エンティティ間の関連性に基づく意外な情報の発見

佃　洸摂† 大島　裕明† 山本　光穂†† 岩崎　弘利†† 田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻　〒 606–8501 京都府京都市左京区吉田本町
†† 株式会社デンソーアイティーラボラトリ　〒 150–0002 東京都渋谷区渋谷 2-15-1 渋谷クロスタワー 28F
E-mail: †{tsukuda,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp, ††{miyamamoto,hiwasaki}@d-itlab.co.jp

あらまし　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．本稿において対象とする意外な情報とは，例えば「落合博満はガンダムのファンである.」のようなものである．我々は,「落合博満」や「ガンダム」のように，意外な情報を形成するうえで欠かせない語をエンティティと呼ぶ．本手法では，そのようなエンティティの組を大規模に収集し，各組についてそれらのエンティティの同位語間の関連の強さを Wikipedia の単語間のリンク構造に基づいて求める．これにより，通常はクエリの同位語集合が関連を持たない同位語集合およびエンティティが求められる．

キーワード　同位語，Wikipedia，HITS

1. はじめに

近年，インターネットと Web 検索エンジンの普及に伴い，様々な情報を得ることができるようになってきた．ユーザが検索エンジンを用いてあるクエリを入力して Web ページの検索を行う際，検索エンジンが返す Web ページ集合の中には，クエリに関する様々な情報が含まれる．そのような情報の中には，クエリに関する一般的な情報や，クエリに関する “意外な情報” などがある．例えば，クエリが “落合博満” という語であった場合，“落合博満は元プロ野球選手である．” 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情報であると言える．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の中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うもれてしまい，発見が困難であるという問題がある．そこで我々はある語に関して一般にはあまり知られていない情報を “意外な情報” と呼び，意外な情報を発見することを目的とする．

ユーザがあるクエリを入力して Web 検索を行う際に，意外な情報をユーザに提示することは以下の 2 つの点で有用であると考える．1つは，そのクエリについて調べようとしているユーザに対して意外な情報を提示することで，ユーザのそのクエリに関する興味をより引き立てる効果があるという点である．もう1つは，そのクエリについてある程度知っているユーザに対して意外な情報を提示することで，ユーザのそのクエリに関する知識をより深めることを支援できるという点である．また，ユーザがクエリを入力して検索しているとき以外にも，ある地域をドライブしているユーザに対して，その周辺地域に関する意外な情報を提示したり，Web 上でニュースを閲覧しているユーザに対して，そのニュースに関する意外な情報を提示することで，周辺地域やニュース記事の内容に関してより興味を喚起できると考えられる．

本研究では，1 つの情報を “主題語” と “関連エンティティ” という観点から捉える．例えば “落合博満” に関する “落合博満は元プロ野球選手である．” という情報において，主題語は “落合博満”，関連エンティティは “元プロ野球選手” である．主題語および関連エンティティの定義については 3 節で述べる．そして，主題語と関連のある語（関連エンティティ）を大量に収集し，主題語と各関連エンティティの関連性に着目し，意外な情報を発見する手法を提案する．我々は関連エンティティを収集する情報源として Wikipedia を用いる．本稿では，大きく以下の 2 つの段階に分けて意外な情報を発見する．

（1） 主題語と各関連エンティティの組の意外度に基づくランキング.
（2） 主題語と各関連エンティティの組に対する情報の Web からの発見.

まず，主題語と各関連エンティティの組の意外度に基づくランキング手法を提案する．本稿では，Wikipedia の単語間のリンク構造および単語の上位下位関係に着目して，(“落合博満”, “野球”) という組合せよりも (“落合博満”, “ガンダム”) という組合せの方が意外度が高いと判定する手法について述べる．さらに，意外度の高い語の組合せを特定したあとに，それらの語を含む情報を Web 上から発見するための手法についても述べる．これは，意外度の高い語の組合せとして (“落合博満”, “ガンダム”) という語の組が得られた際に，“落合博満はガンダムマニアである．” という情報を Web 上から発見し，ユーザに提示するための手法である．

2. 関連研究

近年，検索結果や推薦結果の意外性に着目した研究が盛んに行われている．

情報検索では，野田ら [1] が 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の提案を行なっている．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．灘本ら [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 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．Liu らは，ある企業が自社の Web ページには含まれていないが競合関係にある企業の Web ページには含まれている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．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，人が意外であると感じる相関ルールについて調査を行なっている研究があり [4] [5] [6]，それらをもとに意外な相関ルールを発見する研究が行われてきた [7] [8]．これらの研究は，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 ページには適用できない．

