知能と情報（日本知能情報ファジィ学会誌）Vol.21, No.5, pp.815－826（2009）

原著論文

日本語オントロジー辞書システムOntolopediaの構築と興味抽出手法への応用検討†

宮城 良征 *1・當間 愛晃 *2・遠藤 聡志 *2

本研究では，オンライン百科事典Wikipedia（日本語）の文章をコーパスとして用い，汎用的な利用を想定した日本語オントロジー辞書システムOntolopediaを設計・構築し，システムを通して作成された知識地図を応用した評価実験を行った．

評価実験では，Twitterよりユーザーの発言情報を取得し，Ontolopediaで構築した知識地図を利用して興味関心をユーザー毎に抽出・推測し，精度や応用可能性について検証した．Twitterにおける発言データから名詞のみを抽出して生成した発言語リストを解析対象とし，提案手法による興味語の抽出･推測を行った．提案手法による興味語の抽出・推測を行い，以下の３項目について評価を行った．第一に，発言語リスト内のどの単語がより興味関心のある言葉なのかを点数化し，ユーザにとって望ましい単語を上位にランク付けられるかを検証した．比較対象としては，特徴的な単語を抽出するために広く用いられているTF-IDF法を取り上げた．第二に，構築した知識地図を利用する事で，発言語リスト内には現れていないが，発言語リスト上の単語群から興味があるだろうと推測される単語群の抽出を試みた．第三の評価実験では，概念別に語句を分類し，ユーザー毎の興味の偏りを調査することで，ユーザーの興味特性を抽出できるかを検討した．

以上の評価実験により，提案手法による興味語の抽出が従来法と比較して精度よく行える事を示し，発言語リストから興味語の推測が行えることを示した．

キーワード：オントロジー，Wikipediaマイニング，興味語抽出・推測

１．はじめに

インターネットやコンピュータネットワークが普及するに伴い，多くの人が自由に情報を発信・取得できるようになったが，それに比例してネットワーク上には膨大な情報が流通し，その中から自分の求めている情報を検索・取得する作業が難しくなった．一方，現在利用されている検索エンジンにおける主流となっている検索方法としてはキーワードマッチング方式が利用されているが，適切なキーワードを思いつけないと探し出すことが困難であり，キーワードが一致したとしても得られたランキング出力結果から目的の情報を選び出すには労力が必要となる．

Google，Yahooなどのようなキーワードに基づく検索エンジンは，Web上から情報を取得する際の主要なツールである．しかし，これらの検索エンジンに対して，以下に示す欠点が指摘されている[1]．

- **結果が検索語に依存しすぎる．**
  検索語を含むページしか出力できない．ある検索キーワードで望んだ結果が得られない場合，関連する文章は異なる用語を使って書かれている場合が多い．意味的に似た検索は，同じ結果を返すべきである．
- **再現率は高いが，精度が低い．**
  主要なページを検索できたとしても，他に関連ページが２万件など多すぎる場合がある．その関連ページには，関連する，やや関連する，あまり関連しない文章が含まれる．情報が多すぎるのも，情報が少なすぎるのと同程度，良くないと考えられる．
- **結果が単一のWebページでしかない．**
  多くの文章に分散している情報が必要な場合，検索を繰り返し行い，文章から情報を手作業で抽出し，一つの文章にまとめなければならない．
- **再現率が低いまたはゼロ．**
  検索に対し何も返ってこない場合がある．また，時間が経つと，同じ検索キーワードでも違う情報が引っかかってくるようになる．

† Construction and Use of Japanese Ontology Dictionary System Ontolopedia in the Extraction of Words of Interest
Yoshiyuki MIYAGI, Naruaki TOMA and Satoshi ENDO

＊1 琉球大学大学院理工学研究科情報工学専攻
Information Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus
＊2 琉球大学工学部情報工学科
The Department of Information Engineering, University of the Ryukyus

これらの欠点を解決するために，パーソナライズドサーチエンジンによる検索支援が効果的に機能することが期待される．パーソナライズドサーチエンジンとは，利用者の嗜好をコンピュータが抽出・推測し，それを元に検索結果を個人に最適化した形で出力するシステムである．利用者の嗜好を抽出するために，mixiやblog等の電子化された個人文書で用いられる単語群を用い，オントロジー辞書の利用により単語間の概念ネットワークを作成し，繋がり上の重みを算出することで「内在する嗜好（興味）」を単語群として推測することを目指す．しかし，汎用的に使用できる日本語オントロジー辞書が一般に公開されていない．

本研究では，このパーソナライズドサーチエンジン開発のために必要となる日本語オントロジー辞書システムを構築し，そのシステムを利用して辞書データ（知識地図）を生成する．オンライン百科事典Wikipedia（日本語）[2]の文章をコーパスとして用い，汎用的な利用を想定した日本語オントロジー辞書システムOntolopedia*1を設計・構築する．また，Twitter[3]よりユーザーの発言情報を取得し，本システムを通して作成された知識地図を利用して興味関心の抽出・推測を行う．

類似の関連研究として，Middletonら[4]では，プロキシサーバやユーザー自身による興味の有無に関する回答からプロファイルを生成し，協調フィルタリング的に推薦する際にコールドスタート問題へ対処するために外部オントロジーを利用している．これに対し，Wikipediaをベースとした外部オントロジーを利用している点では類似しているが，本提案手法ではブログエントリのように記述されたテキスト文章だけで生成可能であり，ユーザーの負担が少ない点が特徴である．また，中辻らの[5]とは，ユーザープロファイルを生成するための解析対象としてブログエントリを採用している点が類似しているが，本手法では抽出・推測される興味語にはエントリ中に記述の無かった語（以下，新規語）を推測可能である点に違いがある．これらに加えて，両手法とも特定ドメインに特化した興味オントロジーを構築しているが，本手法ではドメインを特定しておらず汎用性が高い興味オントロジーを生成することが可能である．

