DBSJ Journal, Vol.11, No.3
February 2013

# 語の認知度と同位語間の関係に基づく意外な情報の発見

## Discovering Unexpected Information on the basis of Popularity of Coordinate Objects and their Relationships

佃 洸摂♥　　大島 裕明♦
山本 光穂♣　　岩崎 弘利♣
田中 克己♦

Kosetsu TSUKUDA　Hiroaki OHSHIMA
Mitsuo YAMAMOTO　Hirotoshi IWASAKI
Katsumi TANAKA

**本稿ではユーザが与えたクエリに対して，そのクエリに関する意外な情報を発見する手法の提案を行う．提案手法では例えば“落合博満”というクエリに対して，“ガンダム”という語の意外度が高い，のようにクエリに対して意外度の高い関連語を求める．それに基づき，“落合博満はガンダムマニアである．”という文を発見し意外な情報としてユーザに提示する．実験では75のクエリを用いて評価を行い，クエリと関連語それぞれの同位語間の関係，および各関連語の認知度を考慮することの有効性を示した．**

**This paper provides and evaluates methods for discovering unexpected information for a keyword query. For example, if the user inputs “Hiromitsu Ochiai,” our system first discovers the unexpected related term “Gundam” and then returns the unexpected information “Hiromitsu Ochiai is a Gundam mania.” Experimental results show that it is effective to consider (i) the relationships of coordinate terms of both the keyword query and related terms, and (ii) the degree of popularity of each related term for discovering unexpected information.**

## 1. はじめに

ユーザが検索エンジンにクエリを入力して検索を行う際，検索エンジンが返すウェブページ集合にはクエリに関するよく知られた情報から意外な情報まで様々な情報が含まれる．例えば，クエリが“落合博満”という語であった場合，“落合博満は首位打者を獲得したことがある．”という情報は多くの人が知っている情報であるが，“落合博満はガンダムマニアである．”という情報は一般にあまり知られていない意外な情報であるといえる．クエリに関する一般的な情報は，検索結果の上位のページに記述されていることが多いため，ユーザは上位のページを閲覧することでクエリに関する一般的な情報を取得できる．一方，クエリに関する意外な情報は検索結果の下位のページに記述されているか，上位のページに記述されていても，そのページ内の他の多くの情報に埋もれてしまい，発見が困難であるという問題がある． 意外な情報を発見することで，ユーザがある地域の観光をしているときに周辺の地域や建物に関する意外な情報を提示し，ユーザの興味を喚起するといったことが可能になる．

本研究では，“主題語”と“関連語”という観点から情報を捉える．例えば，“落合博満”を主題語としたとき，関連語には“首位打者”，“秋田”，“ガンダム”など様々な語があり，“落合博満”と“首位打者”からなる情報のひとつとして“落合博満は首位打者を獲得したことがある．”が挙げられる．主題語および関連語については３章で詳しく述べる．本稿では，大きく以下の３つの段階に分けて意外な情報を発見する．

(1). 入力語（主題語）$q$を与え，その関連語集合$L_q = \{e_1, e_2, \ldots, e_n\}$を収集する．
(2). $q$および$e_i$の同位語の関係と$e_i$の認知度に基づいて$q$に対する$e_i$の意外度を求める．
(3). （２）で求められた意外度の高い関連語を含む情報を求める．

（１）では，主題語の関連語を収集する情報源としてWikipedia[1]の記事を用いる．（２）では，Wikipedia の記事間のリンク構造と語の上位下位関係を用いて主題語と関連語の関連度を測り，“落合博満”という主題語にとって“野球”よりも“ガンダム”という語の方が意外度が高いということを求める．（３）では，意外度が高いと判定された関連語を含む文を Wikipedia の記事から抽出する

我々は人物名，地名，製品名，施設名，および組織名の5つのカテゴリに含まれる75の語を入力語として実験を行った．実験により，主題語と関連語の同位語を考慮すること，および関連語の認知度を考慮することは意外な情報を発見するために有効であるということが明らかになった．

## 2. 関連研究

Nodaら[1]はWikipediaのカテゴリ間の関係を分析することで意外性のある知識を発見する手法を提案した．彼らの手法を用いることで，“麻生太郎”は“日本の内閣総理大臣”というカテゴリに属している一方で“オリンピック射撃競技日本代表選手”というカテゴリにも属しているといった情報を発見することができる．しかし，彼らの手法では，Wikipediaのカテゴリとして作成されていない情報は抽出できないという問題点がある．Nadamotoら[2]は，ブログやSNSといったコミュニティ型コンテンツを対象として，コミュニティ内の議論においてユーザが気づいていない情報を発見する手法を提案している．彼らの手法では，特定の構文パターンを用いているため，発見できる情報のタイプに制約があるが，我々は構文パターンを用いないため，意外な情報をより柔軟に発見することができる．Liuら[3]の研究では，2つのウェブページを与えたときに，一方のウェブページにしか含まれていない意外な情報を，各ウェブページに含まれる語を比較することで発見する手法を提案している．彼らが

