

不確かな順序関係の分析に基づく経験的属性によるオブジェクト検索

内田 臣了[†] 山本 岳洋[†] 加藤 誠[†] 大島 裕明[†] 田中 克己[†]

[†] 京都大学大学院情報学研究科 〒 606-8501 京都府京都市左京区吉田本町

E-mail: †{uchida,tyamamot,kato,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし 本研究では、「持ち運びしやすく使いやすいカメラ」といった、経験的属性をクエリとしたオブジェクトの検索手法を提案する。一般的なオブジェクト検索システムでは、「重さ」や「大きさ」など、主観を排した即物的な探索的属性によるクエリによる検索が実現されている。一方で、「使いやすさ」や「持ち運びやすさ」など主観を交えた経験的属性によるクエリでの検索は、一般に関連文書とクエリ文字列の一致判定により実現されるが、検索精度が十分に得られないという問題がある。本研究では、オブジェクトの探索的属性と、クエリとして入力された経験的属性の間の対応関係に着目することで問題の解決を図る。提案手法では、オブジェクトのレビューテキスト中の単体評価および相対評価からオブジェクトの順序関係を推定し、それらの信頼性の評価を行う。その上で、得られた信頼度付き順序関係を教師データとして Fuzzy Ranking SVM による機械学習を行い、オブジェクトのランキングを生成する。また、収集した正解セットを用いて手法の妥当性を評価する。

キーワード レビュー分析, テキストマイニング, 機械学習

1. はじめに

インターネットの普及に伴い、オブジェクトの実体に触れることなく、自身の要求に適合するオブジェクトを探す場面が増加している。例えば、Amazon.com^(注1) のような EC サイトを利用する場合、我々は実際の商品に触れることなく、購入する商品の比較検討を行う。我々は検索システムを用いることで、様々な属性に基いてオブジェクトをランキング・フィルタリングし、自身の要求を満足するオブジェクトを発見する。

本論文では、オブジェクトのもつ属性を探索的属性と経験的属性の2つに分類する。人がそのオブジェクトの実体に触れずとも評価できる属性を探索的属性という。例えば、「重さ」や「形」、「機能の有無」は探索的属性である。一方、人がそのオブジェクトの実体に触れて初めて評価できる属性のことを経験的属性という。例えば、「使いやすさ」や「持ち運びやすさ」は経験的属性である。

この2つの属性のうち、経験的属性は、オブジェクトを検索するユーザにとって特に重要な観点の1つである。一般的に、ユーザはある利用目的を想定し、オブジェクトの検索を行うと考えられる。ユーザはオブジェクトを比較検討し、その利用目的を達成できるオブジェクトを選択する。つまり、そのオブジェクトの実体に触れることで得られる経験を知ることができれば、ユーザはそのオブジェクトを選択するかどうかの意思決定を容易に行うことができる。

しかし、ユーザにとって、オブジェクトの経験的属性を知ることが容易ではない。一般的なオブジェクト検索システムでは、数値的な指標に基づいた検索、及び関連文書とクエリ文字列の一致判定に基づいた検索が可能である。経験的属性が数値化されることは極めて稀なため、数値的な指標に基づいた検索で経験

的属性をクエリにすることは困難である。例えば、EC サイトにおける商品のスペック情報は検索の指標となる数値であるが、「重さ」や「大きさ」など、そのほとんどは探索的属性である。一方で、関連文書とクエリ文字列の一致判定に基づいた検索も、経験的属性に基づいた検索を十分に実現しているとはいえない。なぜならば、そのオブジェクトのもつ経験的属性が、関連文書に全て記述されるわけではないためである。レビュー文に「持ち運びやすい」と明記されていないカメラでも、高い「持ち運びやすさ」をもつことはあり得る。つまり、関連文書とクエリ文字列の一致判定に基づいたアプローチには再現率の問題が存在する。また、このアプローチでは経験的属性を数値化することが困難である。例えば、「持ち運びやすい」と記述されたオブジェクトが複数存在した場合、どちらがより高い「持ち運びやすさ」をもつか判別できない。

オブジェクトのもつ探索的属性の情報から、その経験的属性を推定することは可能である。しかし、経験的属性の推定は、そのオブジェクトのドメインに関する専門知識を必要とすることが多く、一般的なユーザには難しいことが多い。例えば、撮像素子の大きさが「APS-C」のカメラと「中判」のカメラが存在した場合に、どちらがより「鮮明な写真が撮れる」か判断するのは、専門知識をもたないユーザには困難である。