評価実験では，Twitterにおける発言データから名詞のみを抽出して生成した発言語リストを解析対象とし，提案手法による興味語の抽出・推測を行い，以下の3項目について評価を行う．第一に，発言語リスト内のどの単語がより興味関心のある言葉なのかを点数化し，ユーザにとって望ましい単語を上位にランク付けられるかを検証する．比較対象としては，特徴的な単語を抽出するために広く用いられているTF-IDF法[6]を取り上げている．第二に，構築した知識地図を利用することで，発言語リスト内には現れていないが，発言語リスト上の単語群から興味があるだろうと推測される単語群の抽出を試みる．第三の評価実験では，概念別に語句を分類し，ユーザー毎の興味の偏りを調査することで，ユーザーの興味特性を抽出できるかを検討する．

以上の評価実験により，提案手法による興味語の抽出が従来法と比較して精度よく行えることを示し，発言語リストから興味語の推測が行えることを示す．

2．システム概要

Ontolopediaシステム概要を示す（図1）．Ontolopediaとは，Wikipediaのダンプデータを解析してオントロジーの基礎となるデータを自動生成し，ユーザーがWebインターフェースを介してデータベースを手動修正することで精度の高い日本語オントロジー辞書を構

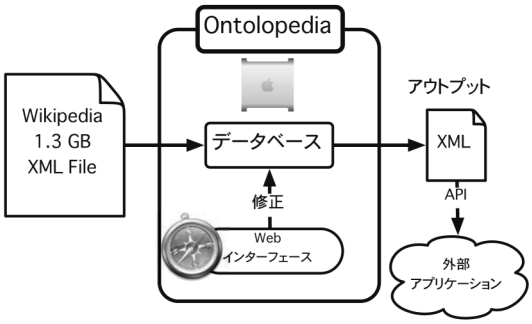


図1　システム概要

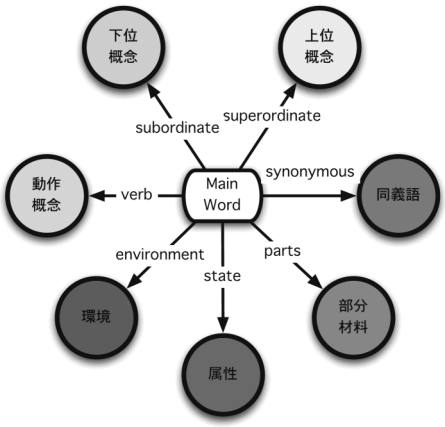


図2　Ontolopedia概念構造

*1 http://www.nal.ie.u-ryukyu.ac.jp/ontolopedia

築するためのWebシステムである．Ontolopediaシステムにより構築されるオントロジーの概念構造(図2)を「Ontolopedia概念構造」と呼び，この構造に基づき構築されたデータ集合を「知識地図」と呼ぶ．なお，3節で詳しくは述べるが，Ontolopediaに格納されたデータベースからXMLファイルを出力することで，外部のアプリケーションから利用することも可能である．

2.1 Wikipediaから知識地図を構築

Wikipediaのコンテンツは全て再配布や再利用のためにデータベース・データの提供が行われている．Wikipediaの提供しているダンプデータには数種類あり，本研究では2006年9月27日に作成された「jawiki-latest-pages-articles.xml.bz2」を使用した．これは，494, 854ページの記事で構成された，1.3GBの一つのXMLファイルである*2．このXMLのデータ解析手順を以下に示す．

1. **各ファイルのtitleタグに囲まれている語句を抽出**：分割した各ファイルのtitleタグで囲まれている語句を抽出する．この抽出した語句を本稿ではMainWordと呼び，MainWordを中心語として概念構造を構築する．
2. **各ファイルのtitleに関連する語句を抽出**：各ファイル毎にページタイトルに関連する語句として「Wiki形式の太文字」「Wiki形式のリンク」「Categoryタグ」の3種類を抽出する．

   - **Wiki形式のリンクを抽出**：pageとpageを接続するために，Wiki形式では‘[[...]]’の間にpageタイトルを記述する．このpageタイトルとは，titleタグではさまれている語句のことである．このリンクが張られている語句は，pageタイトルにとって重要な語句だと考えることができる．しかし，この語句がどの概念にあたるのか，コンピュータに判断させるのは難しいので，「未分類(重要)」に分類する．
   - **Wiki形式の太文字を抽出**：リンクを張られていない言葉でも，強調した言葉は重要な場合がある．言葉を強調する場合，Wikiではシングルクオーテーション3つで語句を囲む．これもWiki形式のリンクと同様に重要だと考え，「未分類」に分類する．
   - **‘Category:’から始まるWiki形式のリンクを抽出**：Categoryへのリンクがある場合，この語句はページに対する上位概念だと考えることができる．そこで，‘Category:’または‘category:’が含まれる場合，これを上位概念に当てはめるようにする．

「未分類」または「未分類(重要)」として自動分類された語句は，OntolopediaのWebインターフェース(図1)を利用して手動で修正を行う．

2.2 概念

図2に概念構造を示す．同図の中心にあるMainWordは，Wikipediaにおけるページ名(語句)が対応する．この語句を中心に概念を形成していく．「上位概念」「下位概念」「類義語」「部分材料」「動作概念」「属性」「環境」の7つから構成される．また，図2の各矢印上の英単語は，XML出力時の各概念を表記するためのタグを示している．

構成する各概念について解説する．図2の中心に位置する‘MainWord’に関する概念を形成するために，以下に述べる概念に分類する．また，実際に分類した例として「飛行機」について概念を形成する場合を考える．

