C-09

オブジェクトに付けられた修飾語と内容の合致度判定

Measuring Correlation between Modifiers for the Names of Objects and their Contents

高橋　良平†　　小山　聡†　　田中　克己†
Ryohei Takahashi Satoshi Oyama Katsumi Tanaka

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，名前から想像される内容と実際の内容が合致していないコンテンツも多い．例えば，“本格カレー”という名前であるがそれほど本格でない料理レシピ，“満腹ツアー”という名前であるがそれほど料理の量が多くないツアー，“わかりやすい解説”と書かれているがそれほどわかりやすくないページなどがある．

このようなコンテンツを閲覧した場合，閲覧者がその分野に関する知識がある程度ある場合には，修飾語と内容があまり合致していないことが分かるため，それほど問題ではない．しかし，閲覧者が他にどのようなコンテンツがあるのかを知らない場合には，そのコンテンツが修飾語と合致していると信じてしまい，より修飾語と合致しているコンテンツを見る機会が失われてしまう可能性がある．

本研究では，コンテンツに記述されたオブジェクトの名前の一部に修飾語を含む場合に，その修飾語から連想される内容とオブジェクトの内容がどれほど合致しているかを判定する．具体的には，その修飾語の根拠となる語と，修飾語と相反する語を求め，その語をどれほど含んでいるかにより判定する．

以下では料理レシピを例に説明する．

2.　動機

ユーザは自分の投稿したコンテンツを多くの人に見てもらうために，記述されたオブジェクトの名前に魅力的な言葉を付けることがある．例えばカレーのレシピを投稿する場合でも，単純に“カレー”という名前で投稿されることはほとんどなく，投稿したレシピが本格的なインドカレーであれば“本格インドカレー”という名前に，ヘルシーなドライカレーであれば“ヘルシードライカレー”という名前にするといったように，“本格”“ヘルシー”などの修飾語を名前に付け加える．このように，投稿したオブジェクトに魅力的な性質が含まれていれば，その性質を表す語を修飾語として名前に付け加えることが多い．

しかし，名前にこのような修飾語を含むオブジェクトであっても，実際には修飾語とオブジェクトの内容があまり合致していない場合もある．例えば，名前に“本格”を含む料理レシピであっても，それほど本格的でないものも多数存在する．これには，単に目立たせるためにその修飾語を付与した，自分の投稿したレシピを客観的に見ていなかった，などの理由が考えられる．

一方，投稿されたコンテンツを閲覧したい場合には，キーワードをクエリとして検索を行うことが多い．その際，オブジェクト名やオブジェクトの分類名などの名詞だけをクエリにするのではなく，クエリに修飾語を付け加えて検索することもある．例えば，本格的なカレーのレシピが欲しい場合は，“カレー　本格”というクエリで検索を行う．

このクエリで，実際に検索を行うと，かなり本格的な料理レシピも検索結果に現れるが，それほど本格的でないものも検索結果に現れる．これは，実際にはそれほど本格的でないものにも“本格”という修飾語が付いており，検索エンジンは単純にページ中に“カレー”と“本格”の２語が含まれていれば適合していると判断して検索結果として出力するためである．

しかし，このクエリの場合，ユーザは本格的な料理レシピを求めている．そのため，各レシピがどのくらい本格的であるのかを検索エンジンが判断し，それほど本格的でないものは検索結果から除外し，本格的な順にランキングして表示する方がより有用であると考えられる．

このようなことを実現するために，本研究では，オブジェクトに修飾語が付けられている場合に，その修飾語とオブジェクトの内容がどれだけ合致しているかを判定する手法を提案する．

3．　関連研究

Web ページに書かれた記述の根拠を Web 上から抽出する研究はいくつか行われている．[1]では，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 上から取得し提示する．また，[2]では，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している．

[3]では，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．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．

[4]では，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．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．

4．　修飾語とオブジェクトの内容の合致度

4.1　修飾語を強める語と弱める語

本研究では，修飾語とオブジェクトの内容の合致度を求めるために，オブジェクト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んでいないかによって判定する．

†京都大学大学院情報学研究科社会情報学専攻

例えば“本格カレー”という名前の料理レシピがあったとき，“本格”という語とこの料理レシピがどれだけ合致しているかを判断することを考える．この料理レシピが，“ターメリック”や“クミンシード”という語を含んでいれば，含んでいない“カレー”のレシピよりも，本格的であると考えることができる．逆に，この料理レシピが“ルー”を使っていれば，あまり本格的でないと考えられる．このとき，“ターメリック”や“クミンシード”は“本格”という修飾語を強める語，“ルー”は“本格”という修飾語を弱める語と見ることができる．すなわち，修飾語を強める語をより多く含むものほどその修飾語とオブジェクトの内容がより合致しており，修飾語を弱める語をより多く含むものほどその修飾語とオブジェクトの内容がより一致していないと判断することができると考えられる．