内田ら[10]は、オブジェクトの関連文書から抽出されるオブジェクト間の順序関係と属性間の依存関係に着目し、経験的属性によるオブジェクトの検索手法を提案した。内田らの手法は、関連文書から抽出される順序関係と依存関係は正しいという仮定のもとに提案されている。しかし、文書に記述されている内容は必ずしも正しいとは限らない。そのため、関連文書から抽出される順序関係と依存関係について信頼性の問題が提起される。また、内田らは、関連文書から抽出された順序関係と依存関係から推定された順序関係を同等に扱っている。明記されていた順序関係は、依存関係から推定された順序関係よりも信頼

(注1) : <http://amazon.com>

できると考えられる。これらの間の信頼度の差を手法に反映すべきである。

本研究の貢献を以下に示す。

- 順序関係と依存関係に対し信頼度を導入し、関連文書から得られた情報の信頼度を反映したオブジェクトのランキング手法を提案した。
- オブジェクトに対する単体評価文から暗黙的な比較先を推定し、順序関係に変換する手法を提案した。

本稿の構成は以下のとおりである。2節では、関連研究を紹介する。3節では、本研究で用いる概念について説明を行う。4節では、提案手法について詳細に述べる。5節では、提案手法に関する実験と評価について述べる。6節では、まとめと今後の課題について述べる。

2. 関連研究

本節では、本研究と関連する研究について言及し、本研究の位置づけについて述べる。

2.1 ウェブ上の文書からの情報抽出

ウェブ上の意見文から有用な情報の抽出を行う研究について述べる。意見文からの情報抽出における主な課題は3つ組（対象、属性、評価）の抽出である。

立石ら[11]は、文書中での共起に注目することで、意見文から3つ組のうち対象物と評価表現を抽出し、その極性を判定した。また、立石らは[11]を発展させた研究として、意見文から属性と評価表現をブートストラップ的に抽出する手法を提案した。さらに、抽出した評価表現の極性を判定し、各属性を軸としたレーダーチャート形式での可視化を行った[12]。Liuら[6]は supervised な相関ルールマイニングを用いてルールを生成し属性を抽出する手法を提案した。また、Liuらも立石らと同様に各属性ごとに評価値の可視化を行った。

単一のオブジェクトの評価情報にとどまらず複数のオブジェクトの比較関係を抽出する研究も行われている。Liら[4]は weakly-supervised なブートストラップ法を用いることで、比較を意図した質問文から比較の対象となったオブジェクトの対を高い精度で抽出した。Liらでは語の一般化、具体化を行って抽出精度の向上を図っている。佐藤ら[8]はウェブ上のブログ記事から4つ組（対象、基準、属性、評価）または3つ組（対象、基準、評価）の抽出を行った。比較表現に注目したパターンマッチングに加え、構文情報とセンタリング理論を用いることで高精度での情報抽出を実現した。

本研究も意見文からの情報抽出に取り組む。本研究では、人手で作成したパターンとのマッチングにより属性間の依存関係とオブジェクト間の順序関係を抽出する。

2.2 ペアワイズ情報の集約によるランキングの生成

意見文中の比較評価を集約し、オブジェクトの順位付けを行う研究について述べる。

Guoら[1]はクラウドソーシングによってある集合内での最大値をとるオブジェクトを発見する手法を提案した。また、Guoらは新たに比較評価をすると推定精度が向上するような未評価のペアを選択する手法を提案した。Guoらの手法は人による

表1 カメラのもつ探索的属性の例

属性値		属性表現
重さ	0.2	軽い
大きさ	0.9	大きい

評価はすべてペアワイズであるという前提のもとで考えられている。

倉島ら[9]は従来の3つ組（対象、属性、評価）を4つ組（評価対象、比較対象、属性、評価）に拡張した。意見文中の比較表現に注目しパターンマッチングを行うことで評価対象と比較対象を抽出し、相関ルール分析を用いてオブジェクト間の優劣を推定しグラフを生成する。生成したグラフに対して PageRank に基づいたアプローチを行うことでオブジェクトのランク付けを行っている。

本研究は、オブジェクト間のペアワイズな関係を集約しランキングを行うという点で Guoら、倉島らの研究に類似している。Guoら、倉島らの研究では比較の主題となるクエリのみを扱っている。しかし、実データを扱うにあたって、比較表現に直接出現する属性はさほど多様でないという問題がある。本研究では、この問題を解決するアプローチとして文中の順接関係に着目し、比較の主題となりにくい属性についてもランキング規則を生成する。また、本研究では比較表現のみでなく、オブジェクト自身の持つ情報も考慮し、ランキング規則の生成を行うという点で前述の研究と異なる。