♥ 学生会員　京都大学大学院情報学研究科社会情報学専攻 博士後期課程　tsukuda@dl.kuis.kyoto-u.ac.jp
♦ 正会員　京都大学大学院情報学研究科社会情報学専攻 {ohshima, tanaka}@dblab.is.ocha.ac.jp
♣ 非会員　株式会社デンソーアイティーラボラトリ {miyamamoto, hiwasaki}@dl-itlab.co.jp

[1] http://ja.wikipedia.org/


**図 1　主題語と関連語およびそれらの同位語間の関係に基づく情報の構造**

**Fig. 1　The structure of information based on a theme term, its related term and their coordinate terms**

2つのウェブページを与えて意外な情報を発見するような状況を想定しているのに対して，我々はあるキーワードに関する意外な情報を発見するような状況を想定している．

## 3. 意外な情報

本章では，我々が対象とする意外な情報について述べる．

まず，本研究では情報を“主題語” と“関連語” という観点からとらえる．主題語とは意外な情報を求める対象となる人物名や地名などの語である．関連語とは，主題語に対して決まるものであり，主題語と何らかの観点において関連のある語である．例えば，“落合博満” という主題語の関連語としては，“元プロ野球選手” や“中日ドラゴンズ”，“秋田県”，“ガンダム” など，様々な語があげられる．

次に，同位語について述べる．同位語とは，共通の上位語をもつような語のことである．例えば，“落合博満”と“王貞治”は，“元プロ野球選手”という共通の上位語をもつため，同位語である．さらに，“落合博満”と“麻生太郎”も，“男性”という共通の上位語を持つため，同位語である．ただし，“落合博満”の同位語としては，“麻生太郎”よりも“王貞治”の方が，同位語としてよりふさわしいと考えられる．この理由として，“落合博満”と“王貞治”は，“元プロ野球選手”の他にも，“男性”や“三冠王を獲得した選手”のように，多くの共通の上位語を持っているという点があげられる．つまり，ある語の同位語の中には，より同位語らしい語と同位語らしくない語が存在する．

以上をもとに，ある情報が与えられたときに，それに含まれる主題語と関連語，さらにそれぞれの同位語がどのような関係のときに人はその情報を意外であると感じるかを“落合博満”が主題語である４つの例を用いて説明する．まず，“落合博満は首位打者を獲得したことがある．”という情報は，多くの人にとって意外ではないと考えられる．この理由は，野球選手が首位打者を獲得することは一般的によく知られているためである．つまり，“落合博満”の同位語らしい語も，関連語として“首位打者”という語をもちうるためである（図1(a)）．次に，“落合博満は秋田県出身である．”という情報と“落合博満はガンダムマニアである．”という情報について考える．この場合，“落合博満”の同位語らしい語の関連語には，“秋田県”や“ガンダム”という語は全く含まれないか，ごく一部の同位語の関連語にのみ含まれる．しかし，この2つの情報があったとき，前者は広くは知られていないが意外性は低く，後者は広くは知られておらずかつ意外性が高いと考えられる．なぜなら，前者の場合，どの野球選手もいずれかの都道府県の出身であり，“落合博満は秋田県出身である．”という情報はその一例でしかないためである．つまり，“落合博満”のより同位語らしい語は，“秋田県”のより同位語らしい語，つまり都道府県名を関連語としてもっているためである（図1(b)）．一方後者の情報の場合，野球選手は一般にアニメ好きであるというイメージがないため，“落合博満はガンダムマニアである．”という情報の意外性は高い．つまり，“落合博満”の同位語らしい語は，関連語として“ガンダム”の同位語らしい語を持っていないためである（図1(c)）．ここで，“落合博満は成田山名古屋別院大聖寺で中日ドラゴンズの優勝祈願をした．”という情報を考えると，“落合博満”の同位語らしい語は“成田山名古屋別院大聖寺”の同位語らしい語を関連語として持たない．しかし，“成田山名古屋別院大聖寺”が一般に広くは知られていない認知度の低い語であるため，そのような情報を聞いても人は意外とは感じないということが考えられる．つまり，主題語に対する各関連語の意外度を測るためには，各関連語の認知度も考慮する必要がある．

以上より，本研究では，“主題語の多くの同位語らしい語が，認知度の高い関連語およびその同位語らしい語と関連を持たないような主題語と関連語を含む情報”は意外であるという仮説を立てる．そして，ある主題語$q$とある関連語$e$が与えられたときに，$q$と$e$の関連の低さを求める関数$\mathrm{Rel}(q,e)$と$e$の認知度の高さを求める関数$\mathrm{Cog}(e)$を定義し，最終的にそれらを組み合わせた関数$Unexp(q,e) = f\big(Rel(q,e), Cog(e)\big)$を定義することで$q$に対する$e$の意外度を測る．