- **上位概念：**‘MainWord’の上位にあたる語句．
  「飛行機はhogeの一つだ．」と表現できるもの(機械，飛行物体)
- **下位概念：**‘MainWord’の下位にあたる語句．
  「hogeは飛行機の一つだ」と表現できるもの(戦闘機，輸送機，旅客機，F-22，ジャンボ．..)
- **同義語：**‘MainWord’と同義の語句．
  「飛行機とhogeは同義である．」と表現できるもの(航空機．...)
- **部分材料：**‘MainWord’を構成する語句．
  「飛行機を構成する要素」(エンジン，主翼，尾翼，胴体．...)
- **属性：**‘MainWord’がどのような様子か．
  「飛行機がどのような様子か」(便利，重い，難しい，言葉がたくさん，楽しい．...)
- **環境：**‘MainWord’はどのような環境にあるか，どのような環境で使用されるか．
  「飛行機がある場所」(空，空港．...)
- **動作概念：**‘MainWord’は何をするのか，されるのか．
  「飛行機に関係する動作．飛行機をhogeする．飛行機にhogeする．」(操縦する，乗る．...)

*2 http://ja.wikipedia.org/wiki/ 赤リンク

表 1 Ontolopediaにて構築された知識地図

概念	個数	edge
MainWord	435,130	
上位概念	23,826	589,450
下位概念	856	905
類義語	376	976
部分材料	877	976
動作概念	37	37
属性	132	139
環境	474	592

表 1 は Ontolopedia にて構築された概念地図のデータ（登録された単語数ならびに単語間に接続されたエッジ数）を示す．

3. Ontolopedia API

外部アプリケーションとの連携を目指し，APIを公開した．このAPIはHTTPを使用したREST（Representational State Transfer）型APIであり，URLを指定してリクエストされたデータを提供する．Ontolopediaが返すデータはXML形式となっている．

現在，データを取得するための 3 種類のAPIを提供している．これらのAPIは Ontolopedia の情報を引き出すのに使用される．

- **概念検索：** MainWord のリストから部分一致で語句を検索することができる．出力は検索結果の語句のリストで返す．
- **完全一致検索：**「概念検索」と同様にMainWordのリストから検索するが，出力は検索に一致した語句の概念を返す．
- **データ取得：** MainWord に付随するユニークなIDを指定して概念データを取得する．

例えば「琉球大学」というMainWordについて概念検索*3・完全一致検索*4・データ取得*5 をするには各々脚注に示すURLを指定することでXML形式で参照結果を取得することができる．

日本語を使用して検索する場合は，検索キーワードをURLエンコード（UTF-8）する必要がある．本システムのドキュメント*6 にて，このAPIを使用するためのRubyライブラリを公開している．

*3 http://www.nal.ie.u-ryukyu.ac.jp/ontolopedia/api/search?keyword =琉球大学
*4 http://www.nal.ie.u-ryukyu.ac.jp/ontolopedia/api/matchfull?keyword =琉球大学
*5 http://www.nal.ie.u-ryukyu.ac.jp/ontolopedia/api/show/4657
*6 http://www.nal.ie.u-ryukyu.ac.jp/ontolopedia/document/api.html

4. 興味語の抽出・推測方法

4.1 既存語ランキング

図 3 を参照して，Ontolopedia を使用して興味関心の度合いをポイントとして算出するためのリスト構築方法について説明する．“リスト:WordCount”は語句とその出現回数をポイントとして扱ったリストである．“リスト:OpPoint”は“WordCount”を元に Ontolopedia から概念を取得し，ポイントを加算して再構築した語句のリストである．これは既存の語句のみのリストであり，新規に出現した語句は含まない．

次に，“OpPoint”の構築方法を述べる．

1. 各ユーザーの“語句”と“語句の出現数”の“リスト:WordCount”を取得する．
2. WordCount を複製して“OpPoint”の初期値として扱う．
3. Ontolopedia API を使用して，ユーザーの発言した語句 B の概念を取得する．
4. 語句 B から上位概念で繋がっている語句 G があるとする．
5. この語句Gがユーザーの発言したリスト内にあれば，ポイントを OpPoint の語句 G に加算する．

3 から 5 を全ての語句について繰り返し，ポイントの降順にソートする．次にポイントの加算方法について説明する．最初に適当にパラメーター param の値

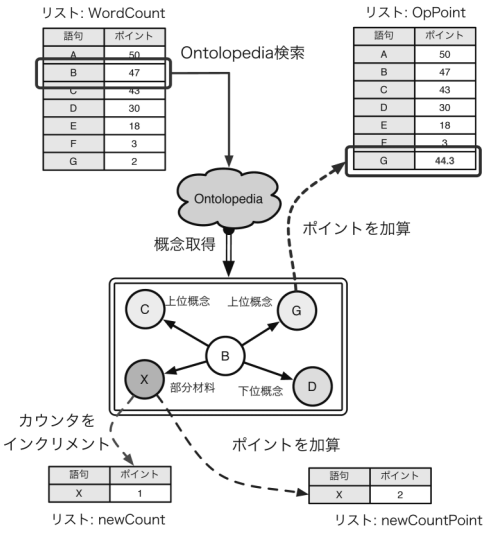


図 3 ユーザーの語句をOntolopediaを使って操作する

を決める．今回は0.3とした．次のリストにあるように，各概念の重みを決める．今回は“同義語＞下位概念＞上位概念＞その他概念”となるようにした．

- 上位概念(superordinate)：3
- 下位概念(subordinate)：5
- 同義語(synonimous)：8
- その他概念(others)：2

MainWordである語句Bの上位概念である語句Gの加算式は次のようになる．

$$\begin{aligned} P_{new}(G) &= P_{new}(G) \\ &+ P_{old}(B) \cdot param \\ &\cdot weight(superordinate) \end{aligned} \quad (1)$$

$P_{new}(X)$ ： リスト：$OpPoint$ の語句 X のポイント
$P_{old}(X)$ ： リスト：$WordCount$ の語句 X のポイント
$param$ ： 任意のパラメータ (0.3)
$weight$ ： 概念の重み

