DEIM Forum 2010 C3-3

Web テキストと修飾表現との適合度判定手法

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　〒 606–8501 京都府京都市左京区吉田本町
†† 北海道大学大学院情報科学研究科複合情報学専攻　〒 060–0814 札幌市北区北 14 条西 9 丁目
E-mail: †{takahasi,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp, ††oyama@ist.hokudai.ac.jp

あらまし　オンライン広告や Web 上で誰でも発信できるレシピ情報などでは，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表現が用いられるが，中には誇張表現も存在する．本論文では，オブジェクトの内容について書かれた部分から，修飾表現と適合する語と，修飾表現と相反する語を抽出することで，Web テキストと修飾表現との適合度を判定する手法を提案する．

キーワード　修飾表現，料理レシピ，旅行ツアー，ランキング

Measuring Relevancy between Web Texts and Modifiers

Ryouhei TAKAHASHI†, Satoshi OYAMA††, Hiroaki OHSHIMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University,　Yoshida–honmachi, Sakyo, Kyoto 606–8501, Japan
†† Division of Synergetic Information Science, Graduate School of Information Science and Technology, Hokkaido University,　Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060–0814, Japan
E-mail: †{takahasi,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp, ††oyama@ist.hokudai.ac.jp

Abstract　To make online advertisements or user-generated content more attractive, people often use modifiers such as “authentic,” “impressive,” “special,” and so on. Some of these are exaggerations. That is, sometimes modifiers that are attached to Web entities do not represent the content appropriately. In this paper, we proposed a method to evaluate the truthfulness of modifiers attached to Web entity names by extracting relevant and conflicting terms from the content texts.

Key words　Modifier, recipe, package tour, ranking

1. はじめに

近年，インターネットの普及により，ユーザが Web 上にコンテンツを投稿することが容易になった．ユーザは自分が投稿したコンテンツに自由に名前を付けることができるが，それにより 2 つの問題が発生する．

1 つ目は，名前から想像される内容と実際の内容が適合していないコンテンツも多いということである．例えば，“本格カレー” という名前であるがそれほど本格的でない料理レシピ，“優雅” という語が名前に含まれているが格安のホテルに宿泊する旅行ツアーなどがある．

2 つ目は，名前の情報だけでは内容を十分に表せていないコンテンツも多いということである．例えば，本格的なレシピであっても，名前に “本格” という語を入れないことも多い．

これらの問題は，他のユーザがコンテンツを閲覧する際に特に問題となる．例えば，修飾表現を含むクエリで検索した場合，通常の検索エンジンはクエリに含まれる語を含むものを適合とみなすため，1 つ目の問題は適合率の低下，2 つ目の問題は再現率の低下の原因となる．また，情報の信憑性という観点では，1 つ目の問題は信憑性の低いコンテンツを閲覧してしまうことの原因となる．

そこで本研究では，Web ページに記述されたオブジェクトと修飾表現との適合度を判定する手法を提案する．我々は，オブジェクトの内容について書かれた部分から，修飾表現と適合する語，修飾表現と相反する語を抽出する手法を提案した [1]．本論文では，これらの語をどれだけ含むかによって修飾表現とオブジェクトの内容との適合度を判定する．これにより，修飾表現を含むクエリで検索した際に，名前に修飾表現を含むかどうかではなく，修飾表現と適合しているかによってオブジェクトをランキングすることができるようになる．また，本手法を名前に修飾表現を含むものに適用すれば，修飾表現の信憑性判定にも利用できる．実装は料理レシピと旅行ツアーの例で行い，それに対して評価実験を行った．

2. 関連研究

ユーザ投稿型コンテンツの品質に関する研究は多数行われている．例えば，Agichtein らは，QA サイトの中から品質の高い QA コンテンツを発見する手法を提案した [2]．また，Fiore らは，出会い系サイトのプロフィールの魅力について評価し，テキストが写真と同じくらいプロフィール全体の魅力に影響を与えることを示した [3]．

Web ページ中に書かれた記述の根拠を Web 上から抽出する研究もいくつか行われている．Lee らは，Web ページ中に実世界のイベントが記述されていた場合，そのイベントが実際に起こったことを示す根拠を Web 上から取得して提示する方式を提案した [4]．また，Murakami らは，ある Web 上の情報を支持する根拠と，その情報と矛盾する主張を支持する根拠を提示することで，情報の信憑性を分析するための支援を行っている [5]．

Kobayashi らは，ブランド名に便乗してつけられた名前を持つ商品に本当に価値があるかどうかを，評価属性に関する記述が Web 上に存在するかどうかで判定している [6]．名前から想像される内容と実際の内容が一致しているかを判定するという点で本研究と類似している．