## 4. 主題語と関連語の組合せの意外度の計算

主題語に対する各関連語の意外度を計算する際の流れは以下のようになる．

1. 主題語$q$を与え，$q$に対する関連語集合$L_q = \{e_1, e_2, \ldots, e_n\}$を収集する．
2. 主題語$q$の上位語集合と同位語集合，各関連語の上位語集合と同位語集合を収集する．
3. 主題語$q$に対する各関連語の関連度$\mathrm{Rel}(q, e_i)$を求める．
4. 各関連語の認知度$\mathrm{Cog}(e_i)$を求める．
5. 主題語に対する各関連語の意外度$\mathrm{Unexp}(q, e_i)$を求める．

次節以降では，それぞれの手法について説明する．

### 4.1 関連語の収集

本研究で我々が関連語を収集する情報源として選択したのは主題語$q$を見出し語とする Wikipedia の記事である．我々が Wikipedia の記事を選択した理由は大きく２つある．１つ目は，Wikipedia の記事には見出し語に関連のある情報のみが記述されているため，ノイズとなる語が比較的含まれにくいという利点がある．２つ目は， Wikipedia の記事中には客観的な内容が記述されているためである．我々が対象とする意外な情報とは，誰かの意見や感想ではなく，客観的な視点から記述されている文であり，この点からも

Wikipediaの記事は関連語の情報源として適しているといえる．これらに加えて，Wikipediaのポリシーとして，見出し語と関連のない語についてはリンクを作成しないことになっている．以上のような理由から，本研究では主題語が見出し語となっている Wikipedia の記事中でリンクが貼られている全ての語をその主題語の関連語集合として収集する．

### 4.2 同位語および上位語の取得

主題語の同位語を得るために，本稿では ALAGIN フォーラムから提供されている上位語階層データ[2]を用いる．このデータは，Wikipedia で記事の見出し語やカテゴリ名となっている名詞句をその上位語，下位語関係に基づいて階層化したものである．これを用いることで，ある語の上位語集合や，ある語と共通の上位語をもつ同位語集合を求めることが可能となる．

### 4.3 主題語と関連語の関連度

本節では，提案手法の詳細を説明する前に，図 6 を用いて提案手法のアイデアを直感的に説明する．図 6 のグラフは下記のノード集合から構成される．以下では，主題語を$q$，語$t$の上位語集合を$hyper(t)$，語$t$の下位語集合を$hypo(t)$，語$t$の関連語集合を$rel(t)$とする．

- $Q = \{q\}$.
- $H_q = \{x | x \in hyper(q)\}$.
- $C_q = \{x | x \in hypo(y), y \in H_q, x \notin Q\}$.
- $L_q = \{x | x \in rel(q)\}$.
- $H_{lq} = \{x | x \in hyper(y), y \in L_q\}$.
- $L_c = \{x | x \in rel(y), y \in C_q, x \notin L_q\}$.

図 6 では，黒丸，白丸，黒三角，白三角のノードはそれぞれ$Q$，$C_q$，$L_q$，$L_c$の語に対応している．そして四角のノードは$H_q$または$H_{lq}$の語に対応している．

また，エッジについては，上位語・下位語の関係にある２語間および，ある語とその関連語の２語間に存在する．以下で，$(n_1, n_2)$はノード$n_1$とノード$n_2$の間にエッジが存在することを表す．

- $(q, x)$ where $x \in H_q$.
- $(x, y)$ where $x \in H_q, y \in C_q,$ and $y = hypo(x)$.
- $(x, y)$ where $x \in C_q, y \in L_c,$ and $y = rel(x)$.
- $(x, y)$ where $x \in C_q, y \in L_q,$ and $y = rel(x)$.
- $(x, y)$ where $x \in L_c, y \in H_{lq},$ and $y = hyper(x)$.
- $(x, y)$ where $x \in H_{lq}, y \in L_q,$ and $x = hyper(y)$.

このグラフにおいて，$\boldsymbol{q}$と$\boldsymbol{x \in L_q}$の間にエッジは存在しない．これは，主題語と主題語の関連語の関係の強さを測るために，主題語からその同位語を経由した際の，主題語から各関連語への辿り着きやすさを求めるためである．つまり，主題語の同位語から辿り着くのが“容易”な関連語は，主題語にとって意外な語ではないと考える．既に述べたように，“首位打者”という語は“落合博満”という語にとって意外ではない．これは，“落合博満→野球監督→野村克也→首位打者”や“落合博満→スポーツ選手→イチロー→首位打者”のように，“落合博満”からその同位語らしい語を通って直接“首位打者”に到達できるパスが沢山あるためであり，このようなとき，容易に辿り着けると考える．一方で，“落合博満”の同位語らしい語から１ステップで“秋田県”に到達できるパスはそれほど多くない．しかしこの場合でも，“落合博満→野球監督→野村克也→京都府→都道府県→秋田県”のように，“落合博満”の同位語らしい語から，“秋田県”の上位語を経由して“秋田県”に到達するパスは多数存在する．しかし，“ガンダム”の場合，“ガンダム”の上位語を考慮したとしても，“落合博満”の同位語らしい語から“ガンダム”に到達するパスはほとんど無いと考えられる．もちろん，“落合博満”の同位語らしくない語を経由すれば，“落合博満→男性→麻生太郎→ローゼンメイデン→アニメ→ガンダム”のようなパスは存在する．このような場合，“落合博満”から“ガンダム”に辿り着くのは“困難”であると考える．