これを，ユーザーの発言した各語句について繰り返し行い，ポイントの高い順に興味関心の度合いが強い単語として抽出する．なお，概念を一つも持たないMainWordの場合には，式(1)による加算が行われず，WordCountの値がそのままOpPointとして採用される．

4.2 新規語推測

Ontolopediaの概念取得によって新規に出現した語句を用いてランキングを作成する方法について説明する．図3の語句Xが新規に出現した語句にあたる．新規語は2種類のランキングリストを作成する．

- 新規に出現した語句の出現回数：newCount
- 新規に出現した語句を概念別にポイント加算：newCountPoint

“newCount”はユーザーの既存語にない語句が出現した場合，カウンタをインクリメントする．このカウンタをポイントとして，リストを降順にソートする．

“newCountPoint”も同様に，ユーザーの既存語にない語句が出現した場合に，計算式を通してポイントを加算する．基本的には“OpCount”を作成する際に使用した計算式と同じだが，“WordCount”のポイントを乗算しない．これを省くことで，元の語句のポイントを重視するのではなく，知識地図の概念構造を重視したリストになる．次に図3の場合の語句Bと語句Xの計算式を示す．この計算式によりポイントが高く算出された単語ほど，新規語として興味関心の度合いが高い単語として推測する．

$$P_{new}(X) = P_{new}(X) + param \cdot weight(others) \quad (2)$$

$P_{new}(X)$ ： リスト：$newCountPoint$ の語句 X のポイント
$param$ ： 任意のパラメータ (0.3)
$weight$ ： 概念の重み

5．評価実験

5.1 実験設定

Ontolopediaで構築した知識地図を利用して興味関心をユーザー毎に抽出・推測できるかどうか調べる．Ontolopediaからの概念取得には，作成したOntolopedia API を利用する．知識地図を利用することで，ユーザーが蓄積した情報に無い新しい語句を引き出せることが期待できる．また，話題に触れている回数は少ないが興味のある事柄を抽出することも期待できる．

Twitter[3]よりユーザーの発言情報を取得し，Ontolopediaで構築した知識地図を利用して興味関心をユーザー毎に抽出・推測し，精度や応用可能性について検証する．検証方法として，2008年10〜12月のTwitter上の発言データを使用して2009年1月に15人を対象にアンケートを行った．15人の被験者は，Twitterユーザーであることを前提とした上で，発言頻度が高いユーザーと低いユーザーがばらけるように選出した．またアンケートは解答用ファイルをWeb経由で配布し，回収した．このアンケートを用いて各ユーザーにとって興味関心が引き出されているか評価する(図4)．

次に，評価用アンケートの内容を説明する．ユーザー毎の発言データから名詞・名詞と思われる語句を抜き出し，発言語リストを作成する．語句の抽出には形態素解析エンジンMeCab[7]のIPA辞書をベースとし，これに加えてOntolopediaにてMainWordとして登録されている単語(Wikipedia上でページ名として登

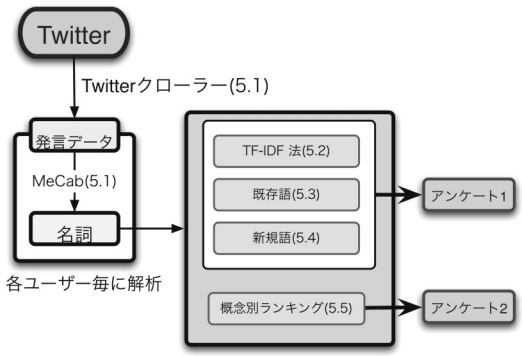


図4 評価実験の流れ（図中の(x.y)等と示しているのは本文中で解説している章節番号を示している）

録されている単語）を名詞として追加登録した辞書を利用して抽出している．抽出した単語集合を利用して次のリストのデータを作成する．

1. TF-IDF 法を用いて作成されたランキングデータ
2. 単語の出現回数（WordCount）を Ontolopedia の知識地図の重みを使用して順位を入れ替えたデータ（OpPoint）
3. ユーザーの発言を元に，Ontolopedia を使用して新規語を取得したデータ（newCount）
4. ユーザーの発言を元に，Ontolopedia を使用して新規語を取得するとともに，知識地図の重みを使用して並び替えたデータ（newCountPoint）
5. ユーザーの発言を元に，Ontolopedia を使用して概念別に語句を分類したデータ

ユーザーへのアンケートには，どのように抽出したのかわからないように順番をランダムに並び替えてリスト化を行った．アンケートは 2 回行い，最初のアンケートは 1，2，3，4 の各データから上位50件をアンケートに使用した．このアンケートについては次の方法で評価する*7．

- 既存語抽出におけるランキング比較（TF-IDF法と Ontolopedia 利用の比較）
- 新規語推測における妥当性比較（Ontolopediaカウンター利用と Ontolopedia 知識地図の重み利用の比較）

2 回目のアンケートは 5 のデータを使用して作成した．各概念において，上位50件を使用してリストを作成した．50件に満たない場合はそのまま使用する．このアンケートを利用してユーザー毎の興味の偏りを調査する．

アンケートの答え方は，リストの各語句に対して，数字または空欄を記入してもらう．曖昧な回答を避けるように，回答は 3 つに絞り，回答「わからない/意味不明な語句」は，不要な文字（“こと”等）や難解な語句（“マクロ解剖学”等）を分類できるようにした．

- 1: 興味あり
- 2: 興味なし
- 空欄: わからない/意味不明な語句

*7 アンケートを実施する際には，二重回答にならないように語句の重複を避けて作成した．

5.2 実験パラメータの設定

式(1)や式(2)における最適なparamや各概念の重みを求めるために次の実験を行った．

計算を行う前にWordCountを複製し，知識地図利用の初期値として利用する．ある語句 A の上位概念として語句 G を取得できた場合，次の式(3)となる．

$$\begin{aligned} P_{new}(G) &= P_{new}(G) + \\ &\quad P_{old}(A) \cdot param \\ &\quad \cdot \ weight(superordinate) \end{aligned} \tag{3}$$