修飾表現によって画像を検索する研究も行われている．Kato らは，抽象的な語をクエリとして画像検索する際に，その語を連想させる具体的な語集合を取得し，それをクエリに利用することで検索精度を向上させている [7]．抽象的な語を具体的な語に変換する点で本研究と類似している．Yusuf らは，日本語の学習者向けに，オノマトペによって写真を検索するシステムを作成した [8]．

修飾表現が内容を端的に表しているという点で，フォークソノミーに関する研究とも関係がある [9] [10]．しかし，タグは多くのユーザによって付けられているのに対して，オブジェクトの名前は 1 人の投稿者だけによって付けられているという点で異なる．

3. オブジェクトと修飾表現との適合度

3.1 修飾表現と適合する語と相反する語

本研究では，オブジェクトの内容についての記述の中から，修飾表現と適合する語と相反する語を抽出し，それらの語をどれだけ含むかによって修飾表現とオブジェクトの内容との適合度を求める．

例えば “和風ハンバーグ” という名前の料理レシピがあった場合，“和風” という修飾表現と，この料理レシピの内容がどれだけ適合しているかを判断することを考える．この料理レシピが，“大根おろし” や “ポン酢” という語を含んでいれば，含んでいない “ハンバーグ” のレシピよりも，より和風であると考えることができる．逆に，この料理レシピが “赤ワイン” や “マッシュルーム” を使っていれば，あまり和風ではないと考えられる．このとき，“大根おろし” や “ポン酢” は “和風” という修飾表現と適合する語，“赤ワイン” や “マッシュルーム” は “和風” という修飾表現と相反する語と見ることができる．すなわち，修飾表現と適合する語をより多く含むものほどその修飾表現とオブジェクトの内容がより適合しており，修飾表現と相反する語をより多く含むものほど，その修飾表現とオブジェクトの内容がより適合していないと判断することができると考えられる．

3.2 範囲の違いによる適合度の違い

修飾表現とオブジェクトの内容が適合しているかを判定する際，比較対象とするオブジェクトの範囲によってその結果は異なる．

例えば，“和風ハンバーグ” という名前のレシピの場合，ハンバーグは日本が起源の料理ではないため，料理全体で比較すれば，この料理レシピは和風とあまり適合していないことになる．しかし，この料理レシピが，“大根おろし” などを含んでいれば，ハンバーグの中では，和風との適合度は高くなると考えられる．

このように，オブジェクトの内容と修飾表現がどれくらい適合しているかというのは相対的なものであるため，同じオブジェクトと修飾表現間の適合度を求める場合でも，比較対象とするオブジェクトの範囲が異なれば，その結果も異なると考えられる．

また，適合する語と相反する語も同様に，範囲によって異なる．例えば，“本格” や “ヘルシー” といった修飾表現の場合，料理の種類によってその修飾表現と適合する語と相反する語というのは変化すると考えられる．本研究では，料理の種類などといったカテゴリに依存して適合・相反する語を，相対的に適合する語，相対的に相反する語と呼び，カテゴリと関係なく適合・相反する語を，絶対的に適合する語，絶対的に相反する語と呼ぶ．

3.3 訓練データの不要な手法の必要性

実際に修飾表現との適合度によってオブジェクトを並び替えることを考えると，修飾表現を含むあらゆるクエリに対して適用できる手法が必要であると考えられる．つまり，本問題では，ユーザがどのような修飾表現に対して本システムを利用するかは事前にはわからない．すべての修飾表現について，事前に訓練データを用意することは不可能であるため，教師付き学習の方法を使用することはできない．そのため，ユーザの入力に応じてその場で適合する語や相反する語を抽出するような手法が必要であると考えられる．

4. 問題の定式化

まず，各オブジェクト o_i は，名前に付けられた修飾表現の集合 (M_i)，オブジェクトが属するカテゴリ集合 (C_i)，オブジェクトの内容を表す語集合 (W_i) の 3 つ組で表されているとする．すなわち，

$$o_i = (M_i, C_i, W_i)$$
$$o_i \in O, M_i \subset M, C_i \subset C, W_i \subset W$$
$$M = \{m_1, m_2, \cdots\}, C = \{c_1, c_2, \cdots\}, W = \{w_1, w_2, \cdots\}$$

である．ここで，O はオブジェクトの全体集合，M は全ての修飾表現の集合で m_k は各修飾表現，C は全てのカテゴリの集合

で c_k は各カテゴリ，W は全ての語の集合で w_k は各語である．

本研究の最終的な目的は，カテゴリ c_k 内における，各オブジェクト o_i と修飾表現 m_j との適合度 $Relevancy(m_j, c_k, o_i)$ を求めることであるが，本研究では，この適合度を以下のように表す．

$$Relevancy(m_j, c_k, o_i) = \begin{cases} p(c_{jk}|o_i)p(c_{j0}|o_i) & (c_k \in C_i) \\ 0 & (c_k \notin C_i) \end{cases} \quad (1)$$

