Webオブジェクトの修飾語表現の信憑性検証

高橋　良平†　小山　聡‡　大島　裕明†　田中　克己†

京都大学大学院情報学研究科社会情報学専攻†　北海道大学大学院情報科学研究科複合情報学専攻‡

1. はじめに

オンライン広告や，ユーザがWebで発信するコンテンツでは，記述対象のオブジェクトをより魅力的に見せるために，様々な修飾語表現が用いられる．しかし，修飾語から想像される内容と実際の内容とが合致していない場合も多い．例えば，名前に「本格」を含む料理レシピは多数存在するが，かなり本格的なものから、それほど本格的でないものまで様々なレベルのものが存在する．

我々は，これまでオブジェクトの内容について書かれた部分から，修飾語の根拠となる語と，修飾語と相反する語を抽出する方式を提案してきた[1]．本論文では，これらの語を用いて，オブジェクトに付けられた修飾語がそのオブジェクトの実際の性質とどの程度合致しているか判定する方式を提案する．

オブジェクトに付けられた修飾語表現が適切であるかどうかを，オブジェクトの内容について書かれた部分に，その修飾語を強める語をどれだけ含むか，その修飾語を弱める語をどれだけ含まないかによって判定する．例えば，「本格」と書かれたカレーのレシピの場合，「ターメリック」や「コリアンダー」のような語を含んでいれば「本格」と合致しており，「ルー」や「レトルト」などの語を含んでいれば合致していないと考えられる．前者を「本格」を強める語，後者を「本格」を弱める語と呼び，これらの語を抽出する．しかし，例えば，「ルー」はカレーの中では非本格的であるが，他の料理では異なる可能性もある．そのため，語が修飾語をどれだけ強めているかは相対的なもので，比較する対象に依存する．

2. 合致度の計算手法

本研究では，オブジェクト集合と修飾語を与えたときに，名前に修飾語を含むオブジェクトについて，オブジェクトの内容と修飾語との合致度を計算する．

2.1 修飾語を相対的に強める語と弱める語の求め方

(1)入力されたオブジェクト集合 O を，オブジェクト o_i の名前 Name(o_i)に，入力された修飾語 m を含むオブジェクト集合Aと含まないオブジェクト集合$\bar{A}$の2つに分ける

$A = \{o_i \in O \mid Name(o_i) \ni m\}$

$\bar{A} = O - A$

(2)A 内で閾値 α 以上の割合で出現し，かつ$\bar{A}$内で閾値 β 以下の割合で出現する語をすべて取り出す

$C = \{t_j \mid \mathrm{DF}_A(t_j) \geq \alpha|A|, \mathrm{DF}_{\bar{A}}(t_j) \leq \beta|\bar{A}|\}$

(3)$t_j \in C$ を満たす各語 t_j に対して，A 内での出現頻度と$\bar{A}$内での出現頻度に関するカイ2乗値を求める

$$\chi^2_A(t_j) = \begin{cases} \sum_{i=1}^{2}\sum_{j=1}^{2} \dfrac{(w_{ij} - a_i b_j / S)^2}{a_i b_j / S} & (if \quad \dfrac{w_{11}}{a_1} > \dfrac{w_{21}}{a_2}) \\ -\sum_{i=1}^{2}\sum_{j=1}^{2} \dfrac{(w_{ij} - a_i b_j / S)^2}{a_i b_j / S} & (otherwise) \end{cases} \quad (1)$$

ここで，

$w_{11} = \mathrm{DF}_A(t_j), w_{12} = |A| - \mathrm{DF}_A(t_j), w_{21} = \mathrm{DF}_{\bar{A}}(t_j)$