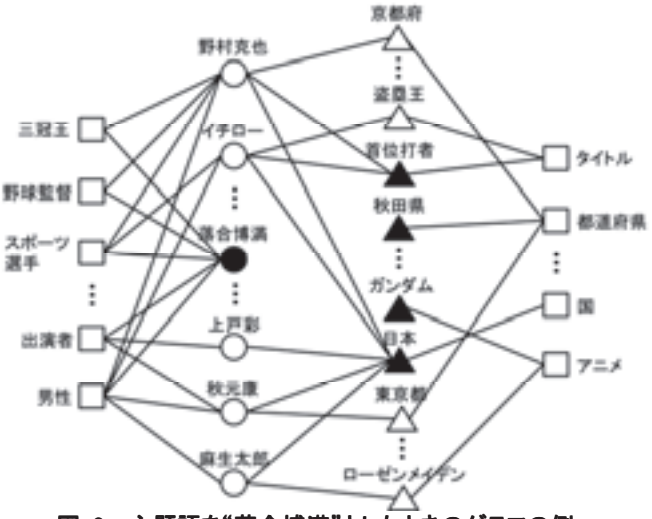


**図 2　主題語を“落合博満”としたときのグラフの例**
**Fig. 2　An example of the graph for a theme term "Hiromitsu Ochiai"**

以上のアイデアをもとに，以降では提案手法の詳細を述べる．まず，主題語と各関連語の関係の強さを評価するために，上記のグラフを構築する．そして，主題語と関連語の関連の有無を，その２語間のパスの有無とみなす．さらに，主題語から関連語へのグラフ上での辿り着きやすさを主題語と関連語の関連の強さとみなす．つまり，主題語から辿り着きにくい関連語ほど，主題語と関連の低い語となる．

次節以降では，グラフを３つのグラフに分割し，各グラフを段階的に評価することで２語の関連度を測る手法について説明する．

#### 4.3.1 主題語に対する同位語らしさの計算

まず，主題語$q$とその上位語および同位語から構成される２部グラフ$G_1 = (Q \cup C_q \cup H_q, E_1)$について考える．ここで，$E_1$は$H_q$と$Q \cup C_q$の間の枝集合である．語$h_i \in H_q$と語$t_j \in Q \cup C_q$が上位下位関係にあるとき，$h_i$と$t_j$の間に枝が存在する．

本研究ではこの２部グラフに次式で表される Co-HITS アルゴリズム[4]を適用し，$C_q$の各語について，$q$との同位語らしさを求める．

$$x_i = (1 - \lambda_h) x_i^0 + \lambda_h \sum_{t_j \in Q \cup C_q} w_{ji}^{th} y_j \quad (1)$$

$$y_j = (1 - \lambda_t) y_j^0 + \lambda_t \sum_{h_i \in H_q} w_{ij}^{ht} x_i \quad (2)$$

ここで，$x_i^0$は語$h_i$の初期値，$y_j^0$は語$t_j$の初期値である．Co-HITS アルゴリズムでは，より多くの枝をもつノードほど，枝の重みは小さくなる．具体的には，$h_i$から$t_j$への枝の

[2] http://nlpwww.nict.go.jp/corpus/

重みは$w_{ij}^{ht} = 1/|hypo(h_i)|$，$t_j$から$h_i$への枝の重みは$w_{ji}^{th} = 1/|hyper(t_j)|$と表される．さらに，Co-HITS アルゴリズムはノードの初期値も考慮する．本手法では$q$の初期値を 1，他のノードの初期値を 0 とする．$\lambda_h \in [0,1]$および$\lambda_t \in [0,1]$は$x_i^0$，$y_j^0$の重要度を表すパラメータである．$G_1$では$q$のみが初期値 1 をもつので，$\lambda_h$および$\lambda_t$の値によって Co-HITS の結果に大きな差がなかったため，本研究では$\lambda_h = \lambda_t = 1$とした．