一方，推薦システムについても，利用者の嗜好に適したアイテムの推薦だけでなく，利用者にとって意外な情報を推薦することの重要性が注目されている [9]．Sarwar ら [10] は協調フィルタリングを用いた推薦システムにおいて，意外性のある推薦手法の提案を行った．彼らは従来のユーザに基づいた協調フィルタリングではなく，アイテムに基づいた協調フィルタリングを優先することで，意外性のあるアイテムの推薦を可能とした．Kamahara ら [11] は，テレビ番組を対象とし，ユーザが属するコミュニティの情報をもとに，ユーザの嗜好との部分的な類似性をもとに意外なテレビ番組を推薦する手法を提案している．また，村上ら [12] は推薦結果の意外性を，“評価対象である推薦システムの予測結果とプリミティブな推薦システムの予測結果との差異にある” と仮定し，推薦結果の意外性を評価する指標を提案している．

3. 意外な情報

本節では，本稿で我々が対象とする意外な情報の定義について述べる．

本稿で我々が対象とする情報は “主題語” と “関連エンティティ” から構成されるような情報である．例えば “落合博満はガンダムマニアである．” という情報を提示する状況として，ユーザが “落合博満” という語をクエリとして Web 検索を行なっていたり，“落合博満” に関するニュースを閲覧していたり

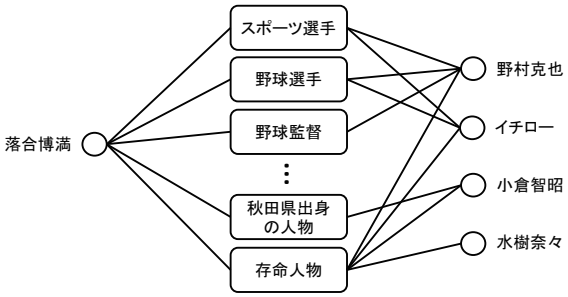

図 1　共通の上位語数による同位語らしさの違い．

といった状況があげられる．すなわち，“落合博満” という語を入力として，意外な情報を発見する必要がある．このように，入力として与えられる語集合を “主題語” と呼ぶ．一方，ある語に対して関係のある語を “関連エンティティ” と呼ぶ．例えば，“落合博満” という語の関連エンティティとしては，“元プロ野球選手” や “中日ドラゴンズ”，“ガンダム” など，様々な語があげられる．

次に，意外な情報の定義を述べる前に，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．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，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，“落合落合” 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（図 1）．

以上を基に，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エンティティ，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，関連エンティティとして “首位打者” という語を持ちうるためである（図 2）．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” のより同位語らしい語の関連エンティティには，“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エンティティにのみ含まれる（図 3）．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，関連エ

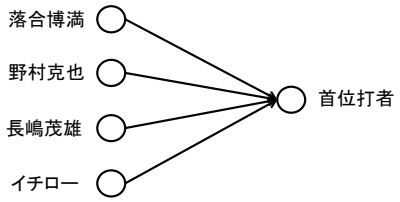

図 2 “首位打者” という関連エンティティは “落合博満” の同位語らしい語の関連エンティティでもある．

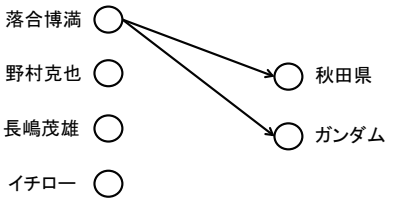

図 3 “秋田県” や “ガンダム” という関連エンティティは “落合博満” の同位語らしい語の関連エンティティではない．

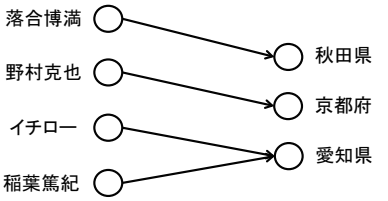

図 4 “落合博満” の同位語らしい語は，関連エンティティとして “秋田県” の同位語らしい語を持つ．

ンティティとして “秋田県” のより同位語らしい語を持っているためである（図 4）．一方後者の場合，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，関連エンティティとして “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 5）．

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エンティティとして持たない（図 6）．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，主題語に対する各関連エンティティの意外度を測るためには，各関連エンティティの認知度も考慮する必要がある．

以上より， 本研究では意外な情報を，“主題語の多くの同位語らしい語が，認知度の高い関連エンティティ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エンティティを含む情報” と定義する．一方，意外でない情報は “主題語の多くの同位語らしい語が，関連エンティティまたはその同位語らしい語と関連を持つような主題語と関連エンティティを含む情報” と定義できる．よって，本稿では，ある主題語 t とある関連エンティティ e が与えられたときに，t と e の関連の低さを求める関数 $Rel(t,e)$ と e の認知度の高さを求める関数 $Cog(e)$