4.2 範囲の違いによる合致度の違い

修飾語とオブジェクトの内容が合致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，“本格インドカレー”という名前のレシピがあったとする．このレシピの名前は以下のように二通りに解釈できる．

一つ目は，“インドカレーの中でも本格的なレシピである”という意味である．つまり，一般的なインドカレーのレシピよりもこのレシピはスパイスを多数含んでいるなどの理由で本格的であるというように解釈できる．

二つ目は，“インドカレーだから本格的”という意味である． 一般的に，インドカレーは普通のカレーよりも本格的である．そこで，その本格さを強調するために本格という語を付けたと考えることができる．

一つ目の解釈の場合では，“カレー”のレシピ集合内で合致度を計算しても，“インドカレー”のレシピ集合内で合致度を計算しても，高い値になるはずである．

一方，二つ目の解釈の場合，カレーの中では本格的であるため，カレーのレシピ集合の中で合致度を計算すれば，比較的高い値となるであろう．しかし，インドカレーの中ではそれほど本格的ではないため，インドカレーのレシピ集合の中で合致度を計算すればそれほど高くない値になると考えられる．

このように，オブジェクトと修飾語がどれくらい合致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語間の合致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．

5． 提案手法

5.1 入力と出力

本研究では，入力として，修飾語 m とオブジェクト集合 O の 2 つを与えたとき，オブジェクト集合 O に含まれるオブジェクト o_i のうち，Name(o_i)⊃m（ただし Name(o_i)は o_i に付けられた名前を単語ごとに分割した単語集合）を満たすものに対し，オブジェクト o_i の内容と修飾語 m の合致度 Score(o_i,m,O)を出力することを目的とする． なお，比較対象とするオブジェクトの範囲は与えられたオブジェクト集合である．

5.2 提案手法の流れ

提案手法の流れは以下のようになっている．

(1)オブジェクト集合 O の中で, 修飾語 m を強める語と弱める語を求める

(2)それらの語を該当オブジェクトがいくつ含むかによって合致度を計算する

なお，本論文では，(1)についてのみ言及する．

5.3 修飾語を強める語の求め方

本研究では修飾語を強める語を以下の 2 つの種類に分けて求める．

一つ目は，比較対象とするオブジェクト集合の中で，名前に対象の修飾語を含むオブジェクトの中には頻繁に現れるが，名前に対象の修飾語を含まないオブジェクトの中にはあまり現れない語である．例えば，“カレー”のレシピ集合の中で“本格”を強める語を求める場合，名前に“本格”と“カレー”の両方を含むレシピ集合内には，“ターメリック”などの語が頻繁に現れるが，“本格”を含まない“カレー”のレシピにはあまり現れないということが考えられる．逆に，“鶏肉”のように，“本格”を強める語でも弱める語でもない場合は, 2 つの間で出現頻度にそれほど差がないと考えられる．

二つ目は，対象とするオブジェクト集合内にはほとんど現れないため，修飾語を強めているのかどうかは判断できないが，他の分野を見ることで修飾語を強めていると判断できる語である．例えば，“ヘルシーカレー”に“長いも”が入っていても，“長いも”が入っているカレーのレシピはほとんどないため，上の方法ではカイ 2 乗検定がうまく働かず，“長いも”が“ヘルシー”を強めているかどうかは判定することができない．このような場合，カレーの中だけで判断するのではなく，レシピ全体を見て，“長いも”がヘルシーであるかを判断し，ヘルシーだということが分かれば，カレーの集合内でも“長いも”がヘルシーという修飾語を強めていると推定できる．

一つ目の種類の語は，カレーの中で相対的に本格的な語を求めていることになる．一方，二つ目の種類の語は，絶対的にヘルシーな語を求めていることになる．

これらの語は以下のようなアルゴリズムで求める．

【アルゴリズム】

(1)オブジェクト集合 O を, 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 B の二つに分ける

$$A = \{o_i \mid Name(o_i) \ni m, o_i \in O\}$$
$$B = \{o_i \mid Name(o_i) \not\ni m, o_i \in O\}$$

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

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

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

$$\chi^2_{AB}(t_j) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(w_{ij} - a_i b_j / S)^2}{a_i b_j / S} \qquad (1)$$

ここで，
$w_{11} = \mathrm{DF}_A(t_j), w_{12} = |A| - \mathrm{DF}_A(t_j),$
$w_{21} = \mathrm{DF}_B(t_j), w_{22} = |B| - \mathrm{DF}_B(t_j), a_1 = |A|, a_2 = |B|$
$b_1 = w_{11} + w_{21}, b_2 = w_{12} + w_{22}, S = b_1 + b_2$　である．
また，$w_{11}/a_1 < w_{21}/a_2$のとき，$\chi^2_{AB}(t_j)$の符号を負にして返すものとする．