### 4.3.2 主題語に対する関連語の関連度の計算1

主題語と各関連語の関連度を求めるために，$C_q$，$L_q$，および$L_c$から構成されるグラフ$G_2$を作成する．このグラフは有向グラフであり，語$y \in C_q \cup L_q \cup L_c$が語$x \in C_q \cup L_q \cup L_c$の関連語であるとき，$x$から$y$に向かって枝が存在する．このグラフに対して，我々は biased PageRank アルゴリズム[5]を適用する．PageRank アルゴリズム[6]では，ページ$u$からページ$v$にリンクが存在する場合，$u$の重要度が$v$に伝播する．$r(u)$を$u$の重要度，$F_u$を$u$がリンクしているウェブページ集合とする．リンクの重要度は等しいと考え，$u$から$v \in F_u$へは$r(u)/|F_u|$の重要度が伝播する．$r(u)$の値もまた$u$をリンクするウェブページによって再帰的に決まる．$B_v$を$v$にリンクを張っているページ集合，$N$をグラフ中のノード数，αをダンピングファクターとすると，$r(v)$は次式により求められる．

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + \frac{1-\alpha}{N} \tag{3}$$

本研究ではα = 0.85とした．ここで，各関連語への$q$からの辿り着きやすさを測るため，biased PageRank の考えを用いて，式（4）の$(1-\alpha)/N$を次式のように変更する．

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + (1-\alpha)\frac{f_{ini}(v)}{\sum_{t \in C_q} f_{ini}(t)} \tag{4}$$

$f_{ini}(v)$はノード$v$の初期値であり，次式により求める．

$$f_{ini}(v) = \begin{cases} \dfrac{f_{co}(v)}{\sum_{t \in C_q} f_{co}(t)} & \text{if } v \in C_q \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$f_{co}(v)$は 4.3.1 項で求められた，$q$に対する$v$の同位語らしさである．式（5）をグラフ$G_2$に適用し，ノードの値が収束したときに$L_q$の中で値が低いノードは$q$と関連度の低い語であるといえる．この段階では，$q$の同位語らしい語と直接関連のある語のみ考慮しているため，次節では関連語の上位語を通して関連のある語も考慮する．

### 4.3.3 主題語に対する関連語の関連度の計算2

最後に，$t \in L_q$の上位語を考慮することで$q$と各関連語の関連度を求める．語$t \in L_q$に対して，我々はまず全ての同位語を収集する．$t$およびその同位語の集合を$C_t$，$t$の上位語集合を$H_t$とし，２部グラフを構築する．$C_t$には$G_2$に含まれる語も存在する．語$u_i \in C_t$と語$v_j \in H_t$が上位下位関係にあるとき，２語の間に枝が存在する．この２部グラフに対して Co-HITS アルゴリズムを適用する．$H_t$に含まれるノードの初期値は全て0である．$C_t$の語が$G_2$に含まれる場合，その語は 4.3.2 項で求められた値を初期値としてもつ．$C_t$の語が$G_2$に含まれない場合，その語の初期値は0である．$u_i$のオーソリティ度を$x_i$，$v_j$のハブ度を$y_j$としたとき，各ノードの値は次式により求められる．

$$x_i = (1-\lambda_u)x_i^0 + \lambda_u \sum_{t_j \in H_t} w_{ji}^{vu} y_j \tag{6}$$

$$y_j = (1-\lambda_v)y_j^0 + \lambda_v \sum_{u_i \in C_t} w_{ij}^{uv} x_i \tag{7}$$

ここで$x_i^0$および$y_j^0$はそれぞれ$u_i$と$v_j$の初期値であり，$w_{ji}^{vu} = 1/|hyper(u_i)|$，$w_{ij}^{uv} = 1/|hypo(v_j)|$である．$H_t$に含まれる語の初期値はいずれも0であるため，$\lambda_v$の値は1とした．$\lambda_u$の値の違いによる影響については５章で検証する．

上記の計算を主題語$q$の各関連語$e_i$について行い，式（7）により求められる$e_i$の値を$Rel(q, e_i)$とする．

### 4.4 関連語の認知度の計算

関連語の認知度を測る方法として，本研究では次の２つを用いる．１つ目は，Wikipedia の全記事集合に対してリンク関係をエッジとするグラフを構築して PageRank アルゴリズム[6]を適用し，PageRank の値を認知度とする方法である．PageRank アルゴリズムでは，多くの記事から参照されている記事は高い PageRank の値をもつため，我々はそのような記事の見出し語の認知度は高いと仮定する．語$e_i$が見出し語である記事の PageRank の値を$PageRank(e_i)$とする．２つ目は，関連語をクエリとしてウェブ検索をした際の検索結果数をその語の認知度とみなす方法である．本研究では Yahoo! ウェブ検索 API[3]を用いて各関連語に対するウェブ検索の検索結果数を取得する．語$e_i$のウェブ検索の検索結果数を$Hit(e_i)$とする．

