情報処理学会論文誌　データベース　Vol.7　No.1　1–17 (Mar. 2014)

語の認知度と語間の関係の非典型度に基づく Wikipedia からの意外な情報の発見

佃 洸摂[1,†1,a)]　大島 裕明[1,b)]　山本 光穂[2,c)]　岩崎 弘利[2,d)]　田中 克己[1,e)]

受付日 2013年9月20日, 採録日 2014年1月7日

概要：本稿ではユーザが与えた 1 語のクエリに対して，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，クエリに対して意外度の高い関連語を発見し，クエリと意外度の高い関連語を基に意外な情報を発見する．その際，クエリの関連語の中でもクエリとの関係が非典型的であり，かつ認知度が高い関連語ほど意外度が高いという仮説に基づいて関連語の意外度を求める．たとえば提案手法により，“落合博満” というクエリに対して “ガンダム” という関連語の意外度が高いことが分かり，これを基に “落合博満はガンダムマニアである.” という意外な情報を発見できる．提案手法では，Wikipedia から得られる情報のみを用いてクエリに対する意外な情報の発見を行う．実験では 75 語のクエリを用いて評価を行い，クエリと関連語間の関係の非典型度，および関連語の認知度を考慮することの有効性を示した．

キーワード：意外な情報検索，典型度，Wikipedia

Discovering Unexpected Information Based on Popularity of Terms and Atypicality of Relationships between Terms

Kosetsu Tsukuda[1,†1,a)]　Hiroaki Ohshima[1,b)]　Mitsuo Yamamoto[2,c)]
Hirotoshi Iwasaki[2,d)]　Katsumi Tanaka[1,e)]

Received: September 20, 2013, Accepted: January 7, 2014

Abstract: In this paper, we propose methods for discovering unexpected information for a given query. Our method first finds unexpected related terms for the query and then finds unexpected information based on the query and unexpected related terms. We hypothesize that a related term that has both an atypical relationship with a query and a high popularity is unexpected to the query. For example, given the query “Hiromitsu Ochiai,” our method detects that “Gundam” is an unexpected term and retrieves unexpected information: “Hiromitsu Ochiai is a Gundam mania.” Our method finds unexpected information solely based on information in Wikipedia. Experimental results using 75 queries show that these two proposed factors are effective for discovering unexpected information.

Keywords: unexpected information retrieval, typicality, Wikipedia

1. はじめに

Google [*1]や Yahoo [*2]，Bing [*3]などの既存の Web 検索エンジンでは，ユーザが入力したクエリとの適合度や人気度によって Web ページをランキングし，検索結果を返す．クエリに適した文書の検索を可能にするために，これまで

[1] 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] 株式会社デンソーアイティーラボラトリ
Denso IT Laboratory, Shibuya, Tokyo 150–0002, Japan
[†1] 現在，日本学術振興会特別研究員（DC1）
Presently with JSPS Research Fellow (DC1)
[a)] tsukuda@dl.kuis.kyoto-u.ac.jp
[b)] ohshima@dl.kuis.kyoto-u.ac.jp
[c)] miyamamoto@d-itlab.co.jp
[d)] hiwasaki@d-itlab.co.jp
[e)] tanaka@dl.kuis.kyoto-u.ac.jp

[*1] http://www.google.com
[*2] http://www.yahoo.com
[*3] http://www.bing.com

様々な検索手法が提案されてきた．たとえば，BM25 [24] はテキスト情報に基づいた手法であり，HITS [14] や PageRank [5] は Web ページ間のリンク構造に基づいた手法である．また，これらの手法を拡張することで，クエリにより適合した文書を検索することが可能になってきた [8], [11], [12], [16], [25], [27].

しかし，これらの手法を用いてクエリに関する意外な情報を発見するのは困難である．相関ルールの抽出に関する研究では，意外なルールを発見することを目的とした研究は多数行われてきたが [4], [22], [23], [29]，Web から意外な情報を発見する研究はほとんど行われていない [17], [19], [20], [21]．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 Web 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．たとえば，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 “落合博満” と “ガンダム” については知っているが，それらの間に関係があることを知らないユーザにとっては意外な情報になりうる．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，クエリに関する一般的でない情報は検索結果の下位のページに記述されているか，あるいは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．

意外な情報には様々なものが存在する．情報を提示するユーザの興味を引くか否かという観点と，情報が有用であるか否かという観点を考えたときに，ユーザの興味を引く有用な情報を発見し提示することは以下のような状況で有益になりうる．たとえば，ユーザがある人物名をクエリとして Web 検索を行っているときにその人物に関する意外な情報を提示することで，その人物に対するユーザの興味をより喚起することができる．同様に，ニュース記事を閲覧しているユーザに対して，その記事で取り上げられている人物や出来事に関する意外な情報を提示することでその記事への興味を喚起できる．また，ユーザがある地域の観光やドライブ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することもユーザの興味の喚起につながる．そこで，本研究では人物名や施設名，地名などを入力として，入力された語に関する意外な情報を発見することを目的とする．

本研究では，“主題語” と “関連語” という観点から情報をとらえる．たとえば，“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 1 つとして “落合博満は首位打者を獲得したことがある.” があげられる．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 $L_q = \{e_1, e_2, \cdots, e_n\}$ を収集する．
(2) q と e_i の関係の非典型度と 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[*4] 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係の非典型度および関連語の認知度を測り，たとえば “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 の記事から抽出し，意外な情報としてユーザに提示する．提案手法の特徴の 1 つとして，Wikipedia から得られるリンク情報や語の上位下位関係のみを用いて意外な情報を発見する点があげられる．また，本研究では “Wikipedia 上の情報を統合的に扱うことで，主題語として与えられた語に対する世の中の一般的な人の認識を表すことができる” と仮定する．したがって，Wikipedia 上の情報を統合的に扱う提案手法によって得られる主題語に関する意外な情報とは，世の中の一般的な人にとって意外な情報である．また，本研究では発見された意外な情報の有用性については考慮しない．我々の最終的な目標はユーザの興味を引く有用な意外な情報を発見することであるが，本稿ではそのための最初のアプローチとして，有用性を考慮せず，一般のユーザにとって意外な情報を発見することを目的とする．

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 語を入力語として実験を行った．実験により，主題語と関連語の関係の非典型度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．

本稿の以降の構成は下記のとおりである．2 章では関連研究について述べる．3 章では本研究が対象とする意外な情報について述べ，4 章では主題語に対する各関連語の意外度を求める手法を説明する．5 章で実験結果を示し，6 章ではまとめと今後の課題について述べる．

[*4] http://ja.wikipedia.org/

2. 関連研究

2.1 意外な情報発見に関する研究

情報検索の分野においては，入力されたクエリに対する適合度や人気度に基づいて文書をランキングしてユーザに提示するための研究が多く行われてきた [5], [8], [11], [12], [14], [15], [24], [25], [27]．これらの研究では，上位にランキングされた文書や Web ページには，クエリに関する一般的な情報が含まれることが多い．そのため，既存の検索アルゴリズムではクエリに関する意外な情報を発見するのは困難である．クエリに対する検索結果の順序を逆にしてユーザに提示したとしても，上位の検索結果はクエリに対して不適合な文書がほとんどであり，意外な情報の発見には有効ではないと考えられる．

情報抽出の分野では，Web 上から有用な知識を発見するための様々な手法が提案されてきた [1], [6], [7], [9]．その中には，機械学習を用いるもの [6], [7] や構文パターンを用いるもの [9]，ブートストラップ手法を用いるもの [1] などがある．これらの研究では，より効果的な情報検索を実現するためにコンピュータにとって理解可能な知識体系を構築することが目的であるので，意外な情報よりもよく知られた一般的な情報を抽出する問題に取り組んでいる．