$P_{new}(X)$: リスト：知識地図利用の語句 X のポイント
$P_{old}(X)$: リスト：出現回数の語句 X のポイント
$param$: 任意のパラメータ
$weight$: 概念の重み

この式の特徴として，MainWordとして用いられた語句のポイントは増加しない．関連語として呼び出された場合にポイントが増加する．

このparamと各概念の重みを決定するために筆者自身のデータを用いて決定した．筆者の興味のある単語を20個挙げ，各 param（0.1, 0.3, 0.5, 1）を使用して計算し，順位がどう変化するか調査した．このとき，各概念の重みは次の値とした．

- 上位概念（superordinate）：10
- 下位概念（subordinate）： 5
- 同義語（synonimous）： 8
- その他概念（others）： 2

表 2 に各paramにおける順位を示す．全てのparamにおいて，既存の語句のリストである従来の手法であるTF-IDF 法では発見できなかった“ps3”と“ラーメン”が新しくリストに出現したことがわかる．param が 1 または0.5の時には，興味語句の順位を大幅に減

表 2 param を決めるための実験結果

興味のある語句	TF-IDF 法	0.1	0.3	0.5	1
chumby	383	579	735	962	975
emacs	254	376	584	603	833
flash	148	245	392	499	744
Google	27	16	17	20	23
iPhone	1	2	14	46	99
iPod touch	36	77	162	228	315
mac	60	128	219	295	437
ps3	なし	2144	2144	2144	2144
rails	851	1118	1360	1360	1362
ruby	111	208	317	350	440
twitter	29	60	150	212	303
youtube	179	169	176	175	170
沖縄	13	22	86	126	197
カメラ	106	78	84	80	81
ガンダム	41	25	26	28	32
プログラミング	374	37	18	16	16
マクロス	113	257	402	511	756
マブヤー	なし	なし	なし	なし	なし
ラーメン	なし	3331	3331	3331	3331
琉球大学	2566	1067	632	519	434

表3 上位概念を除いた実験結果

興味のある語句	TF-IDF 法	param 0.3
chumby	383	414
emacs	254	261
flash	148	161
Google	27	39
iPhone	1	2
iPod touch	36	49
mac	60	77
ps3	なし	1692
rails	851	854
ruby	111	113
twitter	29	40
youtube	179	51
沖縄	13	19
カメラ	106	124
ガンダム	41	55
プログラミング	374	442
マクロス	113	170
マブヤー	なし	なし
ラーメン	なし	2985
琉球大学	2566	312

少している．唯一，“プログラミング”の順位だけが上がっていることがわかる．

param0.1とparam0.3の値で順位をそれぞれでソートした場合，語句の並びが，param0.3の方が筆者の興味の度合いに合っていると感じたのでデフォルトのparamを0.3に設定した．

次に，各概念の重みの決め方について述べる．paramを決めるための実験をしながら，出力される語句を眺めると，ある語句の上位概念にあたる語句が多くランキングの上位に出ており，他の概念にあたる語句が上位に出ていなかった．そこで，上位概念を除いて計算を行った．表3にその結果を示す．上位概念を除いた結果でも，新規語として“ps3”，“ラーメン”を推測することができた．また，私の興味語リストの順位も上昇した．このことから，上位概念の重みが大きいと，他の概念の語句がランキング上位に行きづらくなるのでは無いかと考えた．さらに，同義語はユーザーの既存の語句と同じ意味を示すので，重みを大きくし，次に既存の語句をさらに詳しくする説明として下位概念を同義語の次に重みを置くことを考えた．このことから，各概念の重みを“同義語＞下位概念＞上位概念＞その他概念”となるように設定した．

5.3 既存語ランキング比較評価

アンケートをTwitterを使用している15人に行い，

表4 アンケート結果：既存語ランキング比較（興味あり）

id	発言数	単語数	TF-IDF/興味あり	OpPoint/興味あり	興味あり差分
A	19	70	12	10	2
B	83	172	22	25	-3
C	155	255	26	23	3
D	204	376	22	21	1
E	237	713	28	31	-3
F	309	645	10	10	0
G	391	546	17	25	-8
H	443	903	32	30	2
I	619	834	30	33	-3
J	966	1239	41	41	0
K	1043	1653	17	21	-4
L	2462	2890	29	37	-8
M	2980	2027	30	37	-7
N	3216	3358	26	39	-13
O	3221	2739	10	13	-3

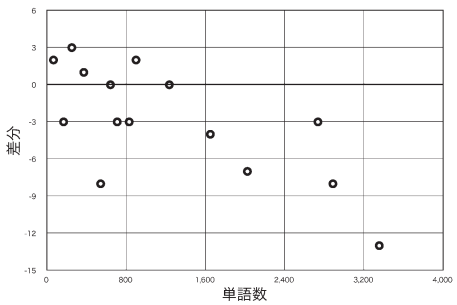


図5 単語数と差分の分布図

既存語と新規語，概念別に集計し評価を行った．表4はTF-IDF法とOpPointの上位50件の“興味あり”と答えた語句のユーザー毎の個数となっている．同表における発言数とはTwitterにおける発言データのポスト数であり，単語数とはMeCabにより抽出された単語数を意味しており，重複を除いたユニーク数を示してる．“興味あり差分”はTF-IDF法の“興味あり”とOpPointの“興味あり”の差分の値である．

TF-IDF法とOpPointの“興味あり差分”と単語数の関係を分布図にしたのが，図5の分布図である．差分の値が負になれば，TF-IDF法よりもOpPointの方がユーザーの興味を適切に抽出していることを示す．

