using wikipedia”, Proceedings of National Conference for Artificial Intelligence (2006).

[5] K. Nakayama, T. Hara and S. Nishio: “Wikipedia mining for an association web thesaurus construction”, Proceedings of International Conference on WISE (2007).

[6] D. P. T. Nguyen, Y. Matsuo and M. Ishizuka: “Relation extraction from wikipedia using subtree mining”, Proceedings of Association for the Advancement of Artificial Intelligence (2007).

[7] D. N. Milne, I. H. Witten and D. M. Nichols: “A knowledge-based search engine powered by wikipedia”, Proceedings of the sixteenth ACM conference on CIKM, New York, NY, USA, ACM (2007).

[8] R. Mihalcea and A. Csomai: “Wikify!: linking documents to encyclopedic knowledge”, Proceedings of the sixteenth ACM conference on CIKM, ACM (2007).

[9] 合原博, 宮田高志, 松本裕治：“医学生物学分野からの専門用語の抽出・分類”, 情報処理学会研究報告. 自然言語処理研究会報告, pp. 41–48 (2000).

[10] 中川裕志, 森辰則, 湯本紘彰：“出現頻度と連接頻度に基づく専門用語抽出”, 自然言語処理, **10**, 1, pp. 27–45 (2003).

[11] Chung: “A corpus comparison approach for terminology extraction”, Terminology, **9**, pp. 221–246(26) (2003).

[12] 木田充洋, 外池昌嗣, 宇津呂武仁, 佐藤理史：“ウェブを利用した専門用語の分野判定”, 電子情報通信学会論文誌, **J89-D**, 11, pp. 2470–2482 (2006).

[13] G. Salton and C. Buckley: “Term-weighting approaches in automatic text retrieval”, Inf. Process. Manage., **24**, 5, pp. 513–523 (1988).