これまでにも意外な情報を発見することを目的とした研究はいくつか存在する [17], [19], [20], [21]．Noda ら [21] は Wikipedia のカテゴリ間の関係を分析することで，（Wikipedia の見出し語，カテゴリ名 1，カテゴリ名 2）で表される 3 つ組の中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，たとえば “麻生太郎” という Wikipedia の見出し語を入力語として与えると，“日本の内閣総理大臣” と “オリンピック射撃競技日本代表選手” という 2 語のカテゴリ名が意外性のある語として出力される．すなわち，“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” という情報を発見することができる．彼らの手法では，複数個の 3 つ組に対して意外か意外でないかを人手でラベル付けした教師データが必要であるのに対して，我々の提案手法では教師データが必要ないという優位性がある．Nadamoto ら [20] 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，コミュニティ型コンテンツで，1 語の Wikipedia の見出し語と，それについて語られている文集合を入力として与える．それらの入力に対して，入力として与えた見出し語の Wikipedia の記事内から，コミュニティ型コンテンツで語られていないトピックで，かつコミュニティ型コンテンツで語られているトピックとは最も類似度の低いトピックに関する文集合を発見し出力する．つまり，与えられた文集合には存在しない意外な情報を発見することを目的としているのに対して，本研究では入力として与えた Wikipedia の見出し語の記事内の文集合中に存在する意外な情報を発見することを目的としている点が異なる．Liu ら [17] の研究では，2 つの Web ページを与えたときに，一方の Web ページにしか含まれていない意外な情報を，各 Web ページに含まれるキーワードを比較することで発見する手法を提案している．彼らが 2 つの Web 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．Mejova ら [19] は我々と同様，入力として与えられた 1 語のクエリに対する意外な語を発見するための手法を提案している．彼らの手法は，ある文書集合の中で，出現する文書数が少なく，かつクエリとの共起度が高い語はクエリにとって意外な語であるという考えに基づいている．つまり，彼らが発見しようとしている意外な語の性質を，本研究で用いている “関係の典型度” と “語の認知度” という考えに基づいて述べると，“クエリとの関係が典型的であるが認知度の低い語” であるといえる．これに対して我々が発見しようとしている意外な語は “クエリとの関係が非典型的であるが認知度の高い語” である．意外な情報の性質には様々なものが存在すると考えられるが，本研究と Mejova らの研究では，対象とする意外な情報の性質が異なる．

また，消費者の商品の購入記録などから相関ルールを抽出する研究においては，頻繁に出現するルールは自明のルールであり有用でないため，意外な相関ルールを発見する研究が行われてきた [4], [22], [23]．しかし，これらの研究では構造化されたデータから意外な情報を発見することを目的としているため，構造化されていない一般の Web ページに彼らの手法を適用するのは困難である．

2.2 典型性に関する研究

認知心理学の分野では，典型性に関する研究が行われてきた [2], [3], [18]．我々はこの中でも Barsalou [3] による研究に着目する．Barsalou [3] は典型性の 3 つの観点について，人が考えるオブジェクトの典型度との関係を調査している．3 つの観点とは，central tendency，ideals，frequency of instantiation である．Central tendency とは，“あるカテゴリにおいて類似したオブジェクトが多いオブジェクトほど典型的である” というものである．たとえば，“哺乳動物” というカテゴリにおいて，“犬” は他の多くの哺乳動物と類似している．それとは逆に，“鯨” は他の多くの哺乳動物とは類似していない．そのため，“犬” は “鯨” よりも典型的な哺乳動物であると人は判断する．次に，ideals とは，“あるカテゴリが満たすべき状態を満たしているオブジェクトほど典型的である” というものである．たとえば，“ダイエットのときに食べる食べ物” というカテゴリにおいて，

オブジェクトが満たすべき条件の 1 つとして “カロリーがゼロである” が考えられる．したがって，“寒天” は “ピザ” よりもダイエットのときに食べる食べ物として典型的であると人は判断する．Frequency of instantiation とは，“カテゴリの構成員として人がよく遭遇するオブジェクトほど典型的である” というものである．たとえば，京都の観光地として “金閣寺” はメディアなどで紹介されることが多く，結果的に人が見聞きすることが多くなる．そのため，“京都の観光地” というカテゴリにおいて “金閣寺” は典型的な観光地であると人は判断する．

本研究で対象とするクエリと関連語の典型度は，これら 3 つの観点の中でも Central tendency に近いものであり，クエリと関連語の関係と類似した関係の多さに基づいて関係の典型度を求めている（詳細は 3 章で述べる）．

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．たとえば，“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，共通の上位語を持つような語のことである．たとえば，“落合博満” と “王貞治” は，“野球選手” という共通の上位語を持つため，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，より同位語らしいと考えられる．このように，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．本研究における同位語らしさについては 4.3.1 項で詳しく述べる．

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，首位打者を獲得するのは野球選手であることを考えると，意外な情報にはなりにくいといえる．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” という語を持ちうるためである（**図 1**）．これは，“落合博満” と “首位打者” の関係と類似した関係が多いため，central tendency において “落合博満” と “首位打者” の関係は典型的であるといえる．

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

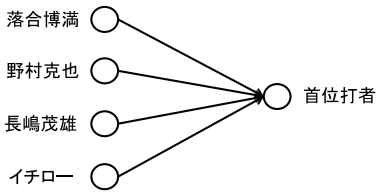


図 1　“首位打者” という関連語は “落合博満” の同位語らしい語の関連語でもある

Fig. 1　Related term “batting champion” is also related to appropriate coordinate terms of “Hiromitsu Ochiai.”

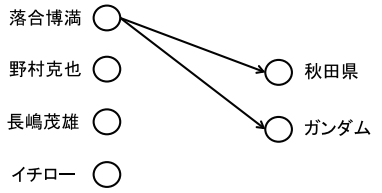


図 2　“秋田県” や “ガンダム” という関連語は “落合博満” の同位語らしい語の関連語ではない

Fig. 2　Related terms “Akita Prefecture” and “Gundam” are not related to appropriate coordinate terms of “Hiromitsu Ochiai.”

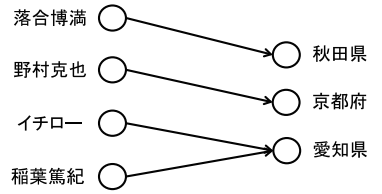


図 3　“落合博満” の同位語らしい語は，関連語として “秋田県” の同位語らしい語を持つ

Fig. 3　Appropriate coordinate terms of “Hiromitsu Ochiai” include appropriate coordinate terms of “Akita Prefecture” as a related term.

考える．この場合，“落合博満” のより同位語らしい語の関連語には，“秋田県” や “ガンダム” という語はまったく含まれないか，ごく一部の同位語の関連語にのみ含まれる（**図 2**）．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持っているためである（**図 3**）．この場合も，“落合博満” と “秋田県” の関係と類似した関係が多いため，central tendency において “落合博満” と “秋田県” の関係は典型的であるといえる．一方後者の情報の場合，野球選手と野球関連の語との関連度に比べると，野球選手とアニメ関連の語との関連度は低いため，“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

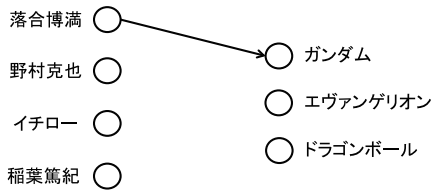


図 4　“落合博満” の同位語らしい語は，関連語として “ガンダム” の同位語らしい語を持たない

Fig. 4　Appropriate coordinate terms of “Hiromitsu Ochiai” do not include appropriate coordinate terms of “Gundam” as a related term.

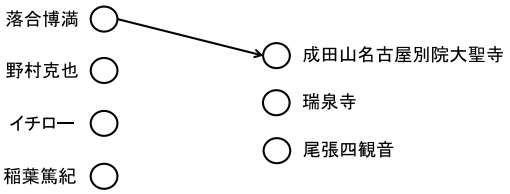


図 5　“落合博満” の同位語らしい語は，関連語として “成田山名古屋別院大聖寺” の同位語らしい語を持たない

Fig. 5　Appropriate coordinate terms of “Hiromitsu Ochiai” do not include appropriate coordinate terms of “Naritasan Nagoya Betsuin Daisyoji Temple” as a related term.