$w_{22} = |\bar{A}| - \mathrm{DF}_{\bar{A}}(t_j), a_1 = |A|, a_2 = |\bar{A}|$

$b_1 = w_{11} + w_{21}, b_2 = w_{12} + w_{22}, S = b_1 + b_2$ である．

(4) $\chi^2_A(t_j)$が有意水準 p におけるカイ2乗値$\chi^2_0(p)$よりも大きい語を，m を 0 内で相対的に強める語として残す

$S_r = \{t_j \in C \mid \chi^2_A(t_j) > \chi^2_0(p)\}$

(5) A 内でβ以下の割合でしか現れず，かつ$\bar{A}$内でα以上の割合で出現する語をすべて取り出す

$D = \{t_j \mid \mathrm{DF}_A(t_j) \leq \beta|A|, \mathrm{DF}_{\bar{A}}(t_j) \geq \alpha|\bar{A}|\}$

(6) $t_j \in D$を満たす各語t_jに対して，A 内での出現頻度と$\bar{A}$内での出現頻度に関するカイ2乗値を式 (1) により求める

(7)カイ 2 乗値が$-\chi^2_0(p)$よりも小さい語を，m を相対的に弱める語として残す

$W_r = \{t_j \in D \mid \chi^2_A(t_j) < -\chi^2_0(p)\}$

2.2 修飾語を絶対的に強める語と弱める語の求め方

(1)データベースが持っているすべてのオブジェクト集合 O'を，名前に m を含むオブジェクト集合 A'と含まないオブジェクト集合$\bar{A}$'に分ける

$A' = \{o_i \in O' \mid Name(o_i) \ni m\}$

$\bar{A}' = O' - A'$

(2)前節の A に含まれるオブジェクトに出現する語を取り出す

$E = \{t_j \mid DF_A(t_j) > 0\}$

(3) $t_j \in E$を満たす各語t_jに対して，A'内での出現頻度とA'内での出現頻度に関するカイ 2 乗値を式(1)により求める

(4)カイ 2 乗値が $\chi^2_0(p)$ よりも大きい語を，m を絶対的に強める語として残す

$S_a = \{t_j \in E \mid \chi^2_{A'}(t_j) > \chi^2_0(p)\}$

(5)カイ 2 乗値が$-\chi^2_0(p)$よりも小さい語を，m を絶対的に強める語として残す

$W_a = \{t_j \in E \mid \chi^2_{A'}(t_j) < -\chi^2_0(p)\}$

Evaluating Credibility of the Modifiers in Web Objects

Ryohei Takahashi† Satoshi Oyama‡ Hiroaki Ohshima† Katsumi Tanaka†

†Department of Social Informatics, Graduate School of Informatics, Kyoto University

‡Division of Synergetic Information Science, Graduate School of Information Science and Technology, Hokkaido University

2.3 オブジェクトと修飾語の合致度

以上で得られた語をもとに，オブジェクトと修飾語の合致度を求める．各語の重みを考慮するため，本研究では以下の式により合致度を求める．

$$Score(o_i,m,O)=\sum_{t_j\in o_i,t_j\in S_r}\log\chi^2_A(t_j)-\sum_{t_j\in o_i,t_j\in W_r}\log(-\chi^2_A(t_j))+\sum_{t_j\in o_i,t_j\in S_a}\log\chi^2_{A'}(t_j)-\sum_{t_j\in o_i,t_j\in W_a}\log(-\chi^2_{A'}(t_j))$$

第 1 項は修飾語を相対的に強める語，第 2 項は相対的に弱める語，第 3 項は絶対的に強める語，第 4 項は絶対的に弱める語に関する値となっている．

3. 評価実験

2 節で使用したパラメータは[1]中の実験により α=0.05, β=1, p=0.1%とした．次に，2.3 節で得られた合致度が，人間の実際の感覚とどれほど合致しているのかを調べるために，評価実験を行った．実験では 6 名の被験者に，クックパッド[2]から取得した料理レシピを提示し，レシピが修飾語とどれほど合致しているかを 11 段階のスコアで評価してもらい，平均値が高い順にレシピを並べたものを，被験者による順位付けとした．なお，被験者にレシピを提示する際には，タイトルやコメントなどの部分は隠し，これらの情報によってスコアが影響されないようにした．評価は，（本格，カレー）（和風，ハンバーグ）（ヘルシー，ハンバーグ）（さっぱり，パスタ）の 4 つで行い，それぞれ 20 個のレシピを提示した．