3. 定義

本節では、本研究の基盤となる概念について説明する。まず、オブジェクトのもつ属性の分類について述べる。次に属性間の依存関係、オブジェクト間の順序関係について定義を述べる。

3.1 探索的属性と経験的属性

Nelson[7]は、消費財を探索財 (search goods) と経験財 (experience goods) の2つに分類している。探索財とは、購入せずとも仕様等を確認することで品質の評価が可能な消費財である。また経験財とは、その商品を購入し、実際に扱って初めて品質を評価できるような消費財である。

本研究では、Nelsonによる消費財の分類に基づき、オブジェクトのもつ属性を**探索的属性**と**経験的属性**の2つに分類する。オブジェクトのもつ属性のうち、ユーザがオブジェクトの実体に触れなくても評価できる属性を探索的属性という。また、ユーザがオブジェクトの実体に触れて初めて評価できる属性を経験的属性という。本研究では、オブジェクトのもつ属性の程度の大きさを属性値として定量的に表現する。属性値は区間 $[0, 1]$ の間の実数値をとるものとする。また、人がその属性の程度を口語的に表現したものを属性表現と呼ぶ。カメラのもつ探索的属性の例を表1に、経験的属性の例を表2に示す。

探索的属性の属性表現は、しばしばオブジェクトの関連文書に記述される。また、探索的属性の一部はオブジェクトのスペック情報としてウェブ上に記述される。例えば、オブジェクトの重さや大きさ、機能の有無などはECサイトに記述される

表 2 カメラのもつ経験的属性の例

属性値		属性表現
携帯性	0.7	持ち運びやすい, 持ち運びに便利
夜景:強さ	0.8	夜景に強い, 夜景が綺麗に撮れる

ことも多い. こうした情報を参照することで, 我々はオブジェクトの探索的属性の属性値について知ることができる.

経験的属性の属性表現もまた, しばしばオブジェクトの関連文書に記述される. しかし, 経験的属性の属性値が記述されることは少ない. そのため, ユーザが, オブジェクトのもつ経験的属性の属性値について直接的に知るのは困難である.

オブジェクトのもつ経験的属性の属性値の推定が本研究の主題の 1 つとなる. 多くのユーザは何らかの利用目的をもってオブジェクトの選択を行うと考えられる. オブジェクトを利用した際に得られる経験を知りたいユーザに対し, そのオブジェクトのもつ経験的属性の属性値を提示することは, 彼らの意思決定に対する有効な支援になる.

3.2 オブジェクト間の順序関係

複数のオブジェクトに対し, ある属性の属性値に関して大小関係をつけることが可能である. これをある属性の基でのオブジェクト間の順序関係と呼ぶ.

オブジェクト間の順序関係を以下のように定義する. 属性 a_k について, オブジェクト o_i が異なるオブジェクト o_j より高い属性値をもつとき, a_k の基で o_i と o_j の間に順序関係が存在しているといい,

$$\pi_{i,j}^k = o_i \succ_{a_k} o_j \quad (1)$$

のように表記する. ただし, 全オブジェクト集合 O について, $o_i, o_j \in O$ である. また, 全属性集合 A について, $a_k \in A$ である.

3.3 属性間の依存関係

ある属性の属性値が変動すると, 異なる属性の属性値がそれに対応し変動することがある. 例えば, オブジェクトの探索的属性「画面の大きさ」が増加すると, そのオブジェクトの経験的属性「画面の見やすさ」は増加すると考えられる. オブジェクトの探索的属性「重さ」が増加すると, そのオブジェクトの経験的属性「持ち運びやすさ」は減少すると考えられる. 本研究ではこのような属性間の関係を依存関係と呼ぶ.

属性間の依存関係について以下のように定義する. 相異なる属性 a_k, a_l について, a_k の属性値の変動に対応して a_l の属性値が変動するとき, a_k から a_l への間に依存関係が存在するという. ただし, 全属性集合 A について, $a_k, a_l \in A$ である. 属性間の依存関係には以下の 4 種類が存在する.

- (1) a_k の属性値が増加すると a_l の属性値が増加するもの
 - (2) a_k の属性値が減少すると a_l の属性値が減少するもの
 - (3) a_k の属性値が増加すると a_l の属性値が減少するもの
 - (4) a_k の属性値が減少すると a_l の属性値が増加するもの
- これら 4 種の依存関係のうち, 1, 2 を正の依存関係と呼び,

$$\delta_{k,l}^+ = a_k \rightarrow_+ a_l \quad (2)$$

表 3 関連文書の例

- (1) 持ちやすく撮影に集中できます
- (2) ファインダーが見やすいのでピントが合わせやすいです
- (3) グリップが抜群に持ちやすいので, 移動の際も手が疲れにくいですね
- (4) ファインダーは X3 に比べて見やすいと思う

のように表記する. また, 3, 4 を負の依存関係と呼び,

$$\delta_{k,l}^- = a_k \rightarrow_- a_l \quad (3)$$

のように表記する.