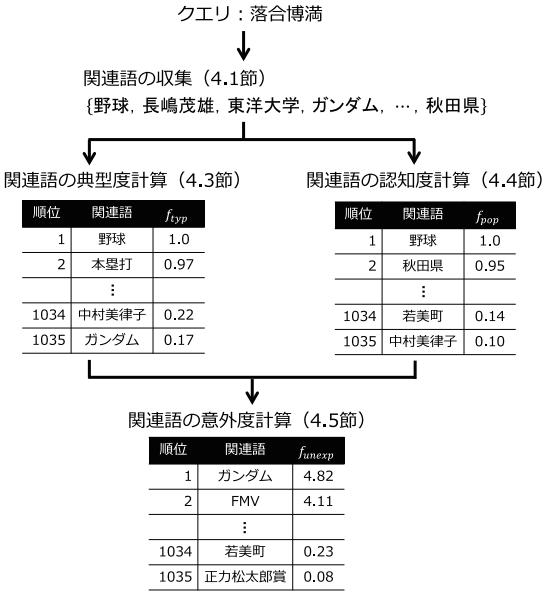
ダム” の同位語らしい語を持っていないためである（図 4）．これは，“落合博満” と “ガンダム” の関係と類似した関係が少ないため，central tendency において “落合博満” と “ガンダム” の関係は非典型的であるといえる．

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（図 5）．この場合も，“落合博満” と “成田山名古屋別院大聖寺” の関係と類似した関係は少ないため，central tendency において “落合博満” と “成田山名古屋別院大聖寺” の関係は非典型的であるといえる．しかし，この情報の意外性は低いと考えられる．この理由として，“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，主題語に対する関連語の意外度を測るためには，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，“主題語と非典型的な関係を持ち，かつ認知度の高い関連語を含む情報” 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 の関係の典型度を求める関数 $f_{typ}(q,e)$ と e の認知度の高さを求める関数 $f_{pop}(e)$ を定義し，最終的にそれらを組み合わせた関数 f：

$$f_{unexp}(q,e) = f(f_{typ}(q,e), f_{pop}(e)) \quad (1)$$

を定義することで q に対する e の意外度を測る．



順位	関連語	f_{typ}
1	野球	1.0
2	本塁打	0.97
	⋮	
1034	中村美律子	0.22
1035	ガンダム	0.17

順位	関連語	f_{pop}
1	野球	1.0
2	秋田県	0.95
	⋮	
1034	若美町	0.14
1035	中村美律子	0.10

順位	関連語	f_{unexp}
1	ガンダム	4.82
2	FMV	4.11
	⋮	
1034	若美町	0.23
1035	正力松太郎賞	0.08

図 6　クエリが “落合博満” のときの意外な関連語を求めるまでの流れ

Fig. 6　Overview of ranking unexpected related terms for a query “Hiromitsu Ochiai.”

4.　主題語と関連語の組合せの意外度の計算

本章では，ある主題語に対して関連語集合を与えたときに，主題語に対する各関連語の意外度を計算する手法について述べる．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 $L_q = \{e_1, e_2, \cdots, e_n\}$ 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 と各関連語の関係の典型度 $f_{typ}(q,e_i)$ を求める．
(4) 各関連語の認知度 $f_{pop}(e_i)$ を求める．
(5) 主題語に対する各関連語の意外度 $f_{unexp}(q,e_i)$ を求める．

クエリを “落合博満” としたときの上記の流れを具体的に示すと図 6 のようになる．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[*5]．情報源としては，q をクエリとしたときの Web 検索結果や，QA サイトで q について言及されているページなども考えられるが，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，一般に語 q をクエリとした

[*5] 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た．

Web 検索の結果中には，クエリとは無関係な文章も多数含まれており，そのような文書集合の中から関連語を収集するとノイズとなる語も多く含まれてしまうためである．そのため，我々が提案する手法でクエリに関する意外な情報の候補が発見された後で，その情報がクエリに関するものであるかを確認する必要が生じる．一方，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おり，ノイズとなる語が比較的含まれにくいため，提案手法により発見された情報はクエリに関する情報であることが高い確率で保証されているという利点がある．したがって，本研究では，クエリに関する意外な情報を発見するという問題だけに着目するために Wikipedia を用いる．

2 つ目は，たとえば語 q に関して QA サイトや一般の Web ページに記述されている情報には主観的なものも含まれるが，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 の記事は関連語を収集する情報源として適しているといえる．これらに加えて，Wikipedia では，見出し語の記事の内容と関係があり，Wikipedia に記事が存在する語にはリンクを作成し，内容と関係のない単なる日本語の単語などに対しては，Wikipedia に記事が存在する場合でもリンクを作成しないようガイドラインで定められている[*6]．そのため，ある見出し語の記事中でリンクが作成されている語があった場合，その語は見出し語の関連語であると見なすことができる．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すべての語をその主題語の関連語集合として収集する．

Wikipedia の記事から関連語を収集するにあたって，本研究では “主題語が見出しである Wikipedia の記事内に記述されている関連語は主題語と一定以上の関連がある” ということを仮定している．逆に，主題語が見出しである Wikipedia の記事内に記述されていない語は主題語との関係が低く，そのような語を含む情報は一般的な情報であり意外な情報にはなりえないと仮定する．つまり，我々が主題語にとって意外な関連語を発見する際は，主題語との関係が一定以上の語の中から発見している．

4.2　同位語および上位語の取得

3 章でも述べたように，本研究では主題語に関する意外な情報を，主題語とその同位語，および主題語の関連語とその同位語をもとに求める．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下位関係抽出ツール[*7]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り，約 20 万語の上位語と約 245 万語の下位語を含む．これを用いることで，ある語の上位語や，ある語と共通の上位語を持つ同位語を求めることが可能となる．ある語の上位語は 1 つとは限らず，複数個存在することもある．たとえば，“落合博満” という語の上位語は，“野球監督” のほかにも “神主打法の選手” や “男性” などがあり，全部で 45 語の上位語を持つ．これら 45 語のうち，“落合博満” と少なくとも 1 つ上位語を共有している語は，なんらかの観点において “落合博満” と同位語であるといえる．

4.3　主題語と関連語の関係の典型度

本節ではまず，**図 7** を用いて提案手法のアイデアを直感的に説明する．図 7 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 7 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 および H_{lq} の語に対応している．

また，枝については，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 の間に枝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 $y = hypo(x)$.

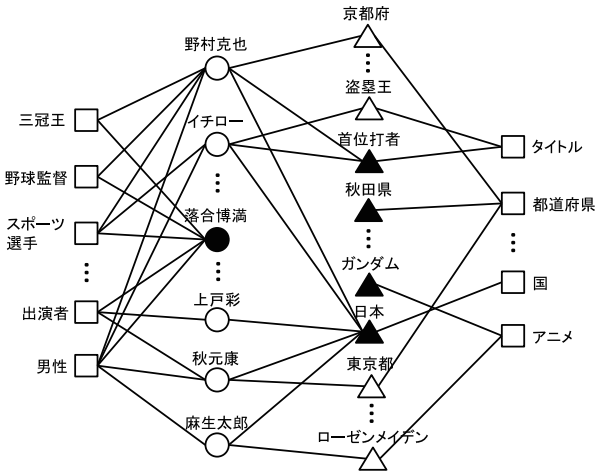


図 7　主題語を “落合博満” としたときに構成されるグラフの例
Fig. 7　An example of the graph for a theme term “Hiromitsu Ochiai.”

*6　http://ja.wikipedia.org/wiki/Wikipedia:記事どうしをつなぐ
*7　http://alaginrc.nict.go.jp/hyponymy/index.html

- (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，q と $x \in L_q$ の間に枝は存在しない．これは，主題語と関連語の関係の典型度を測るために，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たどりつ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たどりつ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．すでに述べたように，“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たくさんあるためであり，このようなとき，容易にたどりつけると考える．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” や “落合博満 → スポーツ選手 → イチロー → 愛知県 → 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ないと考えられる．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” のように “ガンダム” に到達するパスは存在する．このような場合，“落合博満” から “ガンダム” にたどりつくのは困難であると考える．

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典型度を求めるために，上記のグラフを構築する．そして，主題語から関連語への直接のリンクが存在しないグラフ上において，直接のリンクを考慮しない場合に予想される，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見なす．さらに，主題語から関連語へのグラフ上でのたどりつきやすさを主題語と関連語の関係の典型度と見なす．つまり，主題語からたどりつきにくい関連語ほど，主題語との関係が非典型的な語となる．

次節以降では，上記のグラフを 3 つの部分グラフに分割し，主題語と関連語間の関係の典型度を段階的に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

本研究では，次の 2 つの特徴を持つ語をある語 t にとって同位語らしい語とする．

(1-A)：t と多くの上位語を共有している．

(1-B)：t と下位語の数が少ない上位語を共有している．

これらの特徴を，“落合博満” というクエリを例として説明する．“野村克也” と “麻生太郎” の 2 語では，“野村克也” の方が同位語らしいと考えられる．これは，(1-A) を考えることで説明でき，上位下位概念辞書内で “落合博満” と “麻生太郎” は “存命人物” のみを共通の上位語を持つのに対して，“落合博満” と “野村克也” は “存命人物” のほかにも “野球選手” や “野球監督” など多くの上位語を共有しているためである．次に，“松坂大輔” と “小倉智昭” の 2 語では，“松坂大輔” の方が同位語らしいと考えられる．(1-A) だけを考慮すると，“野村克也” と “松坂大輔” は “存命人物” と “野球選手” を上位語として共有し，“落合博満” と “小倉智昭” は “存命人物” と “秋田県出身の人物” を上位語として共有しているため，同位語らしさは同じになる．この場合，(1-B) を考えることで説明でき，“秋田県出身の人物” よりも “野球選手” の方が下位語の数が少ないため，“松坂大輔” の方が同位語らしいといえる．