ここで，$p(c_{jk}|o_i)$ は o_i が c_k 内で m_j と適合している確率，$p(c_{j0}|o_i)$ は，o_i がオブジェクトの全体集合 O 内で m_j と適合している確率を表す．なお，$p(c_{jk}|o_i)$ だけでなく $p(c_{j0}|o_i)$ も用いているのは，相対的に適合する語と相反する語だけでなく，絶対的に適合する語と相反する語を使用するためである．この効果は 6 節の実験で示す．

また，修飾表現と適合する語集合 RW (Relevant Words) と，修飾表現と相反する語集合 CW (Conflicting Words) は以下のように定義する．

$$RW_{jk} = \{w|p(c_{jk}|w) > p(c_{jk}|\overline{w})\}$$
$$CW_{jk} = \{w|p(c_{jk}|w) < p(c_{jk}|\overline{w})\}$$

ここで，$p(c_{jk}|\overline{w})$ は，語 w を含まないオブジェクトが c_k 内で m_j と適合している確率を表す．

また，上記いずれでもない語集合を，修飾表現と無関係な語 IW (Irrelevant Words) とする．

$$IW_{jk} = \{w|p(c_{jk}|w) = p(c_{jk}|\overline{w})\} \quad (2)$$

4.1 適合度の計算

ベイズの定理により，

$$p(c_{jk}|o_i) = \frac{p(c_{jk})p(o_i|c_{jk})}{p(o_i)} \quad (3)$$

となる．

各オブジェクト o_i は内容を表す語集合 W_i で表されており，各語は独立に出現すると仮定すると，multi-variate Bernoulli model により，

$$p(o_i|c_{jk}) = \prod_{w \in W_i} p(w|c_{jk}) \prod_{w \notin W_i} (1 - p(w|c_{jk}))$$
$$p(o_i) = \prod_{w \in W_i} p(w) \prod_{w \notin W_i} (1 - p(w))$$

と書ける．これらを合わせると，

$$p(c_{jk}|o_i) = p(c_{jk}) \prod_{w \in W_i} \frac{p(w|c_{jk})}{p(w)} \prod_{w \notin W_i} \frac{(1 - p(w|c_{jk}))}{(1 - p(w))} \quad (4)$$

となる．

本研究では，

「ある語 w が修飾表現 m_j と適合する語であることと、w を含むオブジェクトの名前に m_j が含まれる割合が w を含まないオブジェクトの名前に m_j が含まれる割合よりも有意に高いことは同値である」

ということを仮定する．また，相反する語の場合はその逆である．

そこで，「語 w を含むオブジェクトの名前に修飾表現 m_j が含まれる割合と語 w を含まないオブジェクトの名前に修飾表現 m_j が含まれる割合は等しい」という帰無仮説 H_0 を立て，この帰無仮説 H_0 を棄却する語のうち，修飾表現を含むものに有意に多く表れる語を修飾表現と適合する語，修飾表現を含まないものに有意に多く表れる語を修飾表現と相反する語として抽出し，帰無仮説 H_0 が棄却されない語は修飾表現と無関係な語とする．

ここで (2) 式より，

$$\begin{aligned} & w \in IW_{jk} \\ \Leftrightarrow\ & p(c_{jk}|w) = p(c_{jk}|\overline{w}) \\ \Leftrightarrow\ & p(c_{jk}) = p(c_{jk}|w)p(w) + p(c_{jk}|\overline{w})p(\overline{w}) = p(c_{jk}|w) \\ \Leftrightarrow\ & \frac{p(w|c_{jk})}{p(w)} = \frac{p(c_{jk}|w)p(w)}{p(c_{jk})p(w)} = 1 \end{aligned}$$

となる．また，$p(c_{jk})$ は同一修飾表現・カテゴリ内では全てのオブジェクトについて正の値をとるため，適合度の順序には影響しない．以上により，

$$p(c_{jk}|o_i) \propto \prod_{w \in W_i \cap (RW_{jk} \cup CW_{jk})} \frac{p(w|c_{jk})}{p(w)} \times \prod_{w \notin W_i, w \in RW_{jk} \cup CW_{jk}} \frac{1 - p(w|c_{jk})}{1 - p(w)} \quad (5)$$

と書ける．この式は，修飾表現と無関係な語に関する値は計算しなくてよいことを示している．

4.2 確率の近似

3.3 節で述べたように，訓練データを用いる方法は使用できないため，$p(w|c_{jk})$ を得ることはできない．そこで，本節では，この確率を近似することを考える．(5) 式は，以下のように一般化できる．

$$p(c_{jk}|o_i) \propto \prod_{w \in W_i \cap (RW_{jk} \cup CW_{jk})} Score_{in}(w) \times \prod_{w \notin W_i, w \in RW_{jk} \cup CW_{jk}} Score_{not}(w)$$