3.4 信頼度

3.2 節でオブジェクト間の順序関係を, 3.3 節で属性間の依存関係をそれぞれ定義した. 4 節では, オブジェクトの関連文書から順序関係と依存関係の推定を行う. しかし, 関連文書から得られる順序関係, 依存関係が必ずしも正しいとは限らないという問題がある. そこで本研究では, 順序関係と依存関係に対し, そのもっともらしさを表す尺度として信頼度を導入する. 信頼度は, 区間 $(0, 1]$ の間の実数値をとるものとする.

4. 経験的属性によるオブジェクト検索

本節では, 我々が提案する経験的属性によるオブジェクトの検索手法について詳細に述べる. 本研究の目的は, オブジェクトの探索的属性から経験的属性の属性値を推定しランキングを行い, 経験的属性によるオブジェクト検索を実現することである. 本手法は, 以下の 6 つのステップから構成される.

- (1) 文書中の属性表現からの属性情報抽出.
- (2) 文書中の比較評価文からの順序関係の抽出.
- (3) 文書中の単体評価文からの不確かな順序関係の推定.
- (4) 文書中の文の順接関係からの不確かな依存関係の抽出.
- (5) 依存関係と順序関係の組み合わせによる新たな順序関係の生成.
- (6) 信頼度付き順序関係を教師としたランキング学習によるランキング規則の推定.

本手法では, オブジェクトの関連文書からオブジェクト間の順序関係や属性間の依存関係を抽出する. 関連文書の例を表 3 に示す.

4.1 属性表現からの属性情報収集

オブジェクトの関連文書には, そのオブジェクトのもつ属性について言及した文, すなわち属性表現が多く記述される. 属性表現からはその属性の属性値を推定することが可能である.

本手法では, オブジェクトの関連文書に対し, 人手で用意した属性表現の抽出パターンとのマッチングを行うことで, 文書中の属性表現からオブジェクトのもつ属性についての情報を収集する. ここで収集する属性の情報は, 属性名と評価値の 2 つ組である. 評価値とは, その属性の属性値をおおまかに高低の 2 値に分類したものである.

属性表現の抽出パターンを表 4 に示す. パターンの末尾に否定の助動詞が接続されていた場合は, 評価値を反転させる. 表 3 の (1) に対して, 表 4 で示した属性表現のパターンを適用した場合, (持ちやすさ, 高) と (撮影の集中しやすさ, 高) が抽

表 4 属性表現の抽出パターン

パターン	属性名	評価値
$\langle [x/\text{名詞}][\text{は} \text{が} \text{に} \text{も}/\text{助詞}]? \langle [y/\text{形容詞}] \rangle$	(x, y)	高
$\langle [x/\text{名詞}][\text{は} \text{が} \text{に} \text{も}/\text{助詞}]? \langle [y/\text{名詞} \cdot \text{形容動詞語幹}][*/\text{助動} \text{詞}] \rangle$	(x, y)	高
$\langle [x/\text{名詞}][\text{は} \text{が} \text{に} \text{も}/\text{助詞}]? \langle [y/\text{動詞}][\text{やすい}/\text{形容詞}] \rangle$	(x, y)	高
$\langle [x/\text{名詞}][\text{は} \text{が} \text{に} \text{も}/\text{助詞}]? \langle [y/\text{名詞}][\text{できる}/\text{動詞}] \rangle$	(x, y)	高
$\langle [x/\text{名詞}][\text{は} \text{が} \text{に} \text{も}/\text{助詞}]? \langle [y/\text{動詞}][\text{にくい}/\text{形容詞}] \rangle$	(x, y)	低
$\langle [x/\text{名詞}][\text{は} \text{が} \text{に} \text{も}/\text{助詞}]? \langle [y/\text{動詞}][\text{づらい}/\text{形容詞}] \rangle$	(x, y)	低