以上の考えに基づいて，主題語との同位語らしさを求めるため，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．

特徴 (1-A) と (1-B) を反映させるため，本研究では HITS アルゴリズム [14] に基づいた手法を用いることで C_q の各語について q との同位語らしさを求める．HITS アルゴリズムは Web ページ間のリンク構造を用いて Web ページの重要度を求めるためのアルゴリズムである．HITS アルゴリズムでは，各 Web ページはハブ度とオーソリティ度と呼ばれる値を持つ．ハブとオーソリティはそれぞれ，良いハブは多くの良いオーソリティをリンクしている Web ページであり，良いオーソリティは多くの良いハブからリンクされている Web ページである，と再帰的に定義されている．

語 h_i のハブ度を x_i，語 t_j のオーソリティ度を y_j とすると，x_i および y_j の値は次式により求められる．

$$x_i = \sum_{t_j \in Q \cup C_q} w_{ji}^{th}\, y_j, \quad y_j = \sum_{h_i \in H_q} w_{ij}^{ht}\, x_i. \tag{2}$$

ここで，w_{ji}^{th} および w_{ij}^{ht} は枝の重みであり，w_{ji}^{th} は t_j から h_i への枝の重みを表す．

HITS アルゴリズムでは，t_j と h_i の間に枝がある場合，いずれの枝の重みも 1 である．2 部グラフ G_1 に HITS アルゴリズムを適用した場合，以下の 2 つの性質を満たす語が主題語の同位語らしい語として求められる．

(1) 主題語と多くの上位語を共有する語．

(2) 下位語の数が多い上位語を主題語と共有する語．

その結果，たとえば “存命人物” のように多くの下位語を持つノードが非常に大きな値を持つことになる．それに

ともない “存命人物” の下位語が不当に大きな値を持ってしまう．本稿ではこれを防ぎ，特徴 (1-B) に沿うようにするため，上位語から下位語に値を伝播させるときに各上位語の下位語の数に応じて枝の重みを工夫する．HITS アルゴリズムにおいて枝の重みを考慮したものとして SALSA アルゴリズム [16] が提案されており，SALSA アルゴリズムでは，より多くの枝を持つノードほど，枝の重みは小さくなる．具体的には，式 (2) において h_i から t_j への枝の重みを $w_{ij}^{ht} = \frac{1}{|hypo(h_i)|}$，$t_j$ から h_i への枝の重みを $w_{ji}^{th} = \frac{1}{|hyper(t_j)|}$ とする．SALSA アルゴリズムを用いるのは q の同位語らしさを求めるためであるので，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．ノードの値が収束したときの $t_j \in C_q$ の値を q の同位語らしさとする．

4.3.2　主題語の同位語を考慮した主題語と関連語の関係の典型度の計算

本項では，主題語の同位語からその関連語へのリンク構造，関連語間のリンク構造，関連語から主題語の同位語へのリンク構造に基づいて，主題語と主題語の各関連語の関係の典型度を推定する．そのために，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．つまり，$x \in L_q$ から $y \in L_q$ などにも枝が存在しうる．G_2 を有向グラフとしたのは，ある語 t_i が見出し語である Wikipedia の記事内で t_j へのリンクが作成されており，t_i が主題語にとって典型度の高い語であれば，t_i と関係のある t_j も主題語にとってある程度典型度の高い語である，というように主題語に対する典型度は語間の関係，つまりリンク構造によって伝播していくと考えたためである．さらに，t_j が見出し語である Wikipedia の記事内に t_i へのリンクが存在しなければ，t_i は t_j にとって一定以上の関係がない語であり，主題語にとっての t_j の典型度は t_i には伝播させないようにもできる．

上述した，主題語と主題語の各関連語の関係の典型度を推定するための我々のアイデアの特徴は次の 2 点である．

(1) グラフ G_2 において，より多くの，主題語にとって関係が典型的と予期される語からリンクされている語は主題語にとって関係が典型的と予期される語である．
(2) グラフ G_2 において，主題語の同位語らしい語および，同位語らしい語からリンクされている語は主題語にとって関係が典型的と予期される語である．

このアイデアを実現するための手法として，我々は biased PageRank アルゴリズム [12] を用いる．biased PageRank アルゴリズムを用いる理由は，biased PageRank アルゴリズムのアイデアの特徴として以下の 2 点があげられるためである．

(1) より多くの重要なページからリンクされているページは重要である．
(2) あらかじめ重要であることが分かっているページおよび，そのページからリンクされているページは重要である．

つまり，biased PageRank の特徴 (1) において “重要なページ” を “主題語にとって関係が典型的と予期される語” と見なしており，我々のアイデアの特徴 (1) と対応している．また，biased PageRank の特徴 (2) において “あらかじめ重要であることが分かっているページ” を “あらかじめ主題語にとって同位語らしいことが分かっている語” と見なしており，我々のアイデアの特徴 (2) と対応している．ここで，主題語にとっての同位語らしさは 4.3.1 項で求められたものである．

biased PageRank アルゴリズム [12] の詳細を述べる前に，PageRank アルゴリズム [5] について述べる．PageRank アルゴリズムはリンク構造を用いて Web ページの重要度を計算するための手法である．PageRank アルゴリズムの基本的なアイデアは，多くの重要な Web ページからリンクされている Web ページは重要である，というものである．つまり，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 Web 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 Web 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 $\alpha = 0.85$ とした．biased PageRank では，式 (3) 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．

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

$f_{co}(v)$ 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．

これにより，biased PageRank をグラフ G_2 に適用した際に，q の同位語らしい語および，その語の関連語の値が高くなり，上記の我々のアイデア (2) が実現される．式 (4) 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 と関係の典型度が低い語であるといえる．

4.3.3　関連語の同位語を考慮した主題語と関連語の関係の典型度の計算

本項では，関連語の同位語を考慮することで，主題語と各関連語の関係の最終的な典型度を求める．そのために，語 $t \in L_q$ に対して，我々はまずすべ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．本項において主題語と主題語の各関連語の関係の典型度を推定するための我々のアイデアの特徴は次の 4 点である．

(1) ある関連語 t の多くの同位語らしい語が，主題語にとって関係が典型的であると予期される語であれば，t は主題語にとって関係が典型的であると予期される語である．
(2) ある関連語 t の多くの同位語らしい語が，主題語にとって関係が非典型的であると予期される語であれば，t は主題語にとって関係が非典型的であると予期される語である．
(3) 4.3.2 項で biased PageRank を適用した結果，主題語にとって関係が典型的と予期される関連語 t は，t の同位語を考慮した場合でも主題語にとって関係が典型的な語である．
(4) 4.3.2 項で biased PageRank を適用した結果，主題語にとって関係が非典型的と予期される関連語 t は，t の同位語を考慮した場合でも主題語にとって関係が非典型的な語である．

このアイデアを手法として実現するために本研究では Co-HITS アルゴリズム [8] を用いるが，その理由は以下の 2 点である．1 点目は，上記の特徴の (1) と (2) では関連語の同位語らしい語を求める必要があるが，これは 4.3.1 項で主題語の同位語らしい語を求めたものと同じ状況であり，その際に SALSA アルゴリズムを用いたためである．SALSA アルゴリズムは後述の式 (5) および (6) において $\lambda_u = \lambda_v = 1$ としたものと同等であり，Co-HITS アルゴリズムの特殊な状況であるといえる．

2 点目の理由を説明するために，まず 2 部グラフ (V_1, V_2, E) を考える．V_1 内のノード間，V_2 内のノード間に枝は存在せず，V_1 内のノードと V_2 内のノードの間にのみ枝は存在する．Co-HITS アルゴリズムでは各ノードの初期値を考慮することができ，パラメータの値を調整することで初期値が高いノードは Co-HITS の適用後も値が高いままに，初期値が低いノードは Co-HITS の適用後も値が低いままにすることができる．これらの性質は上記の特徴 (3) および (4) を実現するのに適しており，“初期値が高いノード” を “4.3.2 項で biased PageRank を適用した結果，主題語にとって関係が典型的と予期される関連語”，“初期値が低いノード” を “4.3.2 項で biased PageRank を適用した結果，主題語にとって関係が非典型的と予期される関連語” と見なしている．これに加えて，上述のパラメータは V_1 内のノードに共通のものと V_2 内のノードに共通のものの 2 種類があり，個別に値を設定できる．この性質は，特徴 (3) および (4) において，関連語の上位語は主題語との関係の典型度を初期値として持たないため上位語集合の初期値は考慮をする必要がなく，関連語とその同位語の初期値のみ考慮すればよいことを実現するのに適している．