- 結果が良くなったユーザー：9人
- 変わらなかったユーザー：2人
- 結果が悪くなったユーザー：4人

過半数のユーザーで，TF-IDF法よりもOntolopediaの知識地図と組み合わせた方が結果が良くなった．これらの結果より，発言した単語のバリエーションが多い方がユーザーの興味を抽出し易いと考えられる．

5.4 新規語抽出評価

表 5 はnewCount（新規語の出現回数）とnewCount-Point（出現回数と知識地図）の上位50件の“興味あり”と答えた語句のユーザー毎の個数となっている．“興味あり差分”は newCount の“興味あり”とnewCount-Pointの“興味あり”の差分の値であり，負の値が大きい程 newCountPoint が良い結果を示している．

表 6 はnewCount とnewCountPoint の差を，回答中の“興味あり”が占める割合によって分類されたユーザー数により示している．

表 5 と表 6 の結果から，ユーザーの蓄積した情報から，Ontolopedia の知識地図を使用することで，新規にユーザーの興味語を推測できることがわかった．また，newCountPoint はnewCount と比較すると，“興味あり”の割合が75％以上の人数が増えている．また，25％未満のユーザーが 0 になっていることがわかる．

既存語と同様にnewCount とnewCountPoint の差分を取る．出現回数のみの場合と，出現回数と知識地図を利用した場合とを比較して，知識地図を利用した方が結果が良くなっているユーザーは 6 人，変わらなかったユーザーが 4 人，悪くなったユーザーが 5 人となっている．このことから，概念別に重みを変えて計算することで，ユーザーの興味を抽出する手助けになることが考えられる．しかし，全員が同じ重みで計算しては newCountPoint のように“興味あり”が減り，興味抽出できないユーザーも表われる可能性がある．ユーザー毎の概念の重みを変化させるために，次に示す興味語抽出における概念別評価を行った．

表 5 newCount, newCountPointの“興味あり”の差分

id	発言数	単語数	newCount/興味あり	newCountPoint/興味あり	興味あり差分
A	19	48	11	14	-3
B	83	143	28	25	3
C	155	318	20	21	-1
D	204	451	31	26	5
E	237	1307	37	39	-2
F	309	530	14	16	-2
G	391	633	38	40	-2
H	443	334	42	38	4
I	619	979	46	39	7
J	966	793	36	36	0
K	1043	1383	25	25	0
L	2462	2318	37	39	-2
M	2980	2027	35	35	0
N	3216	3505	38	38	0
O	3221	2846	19	18	1

表 6 “興味あり”が占める割合により分類した結果

割合	newCount	newCountPoint
75％以上	4	6
50％以上〜75％未満	7	5
25％以上〜50％未満	3	4
25％未満	1	0

5.5 興味語抽出における概念別評価

ユーザー毎に概念別カウンタを作成することで，ユーザーの興味がどの概念に偏っているかを次に示す方法で調査した．図 6 を参照して，概念別ランキングの作成方法を説明する．“newCount”，“newCount-Point”を作成するときと同様に，カウンタをインクリメントするが，この場合は各語句に対して概念別にカウンタを持っている．今回は，既存語と新規語を区別せずに，MainWordである語句 B の概念であればそれを概念カウンタに追加し，カウントする．

これを，ユーザーの発言語全てについて行い，概念別に降順にソートしランキングを作成する．各概念の上位50件を使用して，アンケートを作成し，2 回目のアンケートとして回答してもらった．

1 回目のアンケートにて使用した語句は回答の重複が起こらないように取り除き，集計するときに合算して集計した．集計した結果，ユーザーが幾つかに分類できることがわかった．ユーザー毎に概念分類別の評価を行い，ユーザーの特徴からタイプ分類を行った結果を図 7 〜図10に示す．作図にあたり，検証段階においてOntolopedia の知識地図によって引き出された情報が少ない場合は，参考データにならないと考えたため，推測された語句数が20未満となる概念を興味語として使用しないように切り捨てた．図 7 〜図10において空白となる概念があるのはこのためである．

全ユーザーに共通している結果としては，「同義語」概念に関しては興味を推測しやすい傾向にあることが挙げられる．元になる語句（MainWord）と同じ意味を指す概念なので，ユーザーの興味語である可能性が高

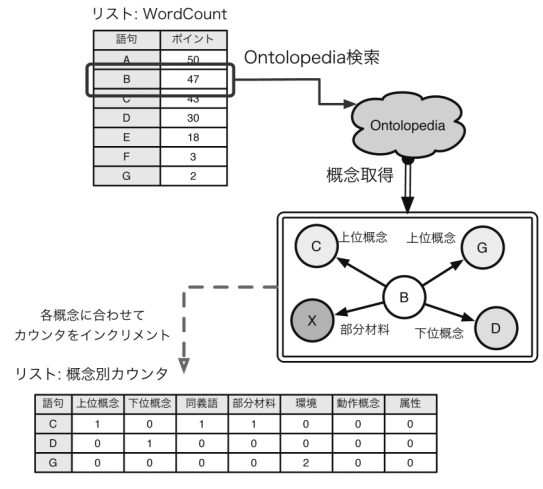


図 6 概念別ランキングを作成する

い．しかし，動作概念においては全ユーザーとも取得できてないことがわかる．Ontolopedia 自体にこの概念へ登録されている語句数が他と比べて少ないことが原因だと考えられる．

次に，ユーザータイプ毎の結果に関して考察する．

- **タイプ1：** 十分な興味語推測が出来なかったユーザの例(15人中 4 人が該当)
  図 7 は先述の理由により推測された語句数が少ない概念を取り除いて表示しており，その結果上位概念のみとなったユーザータイプの評価例を示している．これらのユーザーに共通することは，発言に含まれる単語数が他のユーザーと比べて少なく，図 5 に示した分布における左端 4 人が該当していた．発言単語数が少な過ぎた結果，それを MainWordとして取得できる概念も少なくなり，ほかの概念がデータとして使えるほど語句を取得できない結果となった．評価の良し悪しはユーザ毎に異なり，興味があると評価された割合は25%から50%とばらつきが見られた．
  なお，タイプ 1 以外のユーザに関しては十分な量の新規語推測を行えていたことから，本システムで構築した知識地図のサイズ(表 1)では500単語程度の発言単語が必要であることが分かった．