出される。また、表 3 の (2) に属性表現のパターンを適用した場合、(ファインダーの見やすさ、高) と (ピントの合わせやすさ、高) が抽出される。

4.2 比較評価文からのオブジェクト間の順序関係の収集

本手法では、オブジェクトの関連文書に記述された評価文に注目することでオブジェクト間の順序関係の抽出を行う。文書中に記述される評価文は、比較評価文と単体評価文の 2 つに分類される。比較評価文とは、オブジェクトの属性について言及した文のうち、明示的に比較対象が挙げられ、相対評価が為されているものである。一方、単体評価文とは、オブジェクトの属性について言及した文のうち、比較対象が記述されていないものである。

本節では、比較評価文からオブジェクト間の順序関係を抽出する手法について述べる。得られた順序関係は、4.5 節で述べる順序関係の拡張に用いられるほか、4.6 節で述べるランキング学習の教師データとして用いられる。

ユーザは利用したオブジェクトと他のオブジェクトとの比較を行い、その関係を文書に記述する。ユーザの記述した比較評価文からは、ある属性に関するオブジェクトの優劣、すなわち、ある属性に関するオブジェクトの順序関係を収集できる。また、一般に、評価者は自らの経験に基づき評価を行うと考えられるため、比較評価文に注目することで経験的属性に関する順序関係を収集できる。経験的属性に関する順序関係を集約して得られる知見は、経験的属性値の推定に有用であると考えられる。

Jindal ら [2] は比較文を以下の 4 つに分類している。

- **Non-Equal Gradable:** ある属性に関して、複数のオブジェクト間の優劣関係を表現したもの。
- **Equative:** ある属性に関して、2 つのオブジェクトの等位関係を表現したもの。
- **Superlative:** ある属性に関して、あるオブジェクトの最上位性を表現したもの。
- **Non-Gradable:** ある属性に関して、複数のオブジェクトを比較したものだが、優劣関係については言及していないもの。

本手法では、Non-Equal Gradable な比較文のみを用いて、順序関係の推定を行う。

比較評価文は、オブジェクトの関連文書に対してパターンマッチングを行うことで抽出する。比較評価文のパターンを表 5 に示す。比較評価文のパターンは主語パターンと基準パターンからなる。主語パターン中の s に対応した語は、その比較評

表 5 比較評価文のパターン

主語	$\langle [s/\text{名詞}][\text{の}/\text{助詞}] \langle [方/\text{名詞}][\text{が}/\text{助詞}] \rangle \langle (e_a/\text{属性表現}) \rangle$
基準	$\langle [c/\text{名詞}][\text{より}/\text{助詞}] \rangle \langle (e_a/\text{属性表現}) \rangle$
	$\langle [c/\text{名詞}][\text{と} \text{に}/\text{助詞}] \rangle \langle [\text{比べる}/\text{動詞}] \rangle \langle (e_a/\text{属性表現}) \rangle$
	$\langle [c/\text{名詞}][\text{と}/\text{助詞}] \rangle \langle [\text{比較}/\text{名詞}][\text{する}/\text{動詞}] \rangle \langle (e_a/\text{属性表現}) \rangle$

価文の主語とみなされる。また、基準パターン中の c に対応した語は、その比較評価文の基準とみなされる。属性表現 e_a に対応した属性 a の評価値が高い場合は順序関係 $s \succ_a c$ を、 a の評価値が低い場合は順序関係 $c \succ_a s$ を抽出する。

上記の手法により、文書から比較評価文を発見し順序関係を収集するが、比較文の主語や基準が省略される場合も多い。そうした比較文については、省略により欠落したオブジェクトを推測し補完する必要がある。本手法では、主語または基準いずれかの欠落があり、欠落していないもう一方が関連文書が紐付いたオブジェクトでなかった場合、関連文書が紐付いたオブジェクトで欠落を補完する。これは、関連文書全体を通しての主題は、文書が紐付いたオブジェクトであるという仮定に基づく。

4.3 単体評価文からの順序関係推定

4.2 節では、比較評価文から順序関係を収集する手法について述べた。比較評価文に由来する順序関係は、信頼度が高いという特徴をもつ一方で、4.6 節で用いるのに十分な数を収集するのが難しいという問題がある。この問題を解決するために、単体評価文に基づく順序関係の生成法を提案する。単体評価文は比較評価文と比べて数が多いため、単体評価文からの順序関係の生成は順序関係数の不足を解決する糸口になる。