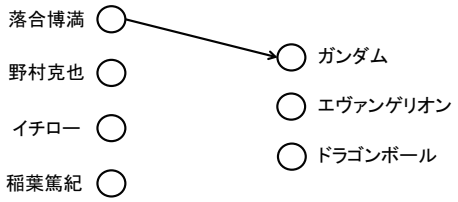

図 5 “落合博満” の同位語らしい語は，関連エンティティとして “ガンダム” の同位語らしい語を持たない．

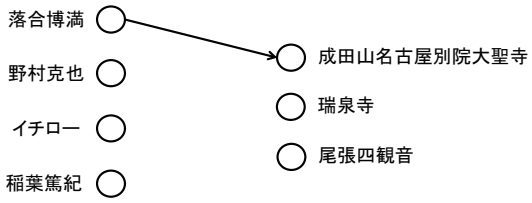

図 6 “落合博満” の同位語らしい語は，関連エンティティとして “成田山名古屋別院大聖寺” の同位語らしい語を持たない．

を定義し，最終的にそれらを組み合わせた関数 f：

$$Unexp(t,e) = f(Rel(t,e), Cog(e)) \qquad (1)$$

を定義することで t と e の組合せの意外度を測ることになる．

4. 主題語と関連エンティティの組合せの意外度の計算

本節では，ある主題語に対して関連エンティティ集合を与えたときに，主題語と各関連エンティティの組合せの意外度を計算する手法について述べる．主題語と各関連エンティティの組合せ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 t を与え，t に対する関連エンティティ集合 $E_t = \{e_1, e_2, \cdots e_n\}$ を収集する．
（2） 主題語 t の上位語集合と同位語集合，各関連エンティティの上位語集合と同位語集合を収集する．
（3） 主題語 t と各関連エンティティの組合せの関連度 $Rel(t,e_i)$ を求める．
（4） 各関連エンティティの認知度 $Cog(e_i)$ を求める．
（5） 主題語と各関連エンティティの組合せの意外度 $Unexp(t,e_i)$ 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 関連エンティティ集合の収集

入力語 t に対する関連エンティティを収集する情報源としては様々なものが考えられる．例えば，語 t をクエリとして Web 検索を行い，検索結果のタイトルやスニペット中の語，あるいはページ内に含まれる語を収集する方法が考えられる．また，語 t に関する意外な情報を発見することが目的であるならば，語 t に “蘊蓄” や “トリビア” といった語を加えて Web 検索結果を取得し，検索結果中に含まれる語を関連エンティティ集合とすることもできる．他にも，QA サイトなど，特定のサービスにおいて語 t について言及されているページのみを収集し，その中から関連エンティティ集合を求めることも考えられる．

本稿において我々が関連エンティティを収集する情報源とし

て選択したのは主題語 t 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，一般に語 t をクエリとした 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エンティティ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．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，例えば語 t 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 の記事は関連エンティティを収集する情報源として適していると言える．さらに，Wikipedia の記事中では，Wikipedia のページが存在する語にはリンクが作成されるが，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は作成しないことになっている．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エンティティであるとみなすことができる．以上のような理由から，本研究では入力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関連エンティティ集合として収集する．

4.2 同位語および上位語の取得

3 節でも述べたように，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エンティティ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語階層データ(注1)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，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も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．例えば，“落合博満” という語の上位語は，“野球監督” の他にも “神主打法の選手” や “男性” など，全部で 45 個の上位語を持つ．これら 45 個のうち，“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” と同位語であると言える．また，“落合博満” とより多くの上位語を共有している語は，“落合博満” のより同位語らしい語であると言える．

4.3 主題語と関連エンティティの組合せの関連度の計算

本研究では，主題語 t と関連エンティティ $e_i \in E_t$ の組合せが意外であるということを，“主題語 t の多くの同位語らしい語が，認知度の高い e_i および e_i の同位語と強い関係を持たない” と定義した．逆に，主題語 t と関連エンティティ e_i の組合せが意外ではないということは，“主題語 t の多くの同位語らしい語が e または e の同位語と強い関係を持つ” と定義した．

本研究では，主題語と関連エンティティの関連の有無を判断するために，語をノード，上位下位関係および語間のリンクをエッジとするグラフを構築する．そして，主題語と関連エンティティの関連の有無を，その 2 語間のパスの有無とみなす．さらに，主題語と関連エンティティの関連の強さを，主題語から関連エンティティへのグラフ上での辿り着きやすさとみなすことで，2 語の組合せの関連度を測る．つまり，主題語から辿り着きにくい関連エンティティほど，主題語と関連の低い語となる．ただし，主題語の同位語と主題語の関連エンティティの関係の強さを測るためには，主題語からその同位語を経由した際の，主題語から各関連エンティティへの辿り着きやすさを求める必要がある．そのため，主題語からその関連エンティティへのエッジは存在しないものとしてグラフを構築する．

以下では，まずグラフを構成するノード集合とエッジ集合について述べる．次にグラフを 3 つの 2 部グラフに分割し，各 2 部グラフに対して HITS アルゴリズム [13] をベースとした手法を適用することで 2 語の関連度を測る手法について説明する．

4.3.1 グラフの構築

まず，グラフは以下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 の関連エンティティ集合を $rel(t)$ とする．

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

また，エッジについては，上位語・下位語の関係にある 2 語の間および，ある語とその関連エンティティの間に存在する．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 $x = hyper(y)$.