### 4.5 関連語の意外度の計算

これまで述べてきた，主題語と各関連語の関連度および各関連語の認知度を用いて，主題語に対する各関連語の意外度を求める．4.3.3 項で求められた関連語の関連度は，値が低いほど主題語との関連度が低く，意外度が高いので，意外度を計算する際にはその逆数をとる．一方，関連語の認知度に関しては，認知度が高いほど意外度は高くなる．これらの考えに基づいて，PageRank アルゴリズムを用いて関連語の認知度を求める場合の主題語$q$に対する関連語$e_i$の意外度$Unexp(q, e_i)$を次式により求める．

$$Unexp(q, e_i) = \frac{1}{Rel(q, e_i)} \cdot PageRank(e_i) \tag{8}$$

また，検索結果数を用いて関連語の認知度を求める場合の主題語$q$に対する関連語$e_i$の意外度$Unexp(q, e_i)$を次式により求める．

$$Unexp(q, e_i) = \frac{1}{Rel(q, e_i)} \cdot \log_{10} Hit(e_i) \tag{9}$$

## 5. 実験

提案手法の有効性を検証するために実験を行った．実験では，次の２つの点を明らかにすることを目的とする．（ⅰ）意外な情報を発見する際の，関連語の認知度を考慮することの重要性．（ⅱ）意外な情報を発見する際の，主題語及び関連語の同位語間の関係を考慮することの重要性．

これらを明らかにするために，我々は６つの提案手法と４つの比較手法を用いた．６つの提案手法では，各関連語の意外度を式（９）または（10）により求める．ここで，式（７）中の$\lambda_u$の値による影響を調べるため，$\lambda_u$の値を0.25, 0.5, 0.75

[3]http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html

とした場合の結果を比較した．これ以降，式（9）を用い，$\lambda_u$の値を0.25，0.5，0.75としたときの手法をそれぞれ$PR_{25}$，$PR_{50}$，$PR_{75}$と記述する．同様に，式（10）を用い，$\lambda_u$の値を0.25, 0.5, 0.75としたときの手法をそれぞれ$HIT_{25}$，$HIT_{50}$，$HIT_{75}$とする．

１つ目の疑問を検証するために，主題語と関連語の関連の強さのみを考慮する手法を用いた．主題語$q$に対する関連語$e_i$の意外度は$Unexp(q,e_i) = 1/Rel(q,e_i)$により計算される．この手法においても，$\lambda_u$の値を0.25，0.5，0.75とし，それぞれの結果を比較した．$\lambda_u$の各値に対する手法を$REL_{25}$, $REL_{50}$, $REL_{75}$とする．次に，２つ目の疑問を検証するために，“落合 博満 ガンダム”のように，主題語と各関連語から検索用クエリを作成し，検索結果数が少ない順に関連語を順位付けする手法を用いた．以後，この手法をTCと記述する．

本研究では，主題語が見出し語である Wikipedia の記事から意外な情報を抽出する．主題語と関連語が与えられると，記事中でその関連語を含む１文を求め，意外な情報とする．

### 5.1 主題語

本実験では，人物名，地域名，製品名，施設名，および組織名の５つのカテゴリそれぞれに対して 15 個，合計 75 個の主題語を用いた．ある情報の意外度を人が評価する際，主題語自体に対する認知度が全く無い場合，その語に関するどのような情報もユーザにとっては意外ではないと考えられる．そのため，我々はまず Wikipedia の記事間の参照関係をもとに全ての記事の PageRank の値を計算し，その値の上位５％にあたる 17,325 個の主題語を抽出した．また，主題語に対する関連語の数が少ないほど，意外な情報を発見できる可能性は低いと考え，抽出した主題語を２つの集合に分けた．一方の集合 A は，150 個以上の関連語をもつ主題語から成り，もう一方の集合 B は関連語の数が 150 個未満の主題語から成る．集合 A には 4,854 個，集合 B には 12,471 個の主題語が含まれる．最後に，実験に用いる主題語として，５つの各カテゴリに属する主題語を集合 A からランダムに 10 個ずつ，集合 B からランダムに５個ずつ選んだ．

### 5.2 実験手順

本実験では，５名の被験者を用いて評価を行った．５名の被験者は，30 代の男性２名，20 代の女性２名，30 代の女性１名であった．評価を行うにあたり，以下に述べる手順で主題語ごとにアンケートを作成した．まず，６つの提案手法と４つの比較手法に対して１つの主題語を与えると，各手法は関連語を意外度が高い順にランキングして返す．我々は各手法の上位５個の関連語を選択して１つの集合とし，関連語とその関連語に対する意外な情報の組をランダムに並べて被験者に提示する．その後，意外度を１から４の４段階でスコア付けしてもらった．スコアが高いほど，その情報の意外度は高いということを表す．１つの主題語に対して１枚のアンケートを作成し，合計で 75 枚のアンケートを作成した．そして，順序効果を考慮して５つの評価用セットを用意した．