本節では、単体評価文を比較対象が明記されていない相対評価として捉え、順序関係の推定を行う。単体評価文には比較対象が記述されていないため、絶対的な評価が為されていると捉えることも可能である。しかし、評価されるオブジェクトによって、同じ表現でも指し示す属性の大きさが明らかに異なる場合がある。例えば、一眼レフカメラに対する「持ち運びやすい」という評価と、トイカメラに対する「持ち運びやすい」という評価は、表現の上では同一だが、指し示す「持ち運びやすさ」の大きさは明らかに異なる。つまり、一眼レフカメラには一眼レフカメラの評価基準が、トイカメラにはトイカメラの評価基準が別々に存在していると考えられる。以上から、我々は単体評価文にも暗黙的な比較対象が存在すると仮定し、単体評価文からの順序関係推定を行う。

比較評価文には比較対象が明記されているため、その対象との間の順序関係を抽出することができた。一方で、単体評価文には比較対象が記述されていないため、評価者が何と比較して評価を行ったか推定する必要がある。推測の結果として得られる比較対象が必ずしも正しいとは限らないため、単体評価文から順序関係を推定する際には、3.4 節で導入した信頼度を考慮する必要がある。

本稿では、そのオブジェクトと明示的に最も比較されたオブジェクトで、単体評価文の暗黙的な比較対象を補完する。また、その信頼度には、他の比較対象との類似度の平均を用いる。こ

表 6 順接関係の抽出パターン

$\langle (e_a/\text{属性表現})[\text{の}/\text{助詞}] \rangle \langle (e_{a'}/\text{属性表現}) \rangle$
$\langle (e_a/\text{属性表現})[\text{から}/\text{助詞}] \rangle \langle (e_{a'}/\text{属性表現}) \rangle$

れは、あるオブジェクトについて人々がよく比較対象としてあげるオブジェクトが比較対象である可能性が高く、かつ、それらは互いに類似しているという仮定に基づいている。

順序関係集合を Π とする。順序関係 $\pi_{i,j}^k = o_i \succ_{a_k} o_j$ が推定されたとき、 $\pi_{i,j}^k$ の信頼度を求める関数 $\text{rel}_{\text{order}} : \Pi \mapsto (0, 1]$ を下式の通り定義する。ここで、 o_i と比較されたオブジェクトの集合を $O_{\text{compared.to}(o_i)}$ と表記する。

$$\text{rel}_{\text{order}}(\pi_{i,j}^k) = \frac{\sum_{x \in O_{\text{compared.to}(o_i)}} \text{sim}(o_j, x)}{|O_{\text{compared.to}(o_i)}|} \quad (4)$$

ここで、オブジェクト間の類似度 $\text{sim}(o_i, o_j)$ は、それぞれを表すベクトルのコサイン類似度で求められる。オブジェクトのベクトル表現については、4.6 節で詳細を述べる。

4.4 属性間の依存関係の収集

本節では、オブジェクトの関連文書から属性間の依存関係を収集する手法について述べる。4.2, 4.3 節で、オブジェクト間の順序関係を収集する手法について述べた。収集した順序関係は、4.6 節で述べるランキング学習の教師データとして用いられるが、その数が不十分であるという問題がある。本手法では、属性間の依存関係に着目し、順序関係を拡張することで数の不足を補う。具体的な手法については、4.5 節で述べる。

接続助詞を用いて構築される順接関係に注目し、属性間の依存関係の生成を行う。接続詞や接続助詞により、文と文の間に論理関係が構築される。特に、文と文の間に順接関係が存在する場合は、「前件が後件の順当な原因・理由になっている」^(注2)と考えられる。そのため、属性表現を含む文節と属性表現を含む文節の間に順接関係が存在する場合は、それらが表す属性の間に依存関係が存在すると考えられる。

関連文書に対し、パターンとのマッチングを行うことで順接関係を抽出し、属性の依存関係を収集する。表 6 に順接関係の抽出パターンを示す。

パターン中の属性表現 e_a と $e_{a'}$ にそれぞれ対応する属性 a , a' の間の依存関係を取得する。 a と a' の評価値が一致していた場合は、正の依存関係 $a \rightarrow_+ a'$ が成立していると考ええる。また、 a と a' の評価値が異なっていた場合は、負の依存関係 $a \rightarrow_- a'$ が成立していると考ええる。

例えば、表 3 の (2) にパターンを適用した場合、正の依存関係「ファインダーの見やすさ」 $\rightarrow_+$ 「ピント:合わせやすさ」が抽出される。同様に、表 3 の (3) にパターンを適用した場合、負の依存関係「グリップの持ちやすさ」 $\rightarrow_-$ 「手の疲れやすさ」が抽出される。

