典型性に基づく Web 検索と分析に関する研究

A Study on Web Search and Analysis based on Typicality

佃 洸摂 ♥

Kosetsu TSUKUDA

本論文では，典型的な情報および，非典型的かつ有用な情報（意外な情報）の検索手法を提案する．ユーザがあるドメインの情報を調べるときに，典型的な情報を提示することはそのドメインの概要を知るうえで有用である．また，非典型的な情報や意外な情報を提示することは，ドメインをより深く知ったり，ユーザの興味を喚起したりするうえで有用である．提案手法では，認知心理学の分野の典型性に関する研究の知見を用いて，似たものの多さという観点に基づく典型度および、出現頻度の高さという観点に基づく典型度の推定，それらに基づく情報の意外度の推定を行う．具体的には，「典型度に基づくオブジェクト集合検索」，「語の認知度と語間の関係の非典型度に基づく意外な情報の発見」，「社会認知量に基づく語間の関係の典型度推定による意外な情報の発見」に関する研究を行い，提案手法・システムの評価を行った．本研究を通して，これまでの情報検索では十分に考慮されていなかった情報の典型性という概念を利用することで，典型度に基づく検索および分析の方法を提案した．

We propose methods for searching typical information and atypical and useful (unexpected) information. Showing typical instances in a category is useful to understand the outline of the category. After understanding the outline of the category, it is helpful to achieve greater understanding of the category by showing atypical examples and unexpected examples. To compute the degree of typicality and the degree of unexpectedness, we consider typicality based on central tendency and frequency of instantiation, which were proposed in cognitive psychology. Specifically, this paper includes the following three research topics: (1) searching for an object set based on typicality; (2) discovering unexpected information based on the popularity of terms and the typicality of relationships between terms; and (3) measuring perceived strength of the relationship between terms to discover an unexpected relationship.

♥ 正会員　産業技術総合研究所情報技術研究部門メディアインタラクション研究グループ研究員　k.tsukuda@aist.go.jp

1. はじめに

本論文では，典型的な情報および，非典型的かつ有用な情報（意外な情報）の検索手法の提案および分析を行う．典型性に基づく検索というのは様々な場面で有用であると考えられる．たとえば，レシピを検索する際に初めてカルボナーラを作るので，まずは典型的なカルボナーラを探すといったことが可能になる．他にも，典型的なパーキンソン病の症状を知ることで自分の病状と照らし合わせることができたり，長崎に初めて旅行をするのでまずは典型的な観光コースを楽しむといったことが可能になる．さらに，たとえば汚職事件について勉強する際に，典型的な汚職事件を学ぶことで理解を早めることも可能になる．このように，典型的な情報を知ることで，対象のカテゴリの概要を把握することが望める．

典型的な情報だけでなく，非典型的な情報や意外な情報も様々な場面で役立つと考えられる．たとえば料理をする際に，典型的なカルボナーラを作ったあとは非典型的なカルボナーラを探すことでより深くカルボナーラについて知ることが可能になる．他にも，熱中症の意外な症状を知ることで，熱中症の対策を早めにとることができたり，東京観光の最中に東京スカイツリーの意外な情報を提示することで興味を喚起することも期待できる．北海道のお土産を買う際も，いつも典型的なお土産を購入するのではなく，非典型的なお土産を探すことでお土産のバリエーションを増やすことも可能になる．このように，非典型的な情報を知ることで，対象のカテゴリを深く知ったり，興味を喚起したりすることが望める．

しかし，通常の検索エンジンを使って，典型性に基づく検索を行うのは容易ではない．その理由として，以下の 4 つがあげられる．

1. 典型性や意外性に基づく検索意図は認知的検索意図と呼ばれるが，こういった意図を表すクエリは多くのユーザにとって入力が困難であることが示されている．Kato ら [2] の研究の結果から考えると，たとえばあるユーザが典型的な京都のお土産を知りたいと思っても，一般的なユーザは「京都土産」といった検索クエリしか入力できないことが予想される．
2. 検索の対象となる Web ページに「典型」や「意外」といった語が含まれるとは限らないため，「汚職事件 典型」や「東京スカイツリー 意外」といったクエリをユーザが入力できたとしても，ユーザが望む Web ページが検索されるとは限らない．
3. たとえば「カルボナーラ 典型」というクエリを入力して，カルボナーラのレシピが掲載された Web ページが検索されたとしても，検索されたレシピが典型的かどうかはカルボナーラに関する十分な知識がないと判断が困難である．
4. 非典型的な情報にはノイズとなるような情報が多数混在する

ため，その中から有用な情報や意外な情報を探すのは困難である．

そこで，本論文では，典型性に基づいて Web 検索および分析を行う方法を提案した．情報の典型性を測る際には，認知心理学の分野で行われた典型性に関する研究 [1] を基に，似たものがどれだけ多いかという観点に基づく典型度および，出現頻度がどれだけ高いかという観点に基づく典型度を用いる．具体的には,「典型度に基づくオブジェクト集合検索」,「語の認知度と語間の関係の非典型度に基づく意外な情報の発見」,「社会認知量に基づく語間の関係の典型度推定による意外な情報の発見」に関する研究を行い，提案手法・システムの評価を行った．以降の章でそれぞれの研究について述べる．