(4) $\chi^2_{AB}(t_j)$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，入力された修飾語 m を O 内で相対的に強める語として残す

$$S_1 = \{t_j \mid t_j \in C, \chi^2_{AB}(t_j) > \chi^2_0(p)\}$$

(5) 集合 A 内でα未満の割合でしか現れず，かつ集合 B 内でγ以下の割合でしか現れない語をすべて取り出す

$$D = \{t_j \mid \mathrm{DF}_A(t_j) < \alpha|A|, \mathrm{DF}_B(t_j) \le \gamma|B|\}$$

(6)検索エンジンが持っているすべてのオブジェクト集合O'を，名前に修飾語 m を含むものの集合 A'と含まないもの集合 B'に分ける

$$A' = \{o_i \mid Name(o_i) \ni m, o_i \in O'\}$$
$$B' = \{o_i \mid Name(o_i) \not\ni m, o_i \in O'\}$$

(7) $t_j \in$D を満たす各語 t_j に対して，集合 A'内での出現頻度と集合 B'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

(8) カイ 2 乗値が有意水準 p におけるカイ二乗値$\chi^2_0(p)$よりも大きい語を，入力された修飾語 m を絶対的に強める語として残す

$$S_2 = \{t_j \mid t_j \in D, \chi^2_{A'B'}(t_j) > \chi^2_0(p)\}$$

5.4　修飾語を弱める語の求め方

修飾語を弱める語の求め方も先ほどとほぼ同じであり，以下のアルゴリズムで求める．

【アルゴリズム】
(1)オブジェクト集合 O を，修飾語 m を含むものの集合 A と含まないものの集合 B の二つに分ける

(2)集合 A 内で閾値 δ 以下の割合でしか出現せず，かつ集合 B 内で閾値 ε 以上の割合で語をすべて取り出す

$$E = \{t_j \mid \mathrm{DF}_A(t_j) \le \delta|A|, \mathrm{DF}_B(t_j) \ge \epsilon|B|\}$$

(3) $t_j \in$E を満たす各語 t_j に対して，集合 A 内での出現頻度と集合 B 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$-\chi^2_0(p)$よりも小さい語を，入力された修飾語 m を弱める語として残す

$$W = \{t_j \mid t_j \in E, \chi^2_{AB}(t_j) < -\chi^2_0(p)\}$$

6．　実験

実験は，投稿型レシピサイト“クックパッド”[5]から取得した約 16,000 件のレシピについて行った．なお，形態素解析には MeCab[6]を使用し，各レシピの名詞，動詞のみを取り出し使用した．

実験では，（“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 つで修飾語を強める語と弱める語を求めた．なお，（“本格”，“カレー”）は，“カレー”のオブジェクト集合内で，“本格”という語を強める語と弱める語を求めたという意味である．

まずは，修飾語を比較範囲内で相対的に強める語を求めた．まず，有意水準 p=0.1%，閾値β=1 と固定し，閾値αを 0.05 から 1 まで 0.05 間隔で変化させることで精度の変化を調べた．なお，精度は 4 つの入力に対してそれぞれ F 値を求め，その平均値を使用した．F 値は以下の式により求める．

$$\mathrm{F} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

得られた語が正解であるかどうかは，主観で判断した．また，再現率を求める際には，この実験中に得られた正解の語を全正解数として計算しているため，本来の F 値よりも高い値となっている．

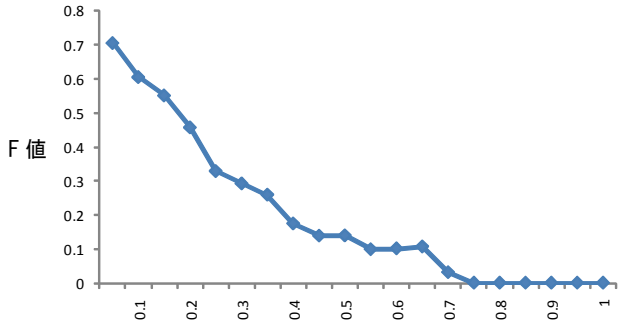

図1　閾値α とF値の平均値の変化

図 1 より，α=0.05 の時に F 値の平均が 0.707 となり最大となった．α=0.04 とすると F 値の平均が 0.704 と下がり，得られた正解の数は増加しなかった．そのためα=0.05 とした．次に，βを 0.05 から 1 まで 0.05 間隔で変化させて F 値の平均値の変化を調べた．