$Score_{in}(w)$ は，語 w が c_k 内で m_j に適合している度合いと考えられる．$Score_{in}(w)$ は，語 w が修飾表現と適合すればするほど 1 より大きい大きな値を取り，修飾表現と相反すればするほど 1 未満の小さな正の値を取る関数であるというようにみなすことができる．

4.2.1 名前による近似

この方法では，「修飾表現 m_j を名前に含むオブジェクトの大部分は m_j と適合している」と仮定し，以下のように近似を行う．

$$Score_{in}(w) \approx \frac{p(w|m_j \in M_i)}{p(w)} \quad (6)$$
$$Score_{not}(w) \approx \frac{1 - p(w|m_j \in M_i)}{1 - p(w)}$$

表 1　カイ 2 乗検定の際の分割表

	語 w を含む	語 w を含まない	計
修飾表現 m_j を名前に含む	x_{11}	x_{12}	a_1
修飾表現 m_j を名前に含まない	x_{21}	x_{22}	a_2
計	b_1	b_2	S

なお，$p(w|m_j \in M_i)$ は，m_j を名前に含むオブジェクトが語 w を含む確率である．

4.2.2　カイ 2 乗値の使用

帰無仮説 H_0 を棄却するかどうか判定する際に使用したカイ 2 乗値も，語 w が c_k 内で m_j に適合している度合いと考えられため，これを $Score_{in}(w)$ の値として使用することも考えられる．すなわち，

$$Score_{in}(w) \approx \begin{cases} \chi^2(w) & (w \in RW_{jk}) \\ 1/\chi^2(w) & (w \in CW_{jk}) \end{cases} \quad (7)$$
$$Score_{not}(w) \approx 1$$

である．

5. 実 装 方 法

5.1　修飾表現と相対的に適合する語と相反する語の求め方

(1) 該当カテゴリ c_k に属するオブジェクトを，修飾表現 m_j を含むオブジェクト集合 O_{jk} と含まないオブジェクト集合 $\overline{O_{jk}}$ の 2 つに分ける

$$O_{jk} = \{o_i | m_j \in M_i, c_k \in C_i\}$$

$$\overline{O_{jk}} = \{o_i | m_j \notin M_i, c_k \in C_i\}$$

(2) 集合 O_{jk} と $\overline{O_{jk}}$ 内に出現する語をすべて取り出す．

$$W_{jk} = \{w | \mathrm{DF}_{O_{jk}}(w) + \mathrm{DF}_{\overline{O_{jk}}}(w) > 0\}$$

ここで $\mathrm{DF}_{O_{jk}}(w)$ は，$\{o_i | o_i \in O_{jk}, w \in W_i\}$ の要素数である．

(3)$w \in W_{jk}$ を満たす各語 w に対して，集合 O_{jk} 内での出現頻度と集合 $\overline{O_{jk}}$ 内での出現頻度に関するカイ 2 乗値を下式により求める

$$\chi^2_{O_{jk}}(w) = \begin{cases} \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(x_{ij} - a_i b_j/S)^2}{a_i b_j/S} & (\frac{x_{11}}{a_1} > \frac{x_{21}}{a_2}) \\ -\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(x_{ij} - a_i b_j/S)^2}{a_i b_j/S} & (\frac{x_{11}}{a_1} < \frac{x_{21}}{a_2}) \end{cases} \quad (8)$$

ここで，

$$x_{11} = \mathrm{DF}_{O_{jk}}(w), x_{12} = |O_{jk}| - \mathrm{DF}_{O_{jk}}(w),$$

$$x_{21} = \mathrm{DF}_{\overline{O_{jk}}}(w), x_{22} = |\overline{O_{jk}}| - \mathrm{DF}_{\overline{O_{jk}}}(w), a_1 = |O_{jk}|$$

$$a_2 = |\overline{O_{jk}}|, b_1 = x_{11} + x_{21}, b_2 = x_{12} + x_{22}, S = b_1 + b_2$$

である（表 1 参照）．

(4) $\chi^2_{O_{jk}}(w)$ が有意水準 p におけるカイ 2 乗値 $\chi^2_0(p)$ より大きい語を，修飾表現 m_j と c_k 内で相対的に適合する語として抽出する

$$RW_{jk} = \{w | w \in W_{jk}, \chi^2_{O_{jk}}(w) > \chi^2_0(p)\}$$

(5)$\chi^2_{O_{jk}}(w)$ が $-\chi^2_0(p)$ よりも小さい語を，修飾表現 m_j と相対的に相反する語として抽出する

$$CW_{jk} = \{w | w \in W_{jk}, \chi^2_{O_{jk}}(w) < -\chi^2_0(p)\}$$

(6) 手順 (4)(5) で得られた語について，$Score_{in}(w)$ と $Score_{not}(w)$ の値を計算し記憶しておく

5.2　修飾表現と絶対的に適合する語と相反する語の求め方

(1) オブジェクトの全体集合 O を，名前に修飾表現 m_j を含むオブジェクト集合 O_{j0} と含まないオブジェクト集合 $\overline{O_{j0}}$ の 2 つに分ける