以下で Co-HITS アルゴリズム [8] の詳細を述べる．$u_i \in C_t$ のオーソリティ度を x_i，$v_j \in H_t$ 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．

$$x_i = (1 - \lambda_u)x_i^0 + \lambda_u \sum_{v_j \in H_t} w_{ji}^{vu} y_j. \tag{5}$$

$$y_j = (1 - \lambda_v)y_j^0 + \lambda_v \sum_{u_i \in C_t} w_{ij}^{uv} x_i. \tag{6}$$

ここで x_i^0 および y_j^0 はそれぞれ u_i と v_j の初期値である．H_t に含まれる語に関しては，事前に語の重要度が決まっていないため，ノードの初期値をすべて 0 とする．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持つ．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．また，$w_{ij}^{uv} = \frac{1}{|hyper(u_i)|}$，$w_{ji}^{vu} = \frac{1}{|hypo(v_j)|}$ 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，λ_v の値は 1 とした．λ_u の値の違いによる影響については 5 章で検証する．λ_u の値が大きいほど，上述の我々のアイデアの特徴 (3) および (4) において，biased PageRank を適用した結果を重視することになる．この場合，Co-HITS アルゴリズムは Taher ら [28] が提案した Personalized PageRank アルゴリズムと同じものになる．

上記の計算を主題語 q の各関連語 e_i について行う．式 (5) により求められる e_i の値を $f_{typ}(q, e_i)$ とする．

4.4　関連語の認知度の計算

ある語の認知度を測る方法の 1 つとして，語をクエリとして Web 検索を行った際のヒット件数を用いて，ヒット件数が多い語ほど認知度が高いとする方法が考えられる．ヒット件数を取得する一般的な方法として，商用検索エンジンが提供する Web 検索 API を使用することがあげられるが，API の 1 日あたりの使用回数に限度が設けられていたり，Yahoo!JAPAN が 2013 年 8 月 14 日で Web 検索 API の提供を終了したり[*8]しているため，多くの語のヒット件数を取得するのは困難である．

そこで本研究では．Wikipedia の全記事集合に対してリンク関係を枝とするグラフを構築して PageRank アルゴリズム [5] を適用し，PageRank の値を認知度とする方法を

[*8] http://techblog.yahoo.co.jp/topics/search_api_close/

表 1　実験に用いた主題語の例
Table 1　Examples of queries used in our experiments.

カテゴリ	250 個以上の関連語を持つ主題語	250 個未満の関連語を持つ主題語
人物名	聖徳太子，タモリ，野比のび太	舟木智介，東久邇成子
地名	モナコ，ライン川，金星	大洲藩，海南島
製品名	エアバッグ，駅弁，あしたのジョー	リズムギター，二足歩行ロボット
施設名	名古屋駅，映画館，東京スカイツリー	アメリカ議会図書館，横浜スタジアム
組織名	ユニクロ，サッカー日本代表，三洋電機	三井グループ，大学生協

用いる．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 の値を持つため，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 $f_{pop}(e_i)$ とする．

4.5　関連語の意外度の計算

これまで述べてきた，主題語と関連語の関係の典型度および関連語の認知度を用いて，主題語に対する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典型度は，値が低いほど主題語との関係が非典型的であり，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，主題語 q に対する関連語 e_i の意外度 $f_{unexp}(q, e_i)$ を次式により求める．

$$f_{unexp}(q, e_i) = \frac{1}{f_{typ}(q, e_i)} \cdot f_{pop}(e_i). \tag{7}$$

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．実験は大きく分けて以下の 3 つから構成される．

(1) 同位語らしさに関する実験.
(2) 語の認知度に関する実験.
(3) 意外な情報の発見に関する実験.

まず，上記の実験 (1) および (3) で共通して用いられる主題語について述べる．本実験では，人物名，地名，製品名，施設名，および組織名の 5 つのカテゴリそれぞれに対して 15 語，合計 75 語の主題語を用いた．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まったくない場合，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すべての記事の PageRank の値を計算し，その値の上位 5%にあたる 17,325 語の主題語を認知度が一定以上ある語として抽出した[*9]．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，250 語以上の関連語を持つ主題語からなり，もう一方の集合 B は関連語の数が 250 語未満の主題語からなる．集合 A には 4,854 語，集合 B には 12,471 語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 語ずつ，集合 B からランダムに 5 語ずつ選んだ．選択された主題語の例を**表 1** に示す．

[*9] ある見出し語の記事の PageRank の値とその見出し語の認知度に高い相関があることは 5.2 節で述べる．

5.1　同位語らしさに関する実験

5.1.1　ベースライン手法

本実験では，以下に述べる 2 つの手法をベースライン手法として用いた．

1 つ目は，本研究で使用した上位下位関係抽出ツールにおいて，主題語との上位語の共有数に応じて同位語をランキングする手法である．この手法は，主題語と多くの上位語を共有する語ほど，主題語の同位語らしい語であるという仮説に基づいている．以下，この手法を上位語共有手法とする．

2 つ目は，4.3.1 項の 2 部グラフに対して SALSA アルゴリズムを用いるのではなく，HITS アルゴリズムを用いる手法である．つまり，式 (2) において $w_{ji}^{th} = w_{ij}^{ht} = 1$ とした手法であり，すべてのノードの値が収束したときの各同位語の値を主題語の同位語らしさと見なす．この手法は，主題語と多くの上位語を共有する語は良い同位語であり，主題語の同位語を多く下位語に持つ語は良い上位語であるという仮説に基づいている．さらに，良い上位語は多くの良い同位語を下位語として持ち，良い同位語は多くの良い上位語を上位語として持つという考えのもと HITS アルゴリズムを用いている．ここでの “良い同位語” とは，主題語にとって同位語らしいという意味である．以下，この手法を HITS 手法とする．

5.1.2　評価方法

4.3.1 項および 5.1.1 項で述べたいずれの手法でも，クエリのすべての同位語について同位語らしさの値は求まる．しかし，本実験で用いたクエリに対して，上位下位概念辞書を用いて得られる同位語の数は平均で 139,631 語と非常に多いため，そのすべての同位語らしさを人手で評価するのは困難である．そこで，同位語らしさの評価では，各クエリに対して以下の手順で評価を行った．まず，クエリの同位語に対して，提案手法と 2 つのベースライン手法で同

位語らしさを求める．次に，各手法での同位語らしさの値の高い上位 100 語をプールし，ランダムに並べる．これらの語に対し，20 代の男性 3 名が独立にクエリの同位語らしさのスコア付けを行った[*10]．スコアをつける際は，クエリとまったく同位語らしくない場合は 0，クエリとそれなりに同位語らしい場合は 1，クエリとかなり同位語らしい場合は 2 とした．ただし，提示された同位語を評価者が知らないためにスコア付けができない場合は，スコアを付けずに “unknown” のラベルを付与するようにした．

評価には Mean Average Precision（MAP）および Normalized Discounted Cumulated Gain（nDCG）[13] を用いた．各手法の nDCG と MAP を求めるために共通する操作として，まず 3 名のうち 2 名以上が “unknown” のラベルを付けた同位語は評価の対象外とした．実験に用いた 3 つの各手法では，クエリの同位語らしい語の上位 100 語のリストが得られているが，評価の対象外となった同位語はそのリストから除き，残った同位語をその同位語らしさの値に応じて再度順位付けする．評価の対象として残った各同位語については，“unknown” のラベルを付けていない評価者の評価値の平均値をそれぞれ同位語らしさの正解値とする．同位語らしさに対する各手法の MAP の値を求めるためには，クエリの同位語を，同位語らしい語として適切な語と適切でない語の 2 種類に分類する必要がある．本実験では，評価値の平均値が 1 以上の語を同位語らしい語として適切な語とし，評価値の平均値が 1 未満の語を同位語らしい語として不適切な語とした．

MAP の場合，値が高いほど，より上位に正解の同位語を多くランキングできていることを表している．nDCG の場合，主題語 q の同位語を正解値が高い順に並べたものを理想的なリストとし，ある手法により求められた同位語らしさの値が高い順に同位語を並べたリストが理想のリストと完全に一致していれば，この手法の q に対する nDCG の値は最も高い 1 となる．理想のリストとの異なりが大きくなるほど，nDCG の値は小さくなる．

5.1.3　結果

まず，評価者間の評価値の一致度を表す quadratic weight による κ 係数 [10] を**表 2** に示す．2 名の評価者間の κ 係数を求める際は，両者ともに “unknown” のラベルを付与していない語のみを対象とした．検定を行った結果，いずれの評価者間でも 1%の有意水準で評価が一致しており，同位語らしさの評価者間での κ 係数の平均値は 0.627 と高い値を示した．