既に述べたように，このグラフにおいて，q と $x \in L_q$ の間にエッジは存在しない．

4.3.2 同位語らしさの評価

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 $G_1 = (H_q \cup T, E_1)$ を作成する．ここで，$T = Q \cup C_q$ であり，E_1 は H_q と T の間の枝集合である．語 $h_i \in H_q$ と語 $t_j \in T$ 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 HITS アルゴリズムを適用し，C_q の各語について，q との同位語らしさを求める．語 h_i の値を x_i，語 t_j の値を 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．

$$x_i = \sum_{t_j \in T} w_{ji}^{th} y_j \qquad (2)$$

(注1)：http://nlpwww.nict.go.jp/corpus/

$$y_j = \sum_{h_i \in H_q} w_{ij}^{ht} x_i \quad (3)$$

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 への遷移確率を表す．HITS アルゴリズムでは，いずれの枝の重みも 1 である．

ただし，上記の 2 部グラフに HITS アルゴリズムを適用した場合，例えば “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．それに伴い “存命人物” の下位語が不当に大きな値を持ってしまう．その結果，多くの下位語を持つ上位語を共有する語が主題語の同位語らしい語となってしまう．本稿ではこれを防ぐため，SALSA アルゴリズム [14] を用いる．SALSA アルゴリズムでは，あるノードから出ている枝の数が多くなるほど，その枝の重みは小さくなる．具体的には，h_i から出ている枝の重みは $w_{ij}^{ht} = \frac{1}{|hypo(h_i)|}$ となる．つまり，下位語から上位語に値を伝播させるときはすべての下位語から出ている枝の重みは同じであり，一方で上位語から下位語に値を伝播させるときは各上位語に対する下位語の数に応じて枝の重みは変わる．

4.3.3 関連エンティティの関連度の評価 1

次に，C_q および L_q，L_c から構成される 2 部グラフ $G_2 = (C_q \cup L, E_2)$ を作成する．ここで，$L = L_q \cup L_c$ であり，E_2 は C_q と L の間の枝集合である．Wikipedia において語 $c_i \in C_q$ が見出しとなっているページ内で語 $l_j \in L$ にリンクが張られているとき，c_i と l_j の間に枝が存在する．

この 2 部グラフに対しても先ほどと同様に HITS アルゴリズムを用いて，各関連エンティティの主題語に対する意外度を求める．このとき，関連エンティティの中に例えば “日本” のように多くの語 $c_i \in C_q$ からリンクされている語があると，その語が高い値を持つことになる．その結果，その高い値を持つ語にリンクしている同位語らしくない語の値が高くなってしまう．これを防ぐために，関連エンティティから同位語に値を伝播する際は SALSA アルゴリズムを用いる．

また，本研究では，より同位語らしい語からリンクされている関連エンティティほど，意外度は低いと考えるため，4.3.2 節で求めた各同位語の同位語らしさを C_q のノードの初期値として Co-HITS アルゴリズム [15] を用いる．x_i^0 を語 c_i の初期値，y_j^0 を語 l_j の初期値，語 c_i の値を x_i，語 l_j の値を 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．

$$x_i = (1 - \lambda_c) x_i^0 + \lambda_c \sum_{l_j \in L} w_{ji}^{lc} y_j \quad (4)$$

$$y_j = (1 - \lambda_l) y_j^0 + \lambda_l \sum_{c_i \in C_q} w_{ij}^{cl} x_i \quad (5)$$

ここで $\lambda_c \in [0.1]$ および $\lambda_l \in [0.1]$ は x_i^0，y_j^0 をどれだけ重要であるとみなすかを表すパラメータであり，値が小さいほど x_i^0 および y_j^0 を重要であるとみなす．

4.3.4 関連エンティティの関連度の評価 2

最後に，L_q および L_c，H_{lc} から構成される 2 部グラフ $G_3 = (L \cup H_{lc}, E_3)$ を作成する．ここで，E_3 は L と H_{lc} の間の枝集合であり，語 $l_i \in L$ と語 $h_j \in H_{lc}$ が上位下位関係にあるとき，l_i と h_j の間に枝が存在する．この 2 部グラフに対して，4.3.3 節で求めた各関連エンティティの主題語との関連度を初期値とした Co-HITS アルゴリズムを適用する．このとき，多くの下位語をもつ上位語は大きな値を持ち，そのような語の下位語が不当に高い値を持つことを防ぐために，上位語から下位語に値を伝播する際は SALSA アルゴリズムを用いる．