2. 典型度に基づくオブジェクト集合検索

2.1 導入

本章では，食材というオブジェクトから構成されるオブジェクト集合としてレシピを対象とし，オブジェクト集合の典型度を推定する手法について述べる。レシピの典型度を求めることで，ユーザはたとえば,「カルボナーラを初めて作るので典型的なレシピを探したい」や「カルボナーラは何度か作ったことがあるので一風変わった非典型的なレシピを探したい」といった意図を反映した検索ができるようになる。

2.2 手法

提案手法では，たとえば「カルボナーラ」というカテゴリであれば，その中での食材の出現頻度と，食材間の共起度を基に，最も典型的なカルボナーラを構成する食材集合 O_T を求める。次に，以下の 2 つの要因に基づいてオブジェクト集合 O の典型度を求める。

- オブジェクト集合 O と O_T の差異 : O_T に含まれるオブジェクトとの差異が小さいほど O の典型度は高くなる．
- オブジェクト集合 O 内のオブジェクト間の相性 : O に含まれる任意の 2 オブジェクト間の相性が良いほど O の典型度は高くなる．

典型度の具体的な計算については佃ら [6] を参照のこと．

2.3 実験

提案手法の有用性を示すために，実験を行った．実験には，COOKPAD [1]の「カルボナーラ」「ナポリタン」「ミネストローネ」「豚汁」「トマトサラダ」「ツナサラダ」の 6 カテゴリに含まれるレシピを用いた．各カテゴリで使用したレシピ数は 72，59，140，76，79，83 であった．

実験では提案手法に加えて，認知心理学における典型性の観点を反映した以下の 2 つの手法を用いた．

- Resemblance に基づく手法 : 各レシピを，食材の有無を要素とするベクトルで表現し，ベクトルの類似度を重みとするグラフを作成して TextRank [3] を適用する．TextRank の値の高いレシピほど，似たものが多い典型的なレシピであるとみなす．
- Frequency に基づく手法 : COOKPAD の各レシピに投稿された「つくれぽ」を利用する．つくれぽとは，あるレシピを実際に作ったことを他のユーザが報告したもので，つくれぽ数が多いほど現実世界でのインスタンスが多く Frequency の高い典型的なレシピであるとみなす．

レシピの典型度の正解値を求めるために，各カテゴリから 40 個のレシピをランダムに選択した．各レシピの食材を 3 名の評価者に見せ，その食材から作られるレシピの典型度を 7 段階で回答してもらい，3 名の評価値の平均値をそのレシピの典型度の正解値とした．

典型度の正解値と，3 つの各手法で求められる典型度との相関を求めたところ，特にレシピ間で食材の類似度の高かった，カルボナーラやナポリタン，豚汁といったカテゴリでは，resemblance に基づいて求めた典型度が高い相関を示した．それに対して，Frequency に基づいて求めた典型度はいずれのカテゴリでも低い相関であった．つまり，世の中で沢山つくられているレシピは，少なくとも食材レベルで見れば，典型的でないものが多いことを示している．

提案手法によって求められる典型度は，いずれのカテゴリでも resemblance に基づく典型度と近い振る舞いをしていたが，出現頻度は高いが共起頻度は低い 2 つの要素を含むオブジェクト集合があったときに，提案手法は要素間の共起を考慮しているので典型度は低くなり，Resemblance に基づく手法では要素間の共起は考慮していないので典型度は高くなるという違いが見られた．たとえば，豚汁というカテゴリで，じゃがいも，里芋，さつまいもを同時に使用したレシピがあったときに，いずれの食材もそれぞれの出現頻度は高いが，それらを同時に使用したレシピは少ない．そのため，評価者はそのようなレシピの典型度を低く評価していた．Resemblance に基づく手法では典型度が高く計算されていたが，提案手法では，より正解の順位に近い結果となっていた．

3. 語の認知度と語間の関係の非典型度に基づく意外な情報の発見

3.1 導入

本章では，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について述べる．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わかり，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することができる．