$$O_{j0} = \{o_i | m_j \in M_i, o_i \in O\}$$

$$\overline{O_{j0}} = \{o_i | m_j \notin M_i, o_i \in O\}$$

(2) $w \in W_{jk}$ を満たす各語 w に対して，$\chi^2_{O_{j0}}(w)$ の値を式 (8) により求める

(3) カイ 2 乗値が有意水準 p におけるカイ 2 乗値 $\chi^2_0(p)$ よりも大きい語を，修飾表現 m_j と絶対的に適合する語として抽出する

$$RW_{j0} = \{w | w \in W_{jk}, \chi^2_{O_{j0}}(w) > \chi^2_0(p)\}$$

(4) カイ 2 乗値が $-\chi^2_0(p)$ よりも小さい語を，修飾表現 m_j と絶対的に相反する語として抽出する

$$CW_{j0} = \{w | w \in W_{jk}, \chi^2_{O_{j0}}(w) < -\chi^2_0(p)\}$$

(5) 手順 (3)(4) で得られた語について，$Score_{in}(w)$ と $Score_{not}(w)$ の値を計算し記憶しておく

5.3　共起頻度の高い語の除去

(4) 式のように変形できるのは,「オブジェクトの内容についての記述中に，各語は独立に出現する」ことを仮定したときであった．しかし，以上で得られた修飾表現と適合する語，相反する語には，互いに独立でない語も含まれる．例えば，ハンバーグ内で和風と適合する語として，“大根おろし” と “大根” の 2 語が得られたとき，実際は 1 つのものを表しているにも関わらず，両方の語に関するスコアを用いていることになる．すなわち，これらの 2 語を同時に含むレシピが不当に高く評価されてしまうこととなる．

そこで，共起度の高い語を同一のものを指し示しているとみなして，カイ 2 乗値が小さい方の語を取り除く．本研究では，共起度を図る指標として，以下の式で表される Jaccard 係数を使用する．

$$Jaccard(w_1, w_2) = \frac{\mathrm{DF}_{RW_{jk}}(w_1 \cap w_2)}{\mathrm{DF}_{RW_{jk}}(w_1 \cup w_2)} \quad (9)$$

分子は集合 RW_{jk} 内で語 w_1 と w_2 をともに含むオブジェクトの数，分母は語 w_1 と w_2 の少なくとも一方を含むオブジェク

トの数である．

そして，この Jaccard 係数の値が，閾値 θ 以上であった場合，カイ 2 乗値が小さい方の語を RW_{jk} から取り除く．具体的には以下のように行う．

(1)RW_{jk} 内で未チェックの語のうち，$\chi^2_{O_{jk}}(w)$ の値が最も小さい語を選び，w_1 とする

(2)RW_{jk} 内の語で，$\chi^2_{O_{jk}}(w_1) \leqq \chi^2_{O_{jk}}(w_2)$ となる語 w_2 について，$Jaccard(w_1, w_2)$ を計算する

(3)$Jaccard(w_1, w_2) \geqq \theta$ となる語 w_2 が 1 つでもあれば w_1 を RW_{jk} から取り除く

(4)$Jaccard(w_1, w_2) \geqq \theta$ となる語 w_2 が 1 つもなければ，語 w_1 をチェック済みの語とし，(1) に戻る

同様のことを，$CW_{jk}, RW_{j0}, CW_{j0}$ の 3 つに対しても行い，語の除去を行う．

6. 実　　験

6.1 修飾表現による料理レシピのランキング

本実験は，投稿型レシピサイト “クックパッド” [11] から取得した約 16,000 件のレシピについて行った．“本格　カレー” などのクエリで検索を行った際を想定し，修飾表現との適合度に基づいて料理レシピのランキングを行う．

料理レシピに付けられた名前の末尾の単語を，その料理レシピが属するカテゴリとみなした．すなわち，“本格カレー” という名前の料理レシピは “カレー” カテゴリであり，“カレーうどん” という名前の料理レシピは “うどん” カテゴリとみなしている．

実験では，まず 6 つの料理名（カレー、ハンバーグ、パスタ、オムレツ、やきそば、炒飯）と 4 つの修飾表現（本格、ヘルシー、和風、さっぱり）を用意した．次に，その料理名と修飾表現を組み合わせ，24 個のクエリを作成した．そのうち，該当のカテゴリ内で名前に該当の修飾表現を含むレシピが 10 件以上存在したクエリのみを使用した．その結果，表 2 に示す 17 個のクエリを使用した．

また，形態素解析には MeCab [12] を使用し，各レシピの材料・作り方の部分に出現する名詞，動詞のみを，オブジェクトの内容を表す語集合 W として使用した．また，実装上の都合により，絶対的に適合する語と相反する語を求める際には，クックパッド上の検索エンジンによって得られた検索結果の数を，その語の出現頻度とした．