以上の 3 つの段階を経て得られる関連エンティティ e_i の値を，主題語 t に対する e_i の関連度 $Rel(t, e_i)$ とする．

4.4 関連エンティティの認知度の計算

本節では，ある主題語に対する各関連エンティティの認知度を求める手法について説明する．ある語の認知度を測る方法として，Wikipedia の記事集合に対してリンク関係をエッジとする構築し PageRank アルゴリズム [16] を適用する方法や，大規模なアンケート調査を行う方法など様々な方法が考えられる．

本稿では Yahoo!ウェブ検索 API (注2) を用いてある語に対する Web 検索の検索結果数を取得することでその語に対する認知度を測る．ある語 e_i の Web 検索の検索結果数を $Cog(e_i)$ とし，検索結果数が多い語ほど，認知度が高い語であるとみなす．

4.5 関連エンティティの意外度の計算

これまで述べてきた，主題語と各関連エンティティの関係の低さおよび各関連エンティティの認知度を用いて，主題語と各関連エンティティの組合せの以外度を以下の方法で求める．これにより，主題語と関連度が低く，認知度の高い語ほど，$Unexp(t, e_i)$ の値が大きくなる．

$$\begin{aligned} Unexp(t, e_i) &= f(Rel(t, e_i), Cog(e_i)) \\ &= \frac{1}{Rel(t, e_i)} \cdot log_{10} Cog(e_i) \end{aligned} \quad (6)$$

5. 意外な情報の発見

本節では，主題語と関連エンティティを入力として，Web 上から意外な情報を発見する手法について述べる．

まず，Yahoo!ウェブ検索 API を用いて主題語と関連エンティティをクエリとして Web 検索を行い，検索結果の上位 k 件のスニペットを取得する．これらのスニペット集合の中から，より重要なスニペットを求めるために，LexRank アルゴリズム [17] を用いて各スニペットの重要度を求める．LexRank アルゴリズムを適用するために，各スニペットに対して MeCab(注3) を用いて形態素解析を行い，各形態素のスニペット集合における $tf \cdot idf$ 値を要素とするベクトルを作成する．$\mathbf{p}$ を各文章の重要度を要素とするベクトルとし，次式で表される LexRank アルゴリズムを用いて再帰的に $\mathbf{p}$ の値を求める．

$$\mathbf{p} = [d\mathbf{U} + (1 - d)\mathbf{B}]^T \mathbf{p}$$

ここで，$\mathbf{U}$ は全要素の値が $\frac{1}{k}$ である k 次正方行列であり，$\mathbf{B}$ は $\mathbf{B}_{ij,}$ がスニペット i とスニペット j の各ベクトルのコサイン類似度であり，各列の値の和が 1 となるように正規化された k 次対称行列である．また，d は $dampingfactor$ であり，本稿

(注2)：http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html
(注3)：http://mecab.sourceforge.net/

表 1 “落合博満” に対する同位語を HITS アルゴリズムと SALSA アルゴリズムを組合せて求めた結果（左）と HITS アルゴリズムのみを用いて求めた結果（右）.

順位	同位語
1	清原和博
2	秋山幸二
3	野村克也
4	二村徹
5	田宮謙次郎
6	平野謙
7	斎藤雅樹
8	野村謙二郎
9	若松勉
10	大沢啓二
11	山本功児
12	星野仙一
13	仁村薫
14	真弓明信
15	権藤博

順位	同位語
1	板東英二
2	金田正一
3	長嶋茂雄
4	広澤克実
5	中畑清
6	赤星憲広
7	岩本勉
8	東尾修
9	江川卓
10	大仁田厚
11	初芝清
12	掛布雅之
13	高橋由伸
14	古田敦也
15	桑田真澄

では $d = 0.15$ とした.

LexRank アルゴリズムを用いることで，各スニペットの重要度を求めることができたが，重要度が高いスニペットのみから情報を抽出した場合，互いに類似した情報となることが多い．我々は，ユーザに情報を提示する際，重要度と多様性を両立させるために MMR アルゴリズム [18] を用いる．MMR アルゴリズムでは，次式により出力する文書を求める．

$$MMR = \underset{d_i \in D \setminus S}{\operatorname{argmax}} [\lambda(Score(d_i)) - (1-\lambda) \max_{d_j \in S} Sim(d_i, dj)]$$

ここで，D はスニペット集合であり，$Score(d_i)$ は LexRank によって求められたスニペット d_i のスコアである．また，S は D のうち出力として選択された文書集合であり，$Sim(d_i, d_j)$ は d_i と d_j のベクトルのコサイン類似度である．$\lambda \in [0, 1]$ は，文書の重要度と，すでに選択された文書の類似度をどの程度重視するかを調整するパラメータであり，本稿では $\lambda = 0.5$ とした.