1 節で言及した通り、属性から属性への依存関係を推定することは、一般的なユーザにとって簡単なタスクではない。その

ため、オブジェクトの関連文書から収集された依存関係には、誤りが含まれると推測される。ゆえに、依存関係について信頼度の計算を行う必要があると考えられる。

依存関係の信頼度を、下式の通り定義する。依存関係集合を Δ とする。順序関係 $\delta_{i,j}^* = \delta_{i,j}^+ \cup \delta_{i,j}^-$ が与えられたとき、 $\delta_{i,j}$ の信頼度を求める関数 $\text{rel}_{\text{dependency}} : \Delta \mapsto (0, 1]$ を下式の通り算出する。ただし、 $\text{appearance}(\delta)$ は、依存関係 δ の関連文書内における被記述回数を表す。また、 P は、記述された依存関係が正しい確率である。

$$\text{rel}_{\text{dependency}}(\delta_{i,j}^*) = 1 - (1 - P)^{\text{appearance}(\delta)} \quad (5)$$

式 (5) は、記述された依存関係のうち少なくとも 1 つが正しいものである確率を表す。

4.5 依存関係を用いた順序関係の拡張

4.2 節で、オブジェクトの関連文書からオブジェクト間の順序関係を抽出する手法について説明した。得られた順序関係を教師データとして用いて、4.6 節で述べるランキング学習を行うことでランキング関数を学習できる。しかし、文書から収集できる順序関係の数は多いとはいえない。特に、経験的な属性表現の出現数は探索的な属性表現に比べて少ないため、ランキング学習の教師データとして用いるには数が不十分であることが多いと考えられる。そのため、ランキング規則の学習を行うにあたって教師データの不足を補う必要がある。

本手法では、属性間の依存関係を用いてオブジェクト間の順序関係の拡張を行い、新たな順序関係の獲得を試みる。以下の仮説に基づき、4.4 節で抽出した属性間の依存関係と 4.2 節で抽出したオブジェクト間の順序関係から新たな順序関係を生成する。

順序関係 (1) と正の依存関係 (2) がともに成立しているとき、順序関係 (6) が成立する。

$$\pi_{i,j}^l = o_i \succ_{a_l} o_j \quad (6)$$

順序関係 (1) と負の依存関係 (3) がともに成立しているとき、順序関係 (7) が成立する。

$$\pi_{j,i}^l = o_j \succ_{a_l} o_i \quad (7)$$

また、その信頼度を求める関数を以下の通り定義する。

$$\text{rel}(\pi_{i,j}^l) = \min(\text{rel}(\pi_{i,j}^k), \text{rel}(\delta_{k,l}^+)) \quad (8)$$

4.6 Fuzzy Ranking SVM によるランキング学習

4.2 節, 4.3 節, 4.5 節でオブジェクト間の順序関係の収集を行った。これらの順序関係を教師データとしたランキング学習により、オブジェクトの探索的属性から、経験的属性に基づくランキング関数を学習する。

各オブジェクトを探索的属性のベクトルで表現する。全オブジェクト集合を $O = \{o_1, o_2, \dots, o_m\}$ 、オブジェクトのもつ探

(注2)：順接 (ジュンセツ) とは - コトバンク, <https://kotobank.jp/word/順接-530004>

索的属性の集合を $A_S = \{a_1^{(s)}, a_2^{(s)}, \dots, a_{n_s}^{(s)}\}$ とすると、オブジェクト o_i を表現した探索的属性のベクトル $\mathbf{o}_i$ は以下のように表現される。

$$\mathbf{o}_i = (o_{1,i}, o_{2,i}, \dots, o_{n_s,i})^T \in \mathbb{R}^{n_s} \quad (9)$$

ここで、 $o_{j,i}$ は、オブジェクト o_i がもつ探索的属性 $a_j^{(s)}$ の属性値であり、 $0 \leq o_{j,i} \leq 1$ を満たす実数である。ただし、 $j = 1, 2, \dots, n_s$ である。

オブジェクトのもつ探索的属性は、ウェブ上に記載されたスペック情報などから収集できる。記載されたスペック値を区間 $[0, 1]$ の実数値に正規化したものを属性値とし、ベクトルの各要素とする。

ある経験的属性 $a^{(e)}$ に関する順序関係集合 $T = \{\tau_1, \tau_2, \dots, \tau_l\}$ を教師データとしたランキング学習により、オブジェクトを $a^{(e)}$ に基づいてランキングする関数を学習する。各教師データは以下のように表される。

$$\tau_k = \mathbf{o}_{x_k} \succ_{a^{(e)}} \mathbf{o}_{y_k} \quad (10)$$

ただし、 $\mathbf{o}_{x_k}, \mathbf{o}_{y_k}$ でそれぞれ表されるオブジェクト o_{x_k}, o_{y_k} について、 $\mathbf{o}_{x_k}, \mathbf{o}_{y_k} \in O$ が成り立つ。また、 τ_k の信頼度を $r_k \in [0, 1]$ とする。