[1] http://cookpad.com/

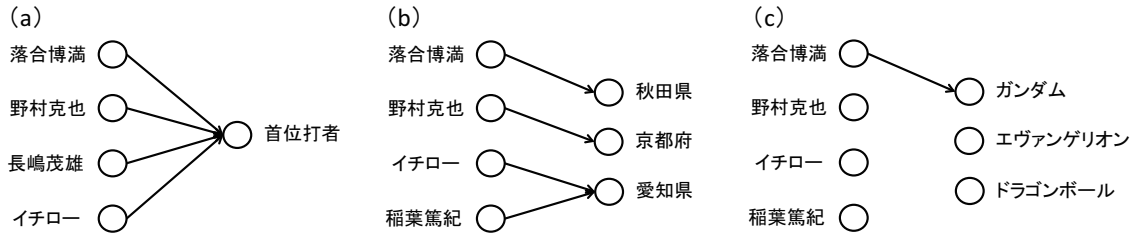

図 1　主題語と関連語およびそれらの同位語間の関係に基づく情報の構造

3.2　意外な情報の性質

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，共通の上位語を持つ語のことである．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，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づらいと言える．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（a））．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言える．

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））．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言える．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言える．

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．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言える．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．

意外度の具体的な計算については佃ら [5] を参照のこと．

3.3 実験

提案手法の有用性を示すために，実験を行った．実験には，Wikipedia [2]の見出し語から，人物名，地域名，製品名，施設名，組織名の５つのカテゴリそれぞれに対して 15 個，合計 75 個の主題語を選択して用いた．

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．

- 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．

そのために，２つの比較手法を用意した．１つ目は，同位語との関係と関連語の認知度を共に考慮しない手法である．この手法では，主題語と関連語を web で AND 検索し，ヒットカウントが少ないほど意外な関係であると見なす．２つ目は，同位語間の関係のみ考慮する手法である．この手法では，クエリと関連語の関係の典型度が低いほど意外度は高いと見なす．

評価の際は，５名の評価者に提案手法と比較手法で求められる各情報の意外度を４段階で評価してもらい，５名の意外度の平均値をその情報の意外度とした．評価の結果，いずれのカテゴリでも提案手法がもっとも高い精度を示し，クエリの同位語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うえで有用であることが示された．

4. 社会認知量に基づく語間の関係の典型度推定による意外な情報の発見

4.1 導入

本章では，オブジェクトと属性値を入力として与えたときに，オブジェクトと属性値の関係の認知度を推定するための手法を提案する．提案手法を用いることで，たとえばオブジェクトを「中国」，属性値を「ワイン」としたとき,「中国」と「ワイン」は実際の関連度は高いがその関係の認知度は低い，といったことが求められる．

4.2 手法

提案手法では，オブジェクト o と属性値 a が与えられたときに，o と a の関係の認知度を推定する．これにより，実際の関連度は高いが関係の認知度は低いオブジェクトと属性値のペアや，実際の関連度は低いが関係の認知度は高いオブジェクトと属性値のペアを意外な情報として提示する．語間の関連度を測ることを目的として，これまでに多数の手法が提案されてきたが，提案手法では，以下の２つの仮説のもと，これらの手法を拡張してオブジェクトと属性値の関係の認知度を推定する．

- オブジェクトの認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる．
- オブジェクトの多くの類似オブジェクトと属性値の関係の認知度が高（低）ければ，オブジェクトと属性値の関係の認知度は実際の関連度よりも高（低）くなる．

関係の認知度の具体的な計算については佃ら [4] を参照のこと．

4.3 実験

実験では Wikipedia の見出し語の中で「国」,「野菜」,「京都の観光地」,「電機メーカー」,「野球選手」の 5 カテゴリに関する 25 語の属性値を対象として，オブジェクトと属性値の関係の認知度推定に関する評価実験を行った．評価のために，クラウドソーシング [3]を用いて語間の関連度の社会的認知度および，情報の意外度を調べた．実験の結果，オブジェクトの認知度および，オブジェクトの類似オブジェクトと属性値の関係の認知度を考慮することで，既存手法に比べて有意に高い精度で関係の認知度を推定できることが明らかになった．

[文献]

[1] L.W. Barsalou. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, pp. 629–654, 1985.

[2] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. Investigating users' query formulations for cognitive search intents. In *Proc. of SIGIR 2014*, pp. 577–586, 2014.

[3] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Vol. 4 of *EMNLP '04*, pp. 404–411, 2004.

[4] 佃洸摂, 大島裕明, 加藤誠, 田中克己．オブジェクト間の意外な共通点の発見．第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014), 2014.

[5] 佃洸摂, 大島裕明, 山本光穂, 岩崎弘利, 田中克己．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．情報処理学会論文誌:データベース (TOD 61), Vol. 7, pp. 1–17, 2014.

[6] 佃洸摂, 中村聡史, 山本岳洋, 田中克己．レシピ検索のためのレシピの構造とその安定度を考慮した追加・削除可能な食材の推薦．電子情報通信学会和文論文誌 A 料理を取り巻く情報メディア技術特集号, Vol. J94-A, pp. 476–487, 2011.

佃 洸摂 Kosetsu TSUKUDA

産業技術総合研究所情報技術研究部門メディアインタラクション研究グループ研究員．2014 年京都大学大学院情報学研究博士後期課程修了．博士（情報学）．情報処理学会会員．

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

[3] 本実験ではランサーズ（http://www.lancers.jp/）を使用した．