次に，2 名以上の評価者によって “unknown” のラベルが付与され，評価の対象外となった語の割合について述べる．すべてのクエリでは平均 48.0%の語が評価の対象外となった．評価の対象外となる語が多すぎると，評価に用いる同位語のランキング結果が元のランキング結果と大きく異なり，手法ごとの正確な評価を行うのが困難になる．そのため，本実験では評価の対象外となった同位語の割合が 70%を超えるクエリはこれ以降の同位語らしさに関する評価では用いないこととした．その結果，40 語のメジャー語と 8 語のマイナー語が同位語らしさに関する評価の対象となった．評価の対象外となった同位語の割合の違いによる影響を調べるため，評価の対象外となった同位語の割合が 40%以下のクエリのみを用いて以下に述べる結果を求めたが，70%としたときの傾向と大きな差はなかったため，閾値の値によらず似たような傾向の結果が得られると考えられる．

3 つの手法に対する MAP の値を**表 3** に示す．全カテゴリに対する MAP の値を見ると，提案手法が最も高い値となった．提案手法と上位語共有手法の間に有意差はなかったが，提案手法と HITS 手法の間には 1%の有意水準で差があった．

nDCG を求める際に, nDCG@10 を求めることのできる主題語は 47 語であったが, nDCG@20 では 43 語, nDCG@30 では 40 語と減少し, nDCG@100 を求めることのできる主題語は 2 語のみであった．異なるクエリ数から求めた nDCG の値を比較するのは適切でないため, 本実験では nDCG@50 を求めることのできる 30 語を対象として nDCG@10 から nDCG@50 までの値を各手法で求めた．3 つの手法に対する nDCG の値を**表 4** に示す．nDCG@10 から nDCG@50 のいずれの場合も提案手法が最も高い値となった．提案手法と上位語共有手法とはいずれの順位でも有意差はなかったが，提案手法と HITS 手法ではいずれの順位でも 1%の

表 2　同位語らしさの評価者間の κ 係数
Table 2　The kappa agreement of coordinate scores between assessors.

	評価者 1・2	評価者 2・3	評価者 3・1
κ 係数	0.707	0.611	0.564

[*10] 3 名のうち 1 名は著者であるが，3 名とも実験の際の条件は同じであった．

表 3　3 つの手法に対する同位語らしさの MAP．太字は 3 手法間での最大値を表す．$*$ と $**$ はそれぞれ HITS 手法と提案手法の間に 5%，1%の水準で有意差があることを表す
Table 3　Performance comparison for three methods measured by MAP. The highest score in each row is shown in bold. Significant differences with HITS method is indicated by $*$ ($\alpha = 0.05$) or $**$ ($\alpha = 0.01$).

	クエリ数	HITS 手法	上位語共有手法	提案手法
人物名	9	0.105	0.443	**0.484****
地名	11	0.284	**0.519**	0.514
製品名	8	0.105	0.232	**0.258**
施設名	9	0.363	**0.700**	**0.700**
組織名	11	0.305	**0.526**	0.516*
全カテゴリ	48	0.240	0.496	**0.501****

表 4　3 つの手法に対する同位語らしさの nDCG．太字は 3 手法間での最大値を表す．** は HITS 手法と提案手法の間に 1%の水準で有意差があることを表す

Table 4　Performance comparison for three methods measured by nDCG. The highest score in each row is shown in bold. Significant differences with HITS method is indicated by ** ($\alpha = 0.01$).

	HITS 手法	上位語共有手法	提案手法
@10	0.312	0.517	**0.544****
@20	0.351	0.551	**0.566****
@30	0.369	0.582	**0.595****
@40	0.387	0.601	**0.619****
@50	0.402	0.621	**0.628****

表 5　同位語らしさに関するカテゴリごとの nDCG．括弧内の数字はクエリ数を表す

Table 5　nDCG of coordinate terms in each category. Figures in parentheses represent the number of queries.

	人物名	地名	製品名	施設名	組織名
@10	0.604 (9)	0.613 (11)	0.391 (8)	0.755 (9)	0.648 (10)
@20	0.581 (8)	0.718 (9)	0.392 (8)	0.735 (8)	0.700 (10)
@30	0.566 (7)	0.737 (9)	0.361 (7)	0.773 (8)	0.690 (9)
@40	0.542 (5)	0.734 (8)	0.423 (7)	0.790 (4)	0.669 (7)
@50	0.570 (5)	0.712 (3)	0.367 (6)	0.784 (3)	0.619 (6)
@60	0.583 (5)	0.722 (3)	0.377 (6)	0.784 (3)	0.576 (5)
@70	0.594 (4)	0.722 (3)	0.608 (2)	0.763 (2)	0.611 (3)
@80	0.647 (3)	0.817 (2)	-	-	0.492 (2)
@90	0.591 (2)	0.826 (2)	-	-	-
@100	-	0.826 (2)	-	-	-

水準で有意差があった．

提案手法により求められる各カテゴリの nDCG の値を**表 5** に示す．表中の括弧内の数字は，その nDCG の値を計算する際に用いられたクエリの数を表す．この結果から，提案手法は地名および施設名カテゴリで特に有用であり，製品名カテゴリでは有用性がやや低いことが分かる．

たとえば “山本浩二” という主題語に対して得られる結果を観察すると，正解データでは広島東洋カープの選手や有名な元プロ野球選手が同位語らしい語と評価される傾向にあったが，HITS 手法では “山本浩二” の上位語の中でも “存命人物” や “CM 出演者” といった多くの下位語を持つ上位語を共有する語の同位語らしさが高くなっており，芸能人が上位に多く現れていた．上位語共有手法では，“存命人物” と “野球選手” の 2 語を共有する語と，“存命人物” と “広島県出身の人物” の 2 語を共有する語の “山本浩二” に対する同位語らしさが等しくなるといったように，上位語の種類を考慮しないことが原因で精度が低くなる例が見られた．提案手法ではこれらの問題を解決できていたため，最も高い精度で同位語を発見できていた．

以上の結果から，2 つのベースライン手法に対して提案手法の方が有用であるといえる．

表 6　認知度の評価者間の κ 係数

Table 6　The kappa agreement of popularity scores between assessors.

	評価者 1・2	評価者 2・3	評価者 3・1
κ 係数	0.775	0.868	0.830

表 7　ベースライン手法と提案手法により求められた認知度と評価者により求められた認知度の正解値とのピアソンの相関係数

Table 7　Pearson's product-moment correlation coefficient between the popularity calculated by a baseline method or our proposed method and the popularity determined by assessors.

	ベースライン手法	提案手法
ピアソンの相関係数	0.816	0.834

5.2　語の認知度に関する実験

本節では，4.4 節で述べた手法により求められる語の認知度と人が判断する語の認知度の一致度合いの評価について述べる．評価は以下の手順で行った．

まず，4.4 節の手法により Wikipedia 上にあるすべての語のスコアを求める．次に，語のスコアが全体の上位 10%以内の語集合，上位 10%から 20%の間にある語集合，のように 10 分割し，各集合から 10 語，合計 100 語をランダムにサンプリングする．これらの語に対し，20 代の男性 3 名が独立に認知度を 5 段階で評価した[*11]．語に対する 3 名の評価値の平均値をその語の認知度の正解値とした．

ベースライン手法として，各語をクエリとして Web 検索を行った際のヒット件数を認知度とする手法を用いた．ヒット件数を取得する際は Bing Search API [*12]を用いた．

評価者間の評価値の一致度を表す quadratic weight による κ 係数 [10] を**表 6** に，ベースライン手法と提案手法で求められた認知度と評価者により求められた認知度の正解値とのピアソンの相関係数を**表 7** にそれぞれ示す．いずれの評価者間でも，1%の有意水準で評価が一致していた．提案手法により求められる認知度と評価者により求められた認知度の正解値とのピアソンの相関係数は 0.834 であり，1%の有意水準で相関が見られた．また，ベースライン手法を用いた際のピアソンの相関係数は 0.816 であり，提案手法を下回った．これらの結果から，評価者間の κ 係数の平均値は 0.824 と高く，かつピアソンの相関係数も 0.834 と高いため，提案手法により語の認知度が高い精度で求められているといえる．

5.3　意外な情報の発見に関する実験

本節の実験では，以下の 2 つの疑問を明らかにすることを目的とする．

[*11] 3 名のうち 1 名は著者であるが，3 名とも実験の際の条件は同じであった．
[*12] http://datamarket.azure.com/dataset/bing/search

- 意外な情報を発見するために，関連語の認知度を考慮することは重要であるか．
- 意外な情報を発見するために，主題語および関連語の同位語間の関係を考慮することは重要であるか．