- **タイプ2：** 特定の概念に興味が偏っていたユーザーの例(15人中 6 人が該当)
  図 8 は，複数の概念において新規語として十分な量を推測でき，概念毎の評価に大きな偏りが見られたユーザータイプの評価例を示しており，最も多くのユーザーがタイプ 2 として分類された．これは，ユーザーの発言単語と表面上には現れない興味語との間には偏りがある可能性が高いことを意味すると考えられる．すなわち，発言された単語を全て同一の重みとして扱うのではなく，例えば図 8 として評価したユーザーの興味語を推測する際には「上位概念」と「部分材料」に属する概念を他概念と比較して相対的に高く設定することで良質な推測結果を得られると推察できる．

- **タイプ3：** 興味の偏りが見られず，殆どの概念で興味ありと解答したユーザの例(15人中 3 人が該当)
  図 9 は，概念毎の評価に大きな差が見られず，かつ，どの概念に対しても高評価をしたユーザータイプを示している．このタイプに属するユーザーは，どの概念においても80%前後を「興味あり」と解答していることから，発言単語と興味語との間に強い正の相関があり，好みのはっきりしているユーザーであると考えられる．

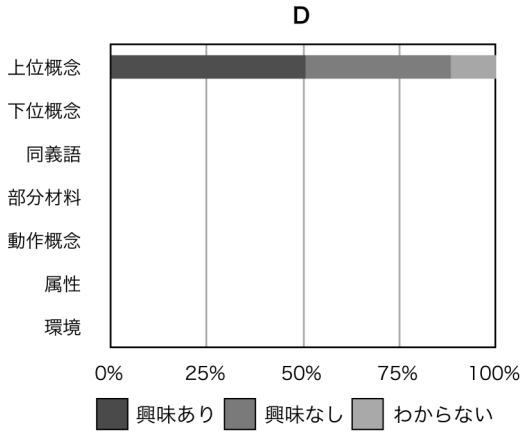


図 7 タイプ 1 の評価例

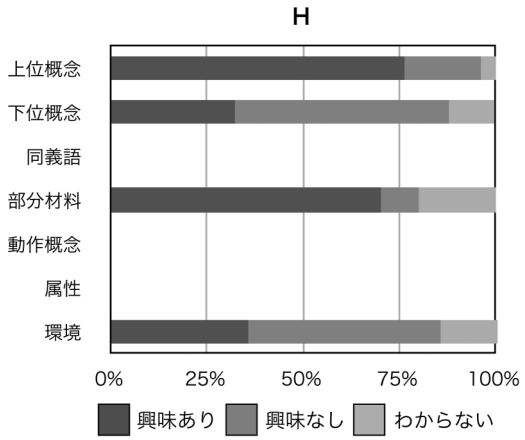


図 8 タイプ 2 の評価例

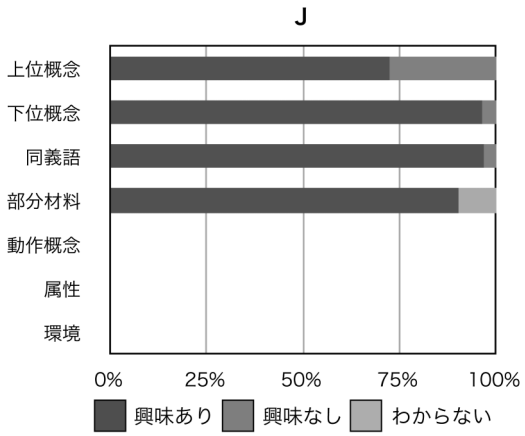


図 9 タイプ 3 の評価例

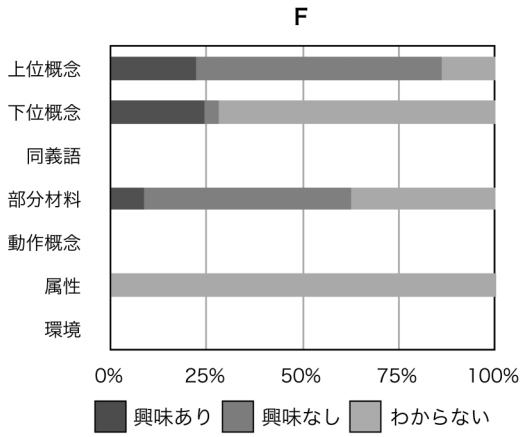


図10 タイプ4の評価例

- **タイプ4：** 興味の偏りが見られず，殆どの概念で興味なしと解答したユーザの例(15人中2人が該当) 図10は，概念毎の評価に大きな差が見られず，かつ，どの概念に対しても低評価をしたユーザータイプを示している．このタイプに属するユーザーはタイプ3とは真逆であり，発言単語と興味語との間に強い負の相関が見られた．新規語を推測することが困難ではあるが，「興味の無い概念」がはっきりとしていることから，それ以外の概念の重みを高くすることで改善することが可能であると思われる．

以上の結果から，ユーザー毎に“興味あり”に対する概念の偏りを観察出来ることが分かった．この偏りを利用してユーザー毎に概念の重み付けを変化させることで，さらに精度の高い結果が得られるものと考えられる．

これらの結果から，ユーザーに特化した興味語を推測するためには，このアンケート結果から推察される興味のある概念を抽出し，その概念を通した関連語を導き出す必要があることが分かった．

6．まとめ

本論文では，汎用日本語オントロジー作成のための日本語オントロジー辞書システム Ontolopedia の構築を行い，そのシステムを使用して構築された知識地図を利用してユーザーが蓄積した情報からユーザーの興味語を抽出・推測する応用技術の検証を行った．