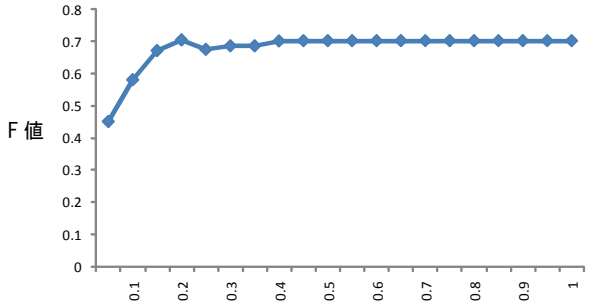

図2　閾値β とF値の平均値の変化

図２より，$\beta \geqq 0.2$ であれば閾値にあまり依存しないと考えられる．精度がほぼ同じならば，より多くの語が得られる方が良いと判断し，$\beta=1$ とした．最後に，有意水準を p=0.1%，0.01%の二つで比較したところ，p=0.1%の時の F 値の平均が 0.702，p=0.01%の時の F 値の平均が 0.728 となり，p=0.01%の時の方が良い結果となった． また，このときの適合率の平均値は，62.4%であった．

次に，修飾語を弱める語を求めた．提案手法では閾値 δ と ε が存在したが，どのような値にしても得られる不正解の数はそれほど変わらないため，より正解が得られるように $\delta=1$，$\varepsilon=0.1$ に設定した．ただし，得られる数が少なかったため，有意水準は 0.1%とした．

表 1 と表 2 に“カレー”の中で“本格”を強める語と弱める語として得られたものの例を示す．

表1 “カレー”の中で“本格”を強める語と弱める語の例

強める語		弱める語	
正解	不正解	正解	不正解
ターメリック	とり	ルー	豚肉
スパイス	ニンニク		
プレーンヨグルト	おろす		
コリアンダー	ホール		
煮込む	トマト		

表2 “ハンバーグ”の中で“和風”を強める語と弱める語の例

強める語		弱める語	
正解	不正解	正解	不正解
大根おろし	おろす	ケチャップ	ソース
みりん	水溶き	ウスターソース	
醤油	片栗粉	赤ワイン	
大根	とろみ	トマト	
だし汁	摩る	チーズ	

最後に，比較範囲内ではほとんど出現しないが，レシピ全体を見て，修飾語を強めると推定できる語を求めた．例えば，“本格”なものとして“八角”や“パルミジャーノ”，“和風”なものとして“鰹節”や“牛蒡”などが得られた．

以上の実験で得られた語の適合率を，表 3 に示す．

表3 各入力における修飾語を強める語と弱める語の適合率

入力	相対的に強める語	絶対的に強める語	弱める語
(“本格”, “カレー”)	56.8	54.3	50.0
(“和風”, “ハンバーグ”)	63.2	56.0	83.3
(“ヘルシー”, “ハンバーグ”)	61.5	67.7	30.0
(“さっぱり”, “パスタ”)	68.2	39.0	50.0

7． 考察

まず，修飾語を相対的に強める語についてであるが，修飾語を強める語と良く共起する語が結果に含まれることで，適合率が下がっている場合が多い．例えば，（“ヘルシー”“ハンバーグ”）では，“豆腐”は“ヘルシー”を強めていると考えられるが，“豆腐”に“重し”を乗せて“水切り”したものを“つぶす”ことが多いため，“重し”“水切り”“つぶす”も“ヘルシー”を強める語として出力されてしまう．これは，単語の出現頻度だけで判定をしているのが原因であるため，語集合の出現頻度を利用することで解決できる可能性がある．

また，修飾語を絶対的に強める語を求める際には，ある分野において修飾語と共起するため出現頻度に有意差がでる語が多かった．例えば“ワカメ”は“さっぱり”という語を強めるとは考えにくいが，サラダや酢の物にワカメが使われることが多かったため，結果的に“さっぱり”と共起する割合が多くなってしまったと考えられる．

8． 終わりに

本研究では，オブジェクトに付けられた修飾語と内容の合致度を求めるために，修飾語を強める語と弱める語を求めた．

実際に，各オブエジェクトについて合致度を求める際には，今回求めた語を使用することになる．その際，単純にこれらの語をいくつ含むかではなく，各語の強弱の重みを考慮する必要があると考えられる．今後は，得られた語をもとに実際にどのように合致度を求めるのかを考えていきたい．

謝辞

本研究の一部は，科学研究費補助金（課題番号 18049041，18049073，21700106），および京都大学 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」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] R. Lee, D. Kitayama, and K. Sumiya, “Web-based evidence excavation to explore the authenticity of local events,” WICOW 2008, pp.63-66.

[2] 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.

[3] T. Kobayashi, H. Ohshima, S. Oyama, and K. Tanaka, “Evaluating brand value on the web,” WICOW 2009, pp.67-74.

[4] M. Kato, H. Ohshima, S. Oyama, and K. Tanaka, “Can Social Tagging Improve Web Image Search?,” WISE 2008, pp.235-249.

[5] クックパッド, http://cookpad.com/

[6] MeCab, http://mecab.sourceforge.net/