参考のために，得られた適合する語・相反する語の例を，表 4 に示す．

6.2 修飾表現による旅行ツアーのランキング

本実験は，旅行ツアーサイト “Yahoo!トラベル” [13] の “海外ツアー” のアジア地域へのツアーから取得した旅行ツアーについて行った．

なお，出発地は異なるが，他の部分は全く同一である旅行ツアーが複数投稿されるという場合があるため，同じ業者が似たような旅行ツアーを複数投稿している場合は，重複と見なして 1 つのみを使用した．重複を削除した結果，使用した旅行ツアーの総数は約 13,000 件となった．

表 2　料理レシピの実験で使用したクエリ

料理名	本格	ヘルシー	和風	さっぱり
カレー	√	√	√	
ハンバーグ		√	√	√
パスタ	√	√	√	√
オムレツ			√	
やきそば		√	√	√
炒飯		√	√	√

表 3　旅行ツアーの実験で使用したクエリ

地域	満喫	便利	優雅	癒し	感動
中国	√	√	√		√
韓国	√	√	√		
バリ		√	√	√	
タイ	√	√	√		
ベトナム	√	√	√	√	
台湾	√	√	√		

また，カテゴリは各旅行ツアーに付けられている地域名を使用した．

クエリを作成する際には，6 つの地域（中国，韓国，バリ，タイ，ベトナム，台湾）と 5 つの修飾表現（満喫，便利，優雅，癒し，感動）を用意し，それらを組み合わせて 30 個のクエリを作成した．そのうち，該当のカテゴリ内で名前に該当の修飾表現を含む旅行ツアーが 5 件以上存在したクエリ計 20 個を使用した（表 3）．

また，各ツアーの特徴やスケジュールについて書かれた部分から名詞のみを抽出し，オブジェクトの内容を表す語集合 W として使用した．

参考のために，得られた適合する語・相反する語の例を，表 5 に示す．

6.3 評 価 実 験

提案手法で得られる適合度が，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は，正解となる人間によるランキングを作成し，そのランキングと比較する必要がある．そ

表 4　料理レシピの実験で得られた適合する語・相反する語の例

和風　ハンバーグ		本格　カレー	
適合する語	相反する語	適合する語	相反する語
大根おろし	ソース	ターメリック	ルー
みりん	ウスターソース	パニール	豚
醤油	トマト	コリアンダー	
だし汁	赤ワイン	クミンシード	
ポン酢	チーズ	ガラムマサラ	

表 5　旅行ツアーの実験で得られた適合する語・相反する語の例

韓国　満喫		中国　感動	
適合する語	相反する語	適合する語	相反する語
プルコギ		万里	
宗廟		天壇	
サムギョプサル		聚徳	
サムゲタン		遺産	
遺産		ダック	

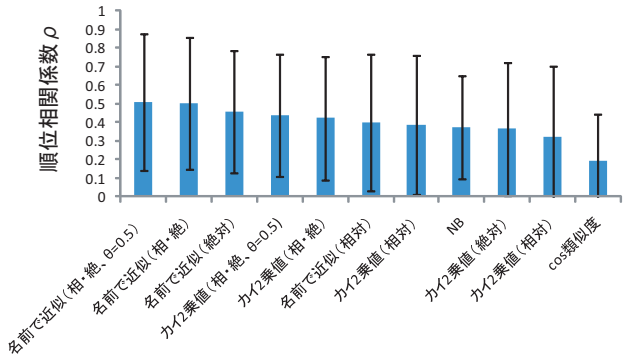

図 1　料理レシピでの順位相関係数の平均値比較

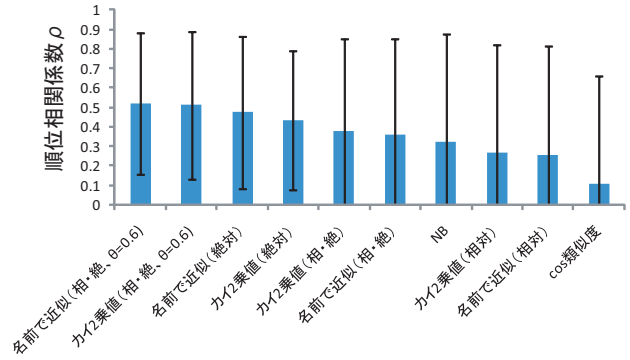

図 2　旅行ツアーでの順位相関係数の平均値比較

のために評価実験を行った．

まず，クエリごとに，レシピの場合は 10 件，旅行ツアーの場合は 5 件のページを各被験者に提示する．次に，各ページの内容が修飾表現とどれほど適合しているかを 7 段階のスコアで評価してもらった．このスコアは数字が大きいほど修飾表現と内容が適合していることを表す．そして，各ページについて，被験者全員の付けたスコアの平均値をとり，その平均値が高い順にページを並べたものを，被験者による順位付けとし，これを正解とする．