最後に，MMR アルゴリズムを用いて出力された上位 r 件に対して，スニペットの生成元となった Web ページから入力に用いた関連エンティティを含む 1 文を抽出し，ユーザに提示する.

6. 実　　験

提案手法により得られる，主題語と関連エンティティの組合せの意外度の妥当性を評価するために，実験を行った.

6.1 主題語と関連エンティティの意外度の評価

まず，4.3.2 節で述べた手法によって得られる，主題語の同位語らしい語を示す．比較手法として，上位語から下位語に値を伝播する際も HITS アルゴリズムを適用する手法を用いた．“落合博満” を入力語としたときの結果を表 1 に示す．“落合博満” の同位語は全部で 32 万 9843 語あるが，HITS アルゴリズムのみを用いた場合，10 位に野球選手ではない “大仁田厚” という語が出現する．また，これ以外にも 22 位に “小倉智昭”，23 位に “地井武男” など，野球選手ではない語が上位にランキングされていた．一方，提案手法では，元プロ野球選手のみが上位にランキングされ，上位 100 語は全て野球選手であった.

表 2 “落合博満” とその関連エンティティの関連度の評価 1.

順位	関連エンティティ
1	ガンダムエクシア
1	バンダイホビーセンター
1	成田山名古屋別院大聖寺
1	林るり子
1	落合博満野球記念館
6	ウイングガンダムゼロカスタム
7	若美町
8	鈴木洋史
9	武藤信二
10	センチュリーベストナイン

表 3 “落合博満” とその関連エンティティの関連度の評価 2.

順位	関連エンティティ
1	バンダイホビーセンター
2	若美町
3	成田山名古屋別院大聖寺
4	ウイングガンダムゼロカスタム
5	落合博満野球記念館
6	武藤信二
7	ガンダムエクシア
8	センチュリーベストナイン
9	最高速度
10	鈴木洋史

続いて，4.3.3 節および 4.3.4 節で述べた手法を用いて，主題語に対する各関連エンティティの関連度の低さを求めた結果を表 2 および表 3 に示す．なお，“落合博満” の関連エンティティは全部で 433 語ある．表 2 において，上位 5 つの語は “落合博満” のいずれの同位語の関連エンティティでもなかった語である．表 2 で上位にある語の中で，“林るり子” や “若美町” といった語は，表 3 では下位にランキングされるのが望ましい語であるが，“若美町” の順位は表 3 でも高いままであった．この理由として，“林るり子” は上位語として “出身の人物”，“歌手”，“芸能人” という語を持つため，それらの上位語を介して “林るり子” の語の値は高くなったが，“若美町” は上位語として “町” という語しか持たないため，今回提案した手法では十分に値を高くすることができなかったことがあげられる.

最後に，4.4 節および 4.5 節の手法を用いて “落合博満” に対する各関連エンティティの意外度を求めた結果を表 4 に示す．各関連エンティティの認知度を考慮することで，“成田山名古屋別院大聖寺” のような認知度の低い語の順位を下げることができている.

6.2 Web 上からの情報発見の評価

最後に，表 4 の上位 3 件の語に対して，5 節の手法を用いて各関連エンティティを含む情報を Web 上から 1 件ずつ取得した．比較手法として，主題語に “トリビア” という語を加えて Web 検索を行い，検索結果の上位 100 件のスニペットに対して 5 節の手法を適用し，出力された上位 3 件のスニペットの元

表 4 “落合博満” とその関連エンティティの意外度の評価.

順位	関連エンティティ
1	バンダイホビーセンター
2	若美町
3	ウイングガンダムゼロカスタム
4	成田山名古屋別院大聖寺
5	落合博満野球記念館
6	最高速度
7	ガンダムエクシア
8	消化試合
9	バックスクリーン
10	鈴木洋史

となったページから，主題語に関して意外であると思われる情報を人手で抽出する手法を用いた．また，“落合博満” 以外にも “谷繁元信”，“エッフェル塔”，“十和田湖”，“祇園” の各主題語に対して提案手法で得られた情報を表 5 に，比較手法で得られた手法を表 6 に示す．表 5 では紙面の都合上，“落合博満” 意外の主題語については意外度が高いと判定された上位 2 件の関連エンティティのみを掲載している．また，5 節のパラメータの値を $k = 100$，$r = 5$ とし，出力された文集合の中から最適な 1 件を人手で選択した．