これらを明らかにするために，我々は 3 つの提案手法と 4 つの比較手法を用いた．3 つの提案手法では，各関連語の意外度を式 (7) により求める．ここで，式 (5) 中の λ_u の値による影響を調べるため，λ_u の値を 0.25，0.5，0.75 とした場合の結果を比較した．これ以降，式 (7) を用い，λ_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．

1 つ目の疑問を検証するために，主題語と関連語の関係の典型度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 の意外度は次式により計算される．

$$f_{unexp}(q, e_i) = \frac{1}{f_{typ}(q, e_i)}. \tag{8}$$

この手法においても，λ_u の値を 0.25, 0.5, 0.75 とし，それぞれの結果を比較した．λ_u の各値に対する手法を TYP_{25}，TYP_{50}，TYP_{75} とする．

次に，2 つ目の疑問を検証するために，検索結果数のみをもとに意外度を求める手法を用いた．この手法では，関連語と主題語の共起度が低いほど，その関連語は主題語にとって意外な語であると仮定する．つまり，“落合博満 AND ガンダム” のように，主題語と関連語をともに含む Web ページを検索対象とした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[*13]．以後，この手法を HIT 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．その関連語を含む文が複数個ある場合は，記事の先頭に最も近い文を抽出する．

次項以降では，まず実験の手順を述べ，次に評価指標について述べる．最後に，実験結果および考察を述べる．

5.3.1　実験手順

本実験では，5 名の被験者が独立に評価を行った．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[*14]．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，3 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．次に，各手法の上位 5 個の関連語をプールし，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，被験者に各情報について “この情報の意外度を 1 つ選択してください” という文を見せ，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，合計で 75 枚のアンケートを作成し，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，その値を各情報の意外度とした．

5.3.2　評価指標

評価指標には nDCG および Normalized Weighted Reciprocal Rank（NWRR）[26] を用いた．

本節の実験で nDCG を計算する際は，主題語 q に関する意外な情報を意外度の正解値が高い順に並べたものを理想的なリストとし，ある手法により求められた意外度の値が高い順に情報を並べたリストが理想のリストと完全に一致していれば，この手法の q に対する nDCG の値は最も高い 1 となる．理想のリストとの異なりが大きくなるほど，nDCG の値は小さくなる．

NWRR は正解文書の逆順位だけでなく正解レベルも考慮した指標であり，かつ最も上位に出現した正解文書のみを評価の対象とする．本実験では，各情報に対して被験者によって評価された 1 以上 4 以下のスコアが与えられているため，その中間値である 2.5 以上のスコアを持つ情報を意外な情報，つまり正解とした．この指標ではペナルティという概念が用いられ，本実験では被験者によって意外度が高いと評価された情報ほど，その情報に対応する関連語のペナルティは小さくなる．ある情報に対して 5 名の被験者によって評価された意外度のスコアを v（$1 \leq v \leq 4$）とすると，ペナルティを $L = 5 - v$ により求める．r_1 を各手法において最も上位に出現した意外な情報の順位，その順位の情報のペナルティの値を L_{r_1} とすると，WRR は次式により求められる．

$$WRR = \frac{1}{r_1 - 1/L_{r_1}}. \tag{9}$$

さらに，次式により正規化を行う．

$$NWRR = \frac{1 - 1/L_{min}}{r_1 - 1/L_{r_1}}. \tag{10}$$

L_{min} はその主題語に関して発見された情報の中で最も意外度が高いと被験者によって評価された情報のペナルティである．

NWRR は各手法において各主題語ごとに値が求められる．主題語 q に対して，ある手法が正解値の最も高い情報を 1 位にランキングできていればその手法の q に対する NWRR の値は最も高い 1 となり，q に対して，ある手法によって意外であると求められた上位 5 個の情報がいずれも被験者によって意外でないと評価された場合，その手法の q に対する NWRR の値は 0 とする．

[*13] 本研究では，サービスが終了する前の Yahoo!Web 検索 API（http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html）を用いて検索結果数を取得した．
[*14] 評価者の中に著者は含まれていない．

5.3.3　結果

まず，評価者間の評価値の一致度を表す quadratic weight による κ 係数 [10] を**表 8** に示す．評価者 2 を除くと，いずれの評価者間でも評価値は 1%の有意水準で評価が一致していた．評価者 2 と 3，評価者 2 と 4 の評価値が一致しているとはいえないこと，また全体の評価者間の一致度が中程度であることから，意外であると感じる情報は被験者によってある程度差があると考えられる．1 章でも述べたように，本研究では世の中の一般的な人にとって意外である情報を求めるための手法を提案したが，各ユーザに特化した意外な情報を発見するための手法を提案することが今後の課題の 1 つとしてあげられる．

5 つの各カテゴリにおける各手法の DCG@5 の値を**表 9** に示す．この結果を見ると，いずれのカテゴリにおいても 3 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分かる．主題語と関連語の関係の典型度のみを考慮し，関連語の認知度を考慮していない TYP_{25}，TYP_{50}，および TYP_{75} 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．HIT 手法はすべてのカテゴリで最も低い値となった．HIT 手法では，Web のヒット件数の低い語，つまり認知度の低い語が意外な語として上位にランキングされる傾向が見られた．認知度が低い語には 2 種類存在し，1 つは主題語やその同位語らしい語に共通のトピックとは無関係であり認知度の低い語（主題語が野球選手であれば，その選手が生まれた村の名前など）であり，もう 1 つは主題語やその同位語らしい語に共通のトピックに関する語であるが認知度の低い語（主題語が野球選手であれば，野球のマイナーな専門用語など）である．前者の場合，提案手法では関連語の認知度を考慮しており，認知度の低い語は意外な語として上位にランキングされにくいことが理由で HIT 手法よりも良い結果となっていた．後者の場合，提案手法では関連語の認知度を考慮していることに加えて，主題語の多くの同位語と関連のある語は意外な語ではないと見なされるため，提案手法により，意外ではない語として下位にランキングすることができていた．

次に，各カテゴリの NWRR の値を**表 10** に示す．全カテゴリの平均値を見ると，PR_{25} が他の手法よりも上位に意外な情報を発見できていたことが分かる．HIT および TYP_{25} は製品名と人物名のカテゴリでそれぞれ NWRR の値が最も高かった．しかし，この 2 つの手法の全カテゴリの平均値は提案手法よりも低く，nDCG の平均値も低かった．この結果から，語の認知度や同位語を考慮しない場合でも，意外度の高い情報を発見できる場合もあるといえる．その一方で，提案手法の結果が示しているように，それらの要素を考慮することで，どのようなカテゴリにおいても意外度が一定以上の情報を発見することができていた．本実験の結果からは，式 (5) における λ_u の値による大きな違いは見られなかった．

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 11** 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求められ，“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求められていた．一般的に，都道府県は特定の病気と関連を持っておらず，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

表 8　意外度の評価者間の κ 係数．** は評価者間の評価値が 1%水準で一致していたことを表す

Table 8　The kappa agreement of unexpectedness scores between assessors.　** represents that inter-assessor agreement was statistically significant at $\alpha = 0.01$.

	評価者 1	評価者 2	評価者 3	評価者 4
評価者 2	0.210**			
評価者 3	0.264**	0.0422		
評価者 4	0.437**	0.164**	0.331**	
評価者 5	0.208**	0.0462	0.206**	0.268**

表 9　5 つのカテゴリに対する各手法の nDCG@5 の値．提案手法と HIT 手法の間の統計的有意差は $*$（$\alpha = 0.05$）または $**$（$\alpha = 0.01$）で表されている．同様に，† と ‡ はそれぞれ TYP_{25} 手法，TYP_{50} 手法との統計的有意差を表す

Table 9　Performance comparison of each category for seven methods measured by nDCG@5. Significant differences with HIT is indicated by $*$ ($\alpha = 0.05$) or $**$ ($\alpha = 0.01$). Similarly, † and ‡ indicate significant differences with TYP_{25} and TYP_{50}, respectively.

手法	人物名	地名	製品名	施設名	組織名	平均
HIT	0.705	0.757	0.773	0.787	0.780	0.760
TYP_{25}	0.805	0.792	0.837	0.800	0.853	0.817
TYP_{50}	0.807	0.803	0.839	0.800	0.857	0.821
TYP_{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**†‡

表 10　5 つのカテゴリに対する各手法の NWRR の値

Table 10　Performance comparison of each category for seven methods measured by NWRR.