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．

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N} \tag{10}$$

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，レシピの場合は N=10，旅行ツアーの場合は N=5 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．

6.4　ベースライン手法

提案手法と比較するために，簡単なベースライン手法を 2 つ用意した．

1 つ目の手法は，修飾表現 m_j とカテゴリ c_k が与えられたとき，そのカテゴリ内で各語の 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求める．次に，名前に修飾表現 m_j を含むオブジェクト集合の特徴ベクトルの重心ベクトル 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ベクトル g とのコサイン類似度を，そのオブジェクトと修飾表現の合致度とする．

2 つ目の手法は，ナイーブベイズによる確率を修飾表現との合致度とするもので，具体的には (4) 式を用いる．各確率は，修飾表現 m_j を名前に含むものを修飾表現 m_j と適合しているとみなして近似する．4.2.1 節の方法との違いは，この手法では全ての語について確率を計算するのに対して，4.2.1 節の方法では，適合する語と相反する語のみについて確率を計算する点である．

6.5　順位相関係数の比較

順位相関係数の比較は，前節で説明した 2 つの手法（cos 類似度，NB）以外に，4.2.1 節と 4.2.2 節で述べた二つのの方法（名前で近似，カイ 2 乗値）をカテゴリ内だけに適用したもの（相対），オブジェクトの全体集合だけに適用したもの（絶対），その二つを組み合わせた手法（相・絶）の手法について行った．ただしいずれの手法も 5.3 節で述べた共起頻度の高い語の除去は行っていない．それぞれの手法の順位相関係数とその標準偏差を，料理レシピと旅行ツアーについてまとめたものを図 1 と図 2 に示す．なお，提案手法については，5.3 節で説明した Jaccard 係数の閾値を変化させたものも行い，そのうち一番結果が良かった閾値について載せている．また，カイ 2 乗検定を行う際の有意水準は，$p = 0.001$ とした．

これらの図より，料理レシピ・旅行ツアーのいずれの場合においても，提案手法が最も順位相関係数が高くなっている．すなわち，単純にカテゴリ内で得られる語を使うだけの方法や，オブジェクト集合全体で得られる語だけを使用するよりも，これらを組み合わせて使う方が，より人間の感覚に近いことが分かる．

また，図 3 に，閾値 θ を 0.1 間隔で変化させたときの，順位相関係数の平均値の変化を示す．なお，このグラフは，料理レシピと旅行ツアーの 37 個の平均をとったものである．その結果，$\theta = 0.6$ のとき，順位相関係数の平均値が最も高くなることが分かった．つまり，共起頻度がある程度高い語を取り除くことで，似たような内容を表す語が削除され，より精度が上がることが分かる．

7.　考　　察

今回提案した手法は，名前に付けられた修飾表現を使用している．そのため，修飾表現を含むオブジェクトの数がどれほどあるかによって精度が変わるのではないかと予想される．図 4 は，カテゴリ内で修飾表現を名前に含むオブジェクトの数と順位相関係数との関係を示したものである．例えば，n=10 の点は，カテゴリ内で修飾表現を名前に含むオブジェクトが 10 個以上あったクエリの場合における，順位相関係数の平均値を示している．