それぞれの結果を見ると，提案手法では 1 位の “バンダイホビーセンター” と 3 位の “ウイングガンダムゼロカスタム” という関連エンティティについては意外な情報が発見できていると言える．2 位の “若美町” という関連エンティティは，“落合博満” の出身町であるため，Web の検索結果からは意外な情報を発見することができなかった．一方，“落合博満” の場合，比較手法でも提案手法とは異なる意外な情報を発見できていた．

また，提案手法では “谷繁元信” という主題語に対して意外度が 2 位であった “パンパース” という関連エンティティから意外な情報を発見することができていた．しかし，比較手法では意外だと思うような情報を発見することができなかった．この結果から，比較手法の有用性は主題語自体の認知度によって大きく異なる可能性があげられる．つまり，`` 落合博満” のように認知度の高い主題語であれば，“トリビア” などの語を用いて明示的に意外な情報を記述しているページが多いが，“谷繁元信” は “落合博満” に比べて認知度は低いと考えられ，明示的に意外な情報として記述されていることは少ないと言える．地物についても，主題語が “エッフェル塔” の場合は提案手法だけでなく比較手法でも意外な情報を発見することができていた．一方，“十和田湖” の場合，提案手法では “鳥インフルエンザ” という関連エンティティに対して意外な情報を発見することができていたが，比較手法では，マイナーすぎる情報や，よく知られた情報のみが発見された．また，主題語が “祇園” の場合，比較手法では意外な情報を発見することができていたが，提案手法では関連エンティティとして得られた語の数が 26 個しかなかったため，意外であると思えるような情報は発見できなかった．このように，提案手法では主題語が見出し語となっている Wikipedia の記事の充実度合いによって得られる関連エンティティの数が大きく異なるため，記事の内容が充実していないような主題語に対しては有効に働かないという問題がある．これを解決するためには，Wikipedia の記事から得られる関連エンティティの数が一定の値より少なければ，Web の検索結果等からも関連エンティティを集めるといった方法が考えられる．

7. ま と め

本稿では，ある入力語に関する意外な情報を発見することを目的とし，入力語とその関連語の組合せの意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が，認知度の高い関連エンティティ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エンティティ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エンティティへの辿り着きづらさと各関連エンティティの認知度を基に主題語と関連エンティティの組合せの意外度を求めた．また，意外度が高いと判定された主題語と関連エンティティの組を与えることで，Web からの意外な情報の発見を行った．

今後は，提案手法により得られた意外な情報がユーザにとって本当に意外な情報であるかについてユーザ実験を通して評価を行う予定である．

謝 辞

本研究の一部は，グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」（拠点リーダ：田中克己）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: Proc. of 4th Int'l AAAI Conference on Weblogs and Social Media, ICWSM'10.

[2] 灘本, 阿辺川, 荒牧, 村上：“コミュニティ型コンテンツのコンテンツホール抽出手法の提案”, 情報処理学会研究報告. データベース・システム研究会報告, **2007**, 65, pp. 265–270 (2007).

[3] B. Liu, Y. Ma and P. S. Yu: “Discovering unexpected information from your competitors' web sites”, Proc. of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '01, pp. 144–153 (2001).

[4] B. Padmanabhan and A. Tuzhilin: “Unexpectedness as a measure of interestingness in knowledge discovery”, Decis. Support Syst., **27**, pp. 303–318 (1999).

[5] B. Liu and W. Hsu: “Post-analysis of learned rules”, Proc. of the thirteenth national conference on Artificial intelligence - Volume 1, AAAI'96, pp. 828–834 (1996).

[6] A. Tuzhilin: “On subjective measures of interestingness in knowledge discovery”, Proc. of the First International Conference on Knowledge Discovery and Data Mining, pp. 275–281 (1995).

[7] B. Padmanabhan and A. Tuzhilin: “Small is beautiful: discovering the minimal set of unexpected patterns”, Proc. of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '00, pp. 54–63 (2000).

[8] B. Padmanabhan and A. Tuzhilin: “A belief-driven method for discovering unexpected patterns”, KDD, pp. 94–100 (1998).

[9] K. Swearingen and R. Sinha: “Beyond algorithms: An hci perspective on recommender systems”, Proc. of the 24th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, pp. 393–408 (2001).

[10] B. Sarwar, G. Karypis, J. Konstan and J. Reidl: “Item-based collaborative filtering recommendation algorithms”,

表 5　提案手法によって発見された情報.