提案手法の評価は，被験者による順位付けと提案手法のスコアによるランキングとの間の，スピアマンの順位相関係数を求めることによって行う．スピアマンの順位相関係数 ρ は以下の式で求められる．

$$\rho=1-\frac{6\sum D^2}{N^3-N} \qquad (2)$$

ここで，D は 2 つのランキングの順位の差であり，例えば片方のランキングで 1 位，もう一方のランキングで 5 位ならば，D=4 となる．また，N は順位付けされるデータの数であり，この場合は 20 個のレシピについて順位付けを行っているため，N=20 である．順位相関係数は，-1 から 1 の値をとり，1 に近いほど 2 つのランキングに正の相関が強いことを表す．

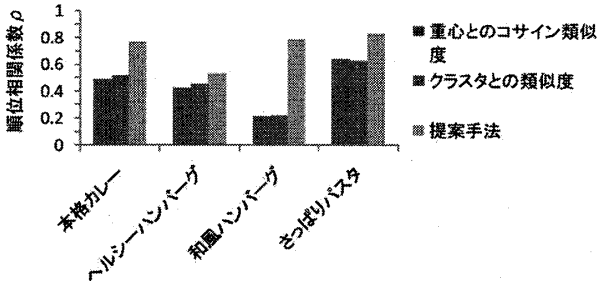

図1 提案手法とベースライン手法の順位相関係数の比較

3.1 ベースライン手法

提案手法と比較するために，ベースライン手法を 2 つ用意した．1 つ目の手法は，修飾語 m とオブジェクト集合 O が与えられたとき，オブジェクト集合 O 内で各語の idf 値を求め，tfidf 値をもとに各オブジェクトの特徴ベクトルを作成する．次に，名前に修飾語 m を含むオブジェクトの特徴ベクトルの重心 g を求める．そして，各オブジェクトの特徴ベクトルと重心 g とのコサイン類似度を，そのオブジェクトと修飾語の合致度とする．

2 つ目の手法は，各オブジェクトの特徴ベクトルを求めるところまでは 1 つ目と同じである．特徴ベクトルを求めた後，名前に修飾語 m を含むオブジェクトを，3 つのクラスタにクラスタリングする．そして，各オブジェクトと一番近いクラスタの重心とのコサイン類似度を，そのオブジェクトと修飾語の合致度とする．

3.2 順位相関係数の比較

提案手法と 2 つのベースライン手法について，被験者による順位付けとの順位相関係数を求めたものを図 1 に示す．図 1 より，全ての入力について，提案手法がベースライン手法よりも上回っていると言える．また，スピアマンの順位相関係数では，N=20 のとき，$\rho > 0.447$ であれば，危険率 5%で相関があると言えるが，全ての入力に対して順位相関係数がこの値を上回っている．以上により，提案手法で得られた合致度は，人間の実際の感覚とある程度合致していると言える．

4. まとめと今後の課題

本研究では，オブジェクトの名前に含まれる修飾語の信憑性を検証する手法を提案した．具体的には，修飾語を強める語と弱める語を抽出し，これらの語を用いてオブジェクトと修飾語の合致度を計算した．

実験は，料理レシピで行い，提案手法のランキングと人間の評価に相関があることを示した．

今後は，料理レシピ以外の例にも適用していきたいと考えている．

謝辞

本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，科学研究費補助金（課題番号 18049041，21700105，21700106），および，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」（研究代表者：田中克己）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]高橋良平, 小山聡, 田中克己, “オブジェクトに付けられた修飾語と内容の合致度判定”, 平成 21 年度情報処理学会関西支部支部大会, C-09, 2009.
[2] クックパッド, http://cookpad.com/