この図より，修飾表現を含むオブジェクトの数が増えるほど，順位相関係数が高くなることが分かる．具体的には，修飾表現

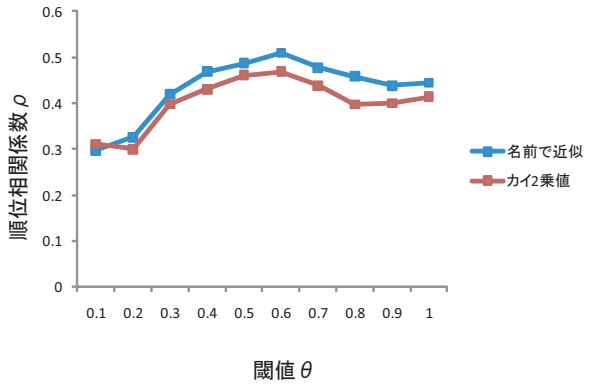

図 3　閾値 θ と順位相関係数の平均値の関係

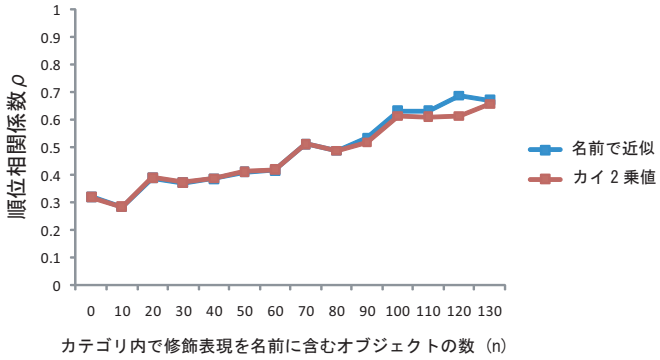

図 4　カテゴリ内で修飾表現を名前に含むオブジェクトの数と順位相関係数の関係

を含むオブジェクト数が 10 個程度しかない場合には順位相関係数は 0.3 程度しかないが，100 を超えると 0.6 を超えるようになる．

修飾表現を名前に含むオブジェクトの数が少ない場合に精度が低い理由の 1 つ目は，修飾表現と適合しないにも関わらず修飾表現が付けられているものが 1 つでもあると，結果に大きな影響を及ぼすことが考えられる．例えば，旅行ツアーの実験では，あるカテゴリ内である修飾表現を含むオブジェクトの半分以上が，ある 1 つの業者によって投稿されていたということがあった．この場合，この業者が好んで使う語が適合する語として抽出されてしまうことになり，精度が悪くなってしまうと考えられる．

2 つ目は，修飾表現があまり付けられないということは，修飾表現と適合するオブジェクトであるにも関わらず，その名前に修飾表現が含まれていないことが多いということを意味するからである．旅行ツアーの実験中の “中国　感動” の例では，この 2 語を含む旅行ツアーが全部で 14 件しか存在しない．被験者によるランキングでは，“桂林” や “九寨溝” に行くツアーが感動的であると評価されたが，これらの語を含むツアーの中で，名前に “感動” を含むものはそれぞれ 1 件ずつしか存在しなかった．そのため，これらの語を “感動” と適合する語と判定できなかった．

8.　まとめと今後の課題

本研究では，Web テキストと修飾表現との適合度を判定する手法を提案した．具体的には，修飾表現を名前に含むオブジェクトの大部分は修飾表現と適合しており，修飾表現を名前に含まないオブジェクトの大部分は修飾表現と適合していないと仮定し，カイ 2 乗検定により，修飾表現と適合する語，相反する語を求め，それらの語をもとに適合度を計算した．

実験は，料理レシピと旅行ツアーについて行い，提案手法によるランキングと被験者によるランキングの順位相関係数を求めた．その結果，カテゴリ内での適合度（相対的な比較）と，オブジェクトの全体集合内での比較（絶対的な比較）の両方を用いる提案手法が，最も人間の感覚と一致していることがわかった．

考察で述べたように，本研究で提案した手法は，修飾表現が付けられたオブジェクトの大部分が修飾表現と相反している場合や，修飾表現を名前に含まないオブジェクトの大部分が修飾表現と適合している場合などには，適用できないと考えられる．今後は，このような場合にも適用できるような方法を考えていきたい．

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，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，課題番号：18049041），および，文部科学省科学研究費補助金若手研究（B）「オンデマンド利用を目的とする Web からの知識発見に関する研究」（研究代表者：大島裕明，課題番号：21700105），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．

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表現と内容の合致度判定,” 平成 21 年度情報処理学会関西支部大会.
[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, “Finding high-quality content in social media,” WSDM 2008, pp.183-194.
[3] A. T. Fiore, L. S. Taylor, G.A. Mendelsohn, and M. Hearst, “Assessing attractiveness in online dating profiles,” CHI 2008, pp.797-806.
[4] R. Lee, D. Kitayama, and K. Sumiya, “Web-based evidence excavation to explore the authenticity of local events,” WICOW 2008, pp.63-66.
[5] K. Murakami, E. Nichols, S. Matsuyoshi, A. Sumida, S. Masuda, K. Inui, and Y. Matsumoto, “Statement Map: Assisting Information Credibility Analysis by Visualizing Arguments,” WICOW 2009, pp.43-50.
[6] T. Kobayashi, H. Ohshima, S. Oyama, and K. Tanaka, “Evaluating brand value on the Web,” WICOW 2009, pp.67-74.
[7] M. Kato, H. Ohshima, S. Oyama, and K. Tanaka, “Can Social Tagging Improve Web Image Search?,” WISE 2008, pp.235-249.
[8] M Yusuf, C Asaga and C Watanabe, “Onomatopeta!:

Developing a Japanese Onomatopoeia Learning-Support System Utilizing Native Speakers Cooperation," Web Intelligence/IAT Workshops 2008, pp.173-177.
[9] A. Hotho, R. Jäschke, C. Scgnutz, and G. Stumme, "Information Retrieval in Folksonomies: Search and Ranking," ECWS 2006, pp.411-426.
[10] S. Bao, G.Xue, X. Wu, Y Yu, B. Fei, Z. Su, "Optimizing web search using social annotations," WWW 2007, pp.501-510.
[11] クックパッド, http://cookpad.com/
[12] Mecab, http://mecab.sourceforge.net/
[13] Yahoo!トラベル, http://travel.yahoo.co.jp/