被験者の評価後，各情報に対する５名の被験者の平均値を求め，その値をその情報の意外度とした．

### 5.3 結果

５つの各カテゴリにおける各手法の nDCG[7]の値を表１に示す．この結果を見ると，いずれのカテゴリにおいても６つの提案手法のいずれかが最も高い nDCG の値をとっていることがわかる．主題語および関連語の同位語のみを考慮し，関連語の認知度を考慮していない$REL_{25}$, $REL_{50}$, および$REL_{75}$の nDCG@5 の全カテゴリの平均値はいずれの提案手法の値よりも低かった．この結果から，関連語の認知度を考慮することは意外な情報を発見する際に有効であるといえる．主題語および関連語の同位語を考慮していないTCは全てのカテゴリで最も低い値となった．この結果は，主題語および関連語の同位語を考慮することの有効性を示している．また，本実験の結果からは，式（7）における$\lambda_u$の値や，関連語の認知度の求め方による大きな違いは見られなかった．

**表 1　5 つのカテゴリに対する各手法の nDCG@5 の値**
**Table 1　Performance comparison of each category for ten methods measured by nDCG**

| 手法 | 人物 | 地域 | 製品 | 施設 | 組織 | 平均 |
|---|---|---|---|---|---|---|
| TC | 0.705 | 0.757 | 0.773 | 0.787 | 0.780 | 0.760 |
| $REL_{25}$ | 0.805 | 0.792 | 0.837 | 0.800 | 0.853 | 0.817 |
| $REL_{50}$ | 0.807 | 0.803 | 0.839 | 0.800 | 0.857 | 0.821 |
| $REL_{75}$ | 0.807 | 0.808 | 0.841 | 0.804 | 0.852 | 0.822 |
| $PR_{25}$ | **0.828** | 0.830 | 0.846 | **0.825** | 0.860 | **0.838** |
| $PR_{50}$ | 0.824 | 0.830 | 0.851 | 0.821 | 0.860 | 0.837 |
| $PR_{75}$ | 0.818 | 0.836 | **0.858** | 0.820 | 0.854 | 0.837 |
| $HIT_{25}$ | 0.798 | 0.832 | 0.843 | 0.824 | 0.856 | 0.830 |
| $HIT_{50}$ | 0.791 | 0.834 | 0.843 | 0.823 | 0.867 | 0.832 |
| $HIT_{75}$ | 0.790 | **0.838** | 0.847 | 0.822 | **0.872** | 0.834 |

提案手法によって発見された情報のうち，被験者によって意外であると判定された情報の例を表２に示す．提案手法では“秋田県”という主題語に対して“生活習慣病”という語が意外な語として発見された．“秋田県”の同位語としては他の都道府県名が同位語らしい語として評価され，“生活習慣病”の同位語としては病名が同位語らしい語として評価されていた．一般的に，都道府県は特定の病気と関連をもっておらず，かつ“生活習慣病”という語の認知度は高いといえる．そのため，提案手法によって“生活習慣病”という語の意外度を高く評価できていた．

最後に，各カテゴリにおいて少なくとも１つ意外な情報を発見できた主題語の数およびその割合を表３に示す．関連語を 150 個以上もつ主題語については 40%の割合で，関連語が 150 個未満の主題語については 24%の割合で意外な情報を発見できていた．意外な情報を発見できなかった主題語についてその原因を分析してみると，主に２つの原因があることがわかった．１つ目は，関連語が多数ある場合でも，Wikipedia の記事の中に意外であると感じる情報がない場合である．特に施設名と組織名のカテゴリではこの傾向が強く見られた．２つ目は，主題語の同位語と関連語の関係の強さに着目するという提案手法の特性にある．例えば，“デジタルカメラ”という主題語の記事の中には，“通常「デジカメ」と略称されるが，「デジカメ」は，日本国内では三洋電機や，他業種各社の登録商標である．”という情報があり，これは意外な情報の候補の１つであるといえる．この情報の中の関連語は“三洋電機”であるが，この関連語は“デジタルカメラ”の同位語である他の多くの電化製品と関連があるため，提案手法では意外な関連語として発見することはできなかった．