手法	人物名	地名	製品名	施設名	組織名	平均
HIT	0.307	0	**0.478**	0	0	0.157
TYP_{25}	**0.513**	0.118	0.319	0.215	0.165	0.266
TYP_{50}	0.506	0.118	0.327	0.199	0.177	0.266
TYP_{75}	0.506	0.163	0.332	0.194	0.177	0.274
PR_{25}	0.434	0.184	0.341	**0.418**	**0.194**	**0.314**
PR_{50}	0.418	0.184	0.361	0.241	**0.194**	0.280
PR_{75}	0.421	**0.199**	0.361	0.241	**0.194**	0.283

表 11　提案手法により発見された意外な情報の例
Table 11　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．
自動販売機	景観	光害の問題や景観に対する悪影響も指摘されている．
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，「$2/3 \div 0.25 \div 0.8 = 10/3$」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 12　5 つの各カテゴリにおいて意外な情報が発見された主題語数
Table 12　The number and the ratio of theme terms that could find unexpected information.

カテゴリ	関連語が 250 個以上	関連語が 250 個未満	合計
人物名	6/10	3/5	9/15
地名	3/10	0/5	3/15
製品名	5/10	2/5	7/15
施設名	4/10	0/5	4/15
組織名	2/10	1/5	3/15

慣病” という語の意外度を高く評価できていた．他にも，“野比のび太” という主題語に対して “一次方程式” という語が意外な語として発見されており，この場合もアニメの登場人物と数学用語との関連の低さおよび “一次方程式” という語の認知度の高さから発見できたといえる．ただし，この例の場合，“野比のび太” は一般に勉強が苦手な人物であることが知られており，その印象とは逆の情報が発見されたため，被験者にはより意外に感じられたと考えられる．このように，情報の意外度を求める際は，主題語特有の性質を考慮することも今後の課題の 1 つとしてあげられる．

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 12** に示す．関連語を 250 語以上持つ主題語については 40%の割合で，また関連語が 250 語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．この割合は十分に高いとはいえないが，主題語をランダムに選んだにもかかわらず，40%もの割合で意外な情報を発見できていたともいえる．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分かった．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．たとえば，“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（2010 年 4 月現在）.” 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

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外度を計算する手法の提案を行った．我々は，主題語と非典型的な関係を持ち，かつ認知度の高い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たどりつきにく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，主題語と関連語の関係の典型度を考慮すること，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．

提案手法では，一般的なユーザにとっての語の認知度に基づいて関連語の意外度を求めた．しかし，たとえば “ガンダム” という語は高齢のユーザの間では認知度はそれほど高くないと考えられる．そのため，今後は年齢や性別，さらには個人にとっての語の認知度を考慮したうえで関連語の意外度を求め，各ユーザに特化した意外な情報の発見を行う予定である．

また，意外な情報には様々なものがあるが，本稿で提案したモデルにより発見できる意外な情報はその一部にすぎない．今後はより多様な意外な情報を発見するために，新たなモデルの提案を行うことを考えている．

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 24240013，24680008，12J03993）および平成 25 年度研究拠点形成費等補助金若手研究者養成費（卓越した大学院拠点形成支援補助金）によるものです．ここに記して謝意を表します．

参考文献

[1] Agichtein, E. and Gravano, L.: Snowball: Extracting relations from large plain-text collections, *Proc. ACM DL 2000*, pp.85–94 (2000).
[2] Barsalou, L.: Ad hoc categories, *Memory & Cognition*, Vol.11, No.3, pp.211–227 (1983).
[3] Barsalou, L.: Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol.11, No.4, pp.629–654 (1985).
[4] Berger, G. and Tuzhilin, A.: Discovering unexpected patterns in temporal data using temporal logic, *Temporal Databases Research and Practice*, Lecture Notes in Computer Science, Vol.1399, pp.281–309 (1998).
[5] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Proc. WWW 2007*, pp.107–117 (1998).
[6] Carlson, A., Betteridge, J., Wang, R.C., Hruschka, Jr., E.R. and Mitchell, T.M.: Coupled semi-supervised learning for information extraction, *Proc. ACM WSDM 2010*, pp.101–110 (2010).
[7] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. and Slattery, S.: Learning to construct knowledge bases from the World Wide Web, *Artif. Intell.*, Vol.118, No.1-2, pp.69–113 (2000).
[8] Deng, H., Lyu, M.R. and King, I.: A generalized Co-HITS algorithm and its application to bipartite graphs, *Proc. ACM SIGKDD 2009*, pp.239–248 (2009).
[9] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A.: Web-scale information extraction in knowitall: (preliminary results), *Proc. WWW 2004*, pp.100–110 (2004).
[10] Fleiss, J.L. and Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurement*, Vol.33, pp.613–619 (1973).
[11] Gyöngyi, Z., Garcia-Molina, H. and Pedersen, J.: Combating web spam with trustrank, *Proc. VLDB 2004*, pp.576–587 (2004).
[12] Haveliwala, T.H.: Topic-sensitive PageRank, *Proc. WWW 2002*, pp.517–526 (2002).
[13] Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques, *ACM Trans. Inf. Syst.*, Vol.20, No.4, pp.422–446 (2002).
[14] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, pp.604–632 (1999).
[15] Lempel, R. and Moran, S.: The stochastic approach for link-structure analysis (SALSA) and the TKC effect, *Proc. 9th International World Wide Web Conference on Computer Networks: The International Journal of Computer and Telecommunications Netowrking*, pp.387–401 (2000).
[16] Lempel, R. and Moran, S.: SALSA: The stochastic approach for link-structure analysis, *ACM Trans. Inf. Syst.*, Vol.19, pp.131–160 (2001).
[17] Liu, B., Ma, Y. and Yu, P.S.: Discovering unexpected information from your competitors' web sites, *Proc. ACM SIGKDD 2001*, pp.144–153 (2001).
[18] Medin, D. and Smith, E.: Concepts and concept formation, *Annual Review of Psychology*, Vol.35, pp.113–138 (1984).
[19] Mejova, Y., Bordino, I., Lalmas, M. and Gionis, A.: Searching for interestingness in Wikipedia and Yahoo!: Answers, *Proc. WWW 2013*, pp.145–146 (2013).
[20] Nadamoto, A., Aramaki, E., Abekawa, T. and Murakami, Y.: Content hole search in community-type content, *Proc. WWW 2009*, pp.1223–1224 (2009).
[21] Noda, Y., Kiyota, Y. and Nakagawa, H.: Discovering Serendipitous Information from Wikipedia by Using Its Network Structure, *Proc. ICWSM 2010*, pp.299–302 (2010).
[22] Padmanabhan, B. and Tuzhilin, A.: A Belief-Driven Method for Discovering Unexpected Patterns, *Proc. ACM SIGKDD 1998*, pp.94–100 (1998).
[23] Padmanabhan, B. and Tuzhilin, A.: Small is beautiful: Discovering the minimal set of unexpected patterns, *Proc. ACM SIGKDD 2000*, pp.54–63 (2000).
[24] Robertson, S.E. and Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, *Proc. ACM SIGIR 1994*, pp.232–241 (1994).
[25] Robertson, S., Zaragoza, H. and Taylor, M.: Simple BM25 extension to multiple weighted fields, *Proc. ACM CIKM 2004*, pp.42–49 (2004).
[26] Sakai, T.: On the Properties of Evaluation Metrics for Finding One Highly Relevant Document, *Information and Media Technologies*, Vol.2, No.4, pp.1163–1180 (2007).
[27] Svore, K.M. and Burges, C.J.: A machine learning approach for improved BM25 retrieval, *Proc. ACM CIKM 2009*, pp.1811–1814 (2009).
[28] Taher, H., Sepandar, K. and Glen, J.: An Analytical Comparison of Approaches to Personalizing PageRank, *Stanford University Technical Report 2003* (2003).
[29] Tuzhilin, A.: On subjective measures of interestingness in knowledge discovery, *Proc. ACM SIGKDD 1995*, pp.275–281 (1995).

佃 洸摂　（学生会員）

2011 年京都大学大学院情報学研究科社会情報学専攻博士前期課程修了．同年同大学院博士後期課程進学．2012 年より日本学術振興会特別研究員（DC1）．

大島 裕明　（正会員）

京都大学大学院情報学研究科社会情報学専攻特定准教授．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）．主に情報検索，ウェブマイニング，デザインの研究に従事．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

（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および情報検索の研究に従事．

岩崎 弘利　（正会員）

（株）デンソーアイティーラボラトリ CTO．1990 年名古屋大学大学院工学研究科博士課程前期課程電気・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，電子情報通信学会各会員．

田中 克己　（正会員）

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院博士前期課程修了．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，ウェブ検索の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，日本データベース学会各会員．

（担当編集委員　宮尾 祐介）