評価実験では，本手法とTF-IDF法により抽出した興味語を比較し，ユーザーの発言した語句集合からより興味の高い語句を精度良く抽出できることを確認した．また，知識地図を用いることで発言語リストには現れていない新たな興味語を，高い精度で推測できることを示した．さらに，その推測精度を高めるためには，アンケート結果から推察される興味のある概念を抽出し，その概念を通した関連語を導き出すことで改善できる見込みがあること示した．

今後の展開としては，抽出・推測できた単語群を利用することで，ユーザー毎に最適化されたパーソナライズドサーチエンジンの一例である検索支援に応用することが考えられる．例えば，ユーザーの入力した検索語をコンピュータが関連語で補完し，ユーザーが欲しているであろう情報を提供できるようになるのではないだろうか．

一方，現時点の Ontolopedia は名詞を中心としてオントロジーを構築したが，名詞以外にも動詞，形容詞も含めて，引き続き汎用日本語オントロジー辞書の構築を行う必要がある．また，現在人手で行われている作業があるため，それらをプログラムを使用して自動化できることに期待する．

本研究の一部は，2008年度琉球大学工学部若手研究者等研究支援経費(プロジェクトコード：08G01002)の助成を受けたものである．

参 考 文 献

[1] G. Antoniou and F. van Harmelen. A Semantic Web Primer The MIT Press, 2004
[2] 多言語オンライン百科事典 Wikipedia 日本語版：http://ja.wikipedia.org/wiki/
[3] Twitter: http://twitter.com
[4] Middleton, S.E., Shadbolt, N.R., De Roure, D.C., Ontological user profiling in recommender systems, ACM Transactions on Information Systems, Vol.22, Issue 1, pp.54-88, 2004
[5] 中辻 真，三好 優，大塚 祥広，ユーザ興味オントロジ抽出によるブログコミュニティ形成手法，日本データベース学会 letters Vol.5, No.1, pp.33-36, 2006
[6] G. Salton and M. McGill. Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1983
[7] Yet Another Part-of-Speech and Morphological Analyzer: http://mecab.sourceforge.net/

(2009年2月17日 受付)
(2009年8月5日 採録)

[問い合わせ先]
〒901-0213 沖縄県中頭郡西原町字千原1番地 琉球大学情報工学科
琉球大学大学院理工学研究科情報工学専攻
宮城 良征
TEL：098-895-8830
FAX：098-895-8727
E-mail：yosshi@eva.ie.u-ryukyu.ac.jp

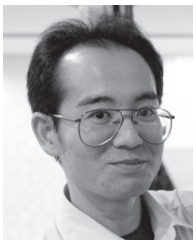
著 者 紹 介

宮城 良征［非会員］

2007年3月琉球大学工学部情報工学科卒業．2009年3月琉球大学大学院理工学研究科終了．現在，株式会社ドワンゴに入社．ソフトウェアエンジニアとしてWebアプリケーションの開発に従事している．

遠藤 聡志［正会員］

1990年北海道大学大学院工学研究科電気工学専攻修士課程修了．同年，北海道大学工学部助手，1995年琉球大学工学部情報工学科講師，1996年同助教授，2004年同教授．複雑系工学に関する研究に従事．日本知能情報ファジイ学会，情報処理学会，人工知能学会，計測自動制御学会各会員．博士(工学)．



當間 愛晃［非会員］

2003年3月琉球大学大学院理工学研究科総合知能工学専攻修了．博士(工学)．2003年4月から2007年3月までの期間は，財団法人沖縄県産業振興公社・デジタルアニメ制作推進員，有限会社トップテクノロジー・非常勤研究員，琉球大学工学部情報工学科・雇上研究員，琉球大学総合情報処理センター・雇上研究員等に従事．2004年10月琉球大学工学部情報工学科助手，2007年4月同助教．複雑系工学・Webインテリジェンスに関する研究に従事．情報処理学会，人工知能学会各会員．

Construction and Use of Japanese Ontology Dictionary System Ontolopedia in the Extraction of Words of Interest

by

Yoshiyuki MIYAGI, Naruaki TOMA and Satoshi ENDO

Abstract：

Ontolopedia, a Japanese-language ontology dictionary system intended for general use, was designed and constructed using the text of the Japanese-language version of the online encyclopedia Wikipedia as the corpus. In the present study, we evaluated the application of a knowledge map created from this dictionary system for accurately extracting and predicting the interests of individual users.

The interests of each user were extracted and predicted from comments acquired via Twitter using the knowledge map constructed from Ontolopedia. Only nouns were extracted from the acquired data. Using the proposed technique and prepared vocabulary lists, words of interest were extracted and predicted for the following three items. First, each word in the vocabulary list was assigned a numerical score based on level of interest, and higher rank for words that generate more interest by users was verified. In order to verify these points, we compared our method with the widely used TF-IDF method. Second, we verified whether extraction of interest words was possible or not. The point of this discussion was carried out a situation in which the interest words don't appear directly in the vocabulary list. Third, we considered whether user interest characteristics could be determined by investigating interest biases with classified concepts on the knowledge map.

As a result, the three kinds of evaluations showed that the proposed method aorded greater accuracy in extracting words of interest compared to conventional methods. Furthermore, we showed that the user interest biases could be observed. This observation result will improve the accuracy in extracting words of interest.

Keywords： Ontology, Wikipedia(ja) mining, Interests Extraction and Prediction

Contact Address：**Yoshiyuki MIYAGI**

Information Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus

The Department of Information Engineering, University of the Ryukyus, 1 Senbara, Nishihara, Nakagami, Okinawa, JAPAN

TEL ：098-895-8830

FAX ：098-895-8727

E-mail ：yosshi@eva.ie.u-ryukyu.ac.jp