## 6. まとめと今後の課題

本稿では，クエリとして与えられた主題語に関する意外な情報を発見することを目的とし，主題語に対する関連語の意

表 2　提案手法により発見された意外な情報の例
Table 2　Examples of discovered unexpected information

| 主題語 | 関連語 | 意外な情報 |
|---|---|---|
| エアバッグ | 消防法 | エアバッグが，火薬の使用が当時の日本の消防法に抵触してしまうことから，日本でエアバッグが開発されることはなかった． |
| 法隆寺 | 文化財防火デー | この火災がきっかけで文化財保護法が制定され、火災のあった 1 月 26 日が文化財防火デーになっている． |
| モナコ | 伊達公子 | クルム伊達公子：現在モナコ在住． |
| 三井グループ | 東京ディズニーランド | 東京ディズニーランド・東京ディズニーシー内には三井住友銀行（旧三井銀行→さくら銀行、以下同）の出張所がある． |
| 秋田県 | 脳卒中 | 特に日本酒の消費量が多く酒の飲み過ぎに加えて，雪国のため保存食である漬け物などの塩分過多が加わり，脳卒中などの生活習慣病での死亡率も高くなっている． |
| 駅弁 | 土瓶 | 1922 年，鉄道省は衛生上の理由により土瓶を禁止したためガラス製のものが登場した． |
| 野比のび太 | 一次方程式 | しかし，「2/3 ÷ 0.25 ÷ 0.8 = 10/3」という難解な答えを正解に導いたり，本来中学校で習うはずの一次方程式「3/8 x = 9/10」を解いたこともあり，100 点を取っている． |

表 3　各カテゴリで意外な情報が発見された主題語の数
Table 3　The number and the ratio of theme terms that could find unexpected information

| カテゴリ | 関連語 150 個以上 | 関連語 150 個未満 | 合計 |
|---|---|---|---|
| 人物 | 6/10 | 3/5 | 9/15 |
| 地域 | 3/10 | 0/5 | 3/15 |
| 製品 | 5/10 | 2/5 | 7/15 |
| 施設 | 4/10 | 0/5 | 4/15 |
| 組織 | 2/10 | 1/5 | 3/15 |

外度を計算する手法の提案を行った．我々は，主題語の多くの同位語らしい語が，認知度の高い関連語およびその同位語らしい語と関連を持たないような主題語と関連語を含む情報は意外な情報であるとし，グラフにおいて主題語から各関連語への辿り着きづらさと各関連語の認知度をもとに主題語に対する関連語の意外度を求めた．また，実験により，主題語と関連語それぞれの同位語を考慮すること，関連語の認知度を考慮することは意外な情報を発見するうえで有効であることを明らかにした．

本稿では Wikipedia を対象として関連語および意外な情報を抽出したが，今後は別の情報源も対象とする予定である．

## [謝辞]

本研究の一部は，文部科学省科学研究費補助金（課題番号 24240013，24680008，243993）によるものです．ここに記して謝意を表します．

## [文献]

[1] Y. Noda, Y. Kiyota and H. Nakagawa: “Discovering serendipitous information from wikipedia by using its network structure”, Proc. of ICWSM 2010, pp. 299-302 (2010).
[2] A. Nadamoto, E. Aramaki, T. Abekawa and Y. Murakami: “Content hole search in community-type content”, Proc. of WWW 2009, pp. 1223-1224 (2009).
[3] B. Liu, Y. Ma and P. S. Yu: “Discovering unexpected information from your competitors’ web sites”, Proc. of ACM SIGKDD 2001, pp. 144-153 (2001).
[4] H. Deng, M. R. Lyu and I. King: “A generalized Co-HITS algorithm and its application to bipartite graphs”, Proc. of ACM SIGKDD 2009, pp. 239-248 (2009).
[5] T. H. Haveliwala: “Topic-sensitive pagerank”, Proc. of WWW 2002, pp. 517-526 (2002).
[6] S. Brin and L. Page: “The anatomy of a large-scale hypertextual web search engine”, Proc. of WWW 2007, pp. 107-117 (1998).
[7] K. Jarvelin and J. Kekalainen: “Cumulated gain-based evaluation of ir techniques”, ACM Trans. Inf. Syst., 20, 4, pp. 422–446 (2002).

**佃 洸摂　Kosetsu TSUKUDA**
京都大学大学院情報学研究科博士後期課程在学中．2011 年京都大学大学院情報学研究科博士前期課程修了．情報処理学会学生会員．日本データベース学会学生会員．

**大島 裕明　Hiroaki OHSHIMA**
京都大学大学院情報学研究科社会情報学専攻助教．2007 年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）．主にウェブ，情報検索，データベースの研究に従事．情報処理学会，電子情報通信学会，日本データベース学会，ACM 各会員．

**山本 光穂　Mitsuo YAMAMOTO**
（株）デンソーアイティーラボラトリ研究企画部シニアエンジニア．2003 年 長岡技術科学大学電気電子システム工学修了．主に車載機器向けサービスの開発及び情報検索の研究に従事．

**岩崎 弘利　Hirotoshi IWASAKI**
（株）デンソーアイティーラボラトリ研究開発部ジェネラルマネージャ．1990 年名古屋大学工学研究科博士課程前期課程電機・電子工学専攻修了．博士（工学）．1990 年日本電装株式会社（現在の（株）デンソー）入社．2000 年より（株）デンソーアイティーラボラトリ出向．車の知能的ユーザインターフェースの研究開発に従事．人工知能学会，電子情報通信学会，情報処理学会各会員．

**田中 克己　Katsumi TANAKA**
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．