主題語	関連エンティティ	発見された情報
落合博満	バンダイホビーセンター	・また，知る人ぞ知る“ガンダムおたく”の落合博満監督（54）は，静岡市にあるガンダムの聖地バンダイホビーセンターを訪れ，プラモデルの生産工場見学に大はしゃぎだった.
	若美町	・どうして中日の落合監督は，秋田県男鹿市 (旧若美町) 出身なのに和歌山県に自分が活躍した写真やボールの出展があるのに地元にないのでしょうか.
	ウイングガンダムゼロカスタム	・また，ガンダムシリーズの熱狂的なファンでも有名な人物であり，好きなモビルスーツは，ウイングガンダムゼロカスタム（機動戦士ガンダム W）.
谷繁元信	東城町	・中国山地の奥深い広島県比婆郡東城町（現・庄原市）の出身.
	パンパース	・入団当初はリードの覚えが悪く，いつまでもオムツの赤ちゃんという意味で「パンパース」というあだ名を付けられていた.
エッフェル塔	パリ万国博覧会	・1889 年のパリ万国博覧会のメイン・モニュメントとして建てられた鉄製の塔です.
	軍事	・建築から 20 年後には取り壊される予定であったが，無線の時代を迎えると，1906 年にアンテナが設置され，第一次世界大戦時にはドイツの無線を傍受するなど軍事拠点としても活躍した後，現在ではパリの街には欠かせない存在となっている.
十和田湖	とわだこ号	・活彩とわだこ号は青森県八戸市と十和田市にある十和田湖を結ぶ会員制バスである.
	鳥インフルエンザ	・秋田県と環境省は 29 日，十和田湖畔で死んでいた白鳥から検出された H5 亜型の鳥インフルエンザウイルスが強毒性だったと発表した.
祇園	花街	・今日，京都には上七軒，祇園甲部，祇園東，嶋原，先斗町，および宮川町の 6 つの花街があり，これらを総称して京都の六花街と呼ぶことがある.
	上七軒	・上七軒は現在の賑わいからすると祇園あたりには水をあけられている感があるが，京都で最も歴史の古い花街とされている.

表 6　比較手法によって発見された情報.

主題語	発見された情報
落合博満	・落合博満野球記念館には落合監督のブリーフ姿のブロがある. ・落合監督といえば仕事一筋で，家のことなど何もできないようなイメージがあるかもしれません．しかし，信子夫人いわく，意外にも監督は家事がなかなか得意なのだそうです. ・落合の巨人へのＦＡ移籍時に長嶋監督は,「2 割 8 分　15 本でもいい．彼の力が必要」と言ったが，移籍 1 年目，本当にその通りの成績になった.
谷繁元信	（該当情報なし）
エッフェル塔	・エッフェル塔を設計したエッフェル氏は，自由の女神の骨組みも設計している. ・もともと，エッフェル塔はパリ万国博覧会のモニュメントとして建設されたが，パリ市民の反発を考慮し万博が終われば解体撤去する約束であった.
十和田湖	・真珠の御木本幸吉とハマチの野網和三郎と並ぶ,『近代日本の養殖家三偉人』の一人である和井内貞行氏が，支笏湖から譲り受けたヒメマスの養殖が明治 36 年に成功. ・典型的な二重式カルデラ湖である十和田湖を取り巻く紅葉が見事で，特別名勝地にも指定されている.
祇園	・「祇園精舎の鐘の声，諸行無常の響きあり」平家物語の冒頭の文章で誰でも知っているフレーズですが，もともと祇園精舎に鐘はなかったんです. ・八坂神社の名は，祇園坂・長楽寺坂・下河原坂・法観寺坂・霊山坂・山井坂・清水坂・産寧坂の八つの坂が近くにある事に由来する.

Proc. of the 10th international conference on World Wide Web, WWW '01, pp. 285–295 (2001).

[11] J. Kamahara, T. Asakawa, S. Shimojo and H. Miyahara: “A community-based recommendation system to reveal unexpected interests”, Proc. of the 11th International Multimedia Modelling Conference, MMM '05, pp. 433–438 (2005).

[12] 村上：“推薦結果の意外性を評価する指標の提案”, 人工知能学会全国大会 (第 21 回) 論文集, 2007 (2007).

[13] J. M. Kleinberg: “Authoritative sources in a hyperlinked environment”, J. ACM, **46**, pp. 604–632 (1999).

[14] R. Lempel and S. Moran: “Salsa: the stochastic approach for link-structure analysis”, ACM Trans. Inf. Syst., **19**, pp. 131–160 (2001).

[15] H. Deng, M. R. Lyu and I. King: “A generalized co-hits algorithm and its application to bipartite graphs”, Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09, ACM, pp. 239–248 (2009).

[16] S. Brin and L. Page: “The anatomy of a large-scale hypertextual web search engine”, Proc. of the seventh international conference on World Wide Web 7, WWW7, pp. 107–117 (1998).

[17] G. Erkan and D. R. Radev: “Lexrank: graph-based lexical centrality as salience in text summarization”, J. Artif. Int. Res., **22**, pp. 457–479 (2004).

[18] J. Carbonell and J. Goldstein: “The use of mmr, diversity-based reranking for reordering documents and producing summaries”, Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, pp. 335–336 (1998).