これらの教師データからランキング関数を学習する手法として、Fuzzy Ranking SVM を提案する。Fuzzy Ranking SVM は、Ranking SVM [3] と Fuzzy SVM [5] を組み合わせた、不確かなペアワイズ教師データに基づくランキング学習法である。教師データとして、 $(\mathbf{x}_i, \mathbf{x}_j, y_{i,j}, s_{i,j})$ の4つ組を元とする集合を受け取る。ここで、 $\mathbf{x}_i, \mathbf{x}_j$ はオブジェクトを表すベクトルである。 $y_{i,j}$ は、 $\mathbf{x}_i$ が表すオブジェクトが $\mathbf{x}_j$ が表す文書より上位のときに1、同位のときに0、下位のときに-1となる変数である。 $s_{i,j}$ は、ファジィ集合におけるメンバシップ値である。Fuzzy Ranking SVM の目的関数及び制約式を以下に示す。ここで、 $\mathbf{w}$ は学習される重み、 C はスラック変数に掛かるコスト、 $\xi_{i,j}$ はスラック変数である。

$$\text{minimize: } \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum s_{i,j} \xi_{i,j} \quad (11)$$

$$\text{subject to: } \forall \{(\mathbf{x}_i, \mathbf{x}_j) : y_{i,j} > 0\} :$$

$$\mathbf{w} \cdot \mathbf{x}_i \geq \mathbf{w} \cdot \mathbf{x}_j + 1 - \xi_{i,j} \quad (12)$$

$$\xi_{i,j} \geq 0 \quad (13)$$

式(10)の形で表される教師データを、 $\mathbf{x}_i \leftarrow \mathbf{o}_{x_k}, \mathbf{x}_j \leftarrow \mathbf{o}_{y_k}, y_{i,j} \leftarrow 1, s_{i,j} \leftarrow r_k$ のように変換し、Fuzzy Ranking SVM に学習させることで、最終的にランキング関数 $f_{a_e} : \mathbb{R}^{n_s} \mapsto \mathbb{R}$ を得る。

f_{a_e} はオブジェクト o_i の探索的属性ベクトル $\mathbf{o}_i$ を入力として受け取り、そのオブジェクトの属性 $a_e^{(k)}$ の属性値を推定する関数である。この関数を用いることで、オブジェクトの探索的属性の属性値を入力すると、経験的属性 a_e に基づいたランキングを行うことが可能になる。

5. 実験

本節では、提案手法を用いて行った実験とその評価について述べ、結果について考察を行う。本実験の目的は、提案手法の有効性について検証することである。

5.1 データセット

価格.com^(注3) より収集したデータを用いて実験を行う。収集したデータのうち、カメラカテゴリに属する商品619件についてのスペック情報をオブジェクトの探索的属性として用いた。また、上記619件の商品を対象としたユーザーレビュー17,061件をオブジェクトの関連文書として用いた。

Ranking SVMでの学習を行うにあたり、オブジェクトを探索的属性のベクトルとして表現する必要がある。本実験では、スペック情報からオブジェクトを51次元のベクトルとして表現した。画素数や重量など、数値を要素にもつフィールドは、値を最大値で割ることで $[0, 1]$ の範囲に正規化した。サブカテゴリやメーカーなど、複数の候補から1つが選ばれるカテゴリカルなフィールドについては、候補それぞれについて次元を設け、対応する次元に1を、それ以外に0を格納するものとした。スペック情報が欠損していた場合は、平均値や最頻値で補完を行い、学習に大きな影響を与えないよう配慮した。

5.2 ベースライン

提案手法と比較するベースラインには、属性の言及数に基づいたランキングを採用した。例えば、「使いやすさ」に関するランキングであれば、レビュー中で「使いやすい」とより多く言及されているオブジェクトを上位に位置づける。これは、一般的なユーザが商品を比較する際に用いるランキング手法であると考えられる。

5.3 評価方法

ユーザーレビューから順序関係が多く得られた属性のうち、経験的属性であると人手で判断したものを評価クエリとして採用した。評価に用いるクエリとその属性に関する順序関係数を表7に示す。「拡張なし」の列は、レビューから得られた順序関係数を表す。「拡張あり」の列のうち、「依存関係」は、依存関係によって拡張された順序関係数を表す。「単体評価文」は、単体評価文を用いて拡張された順序関係数を表す。「依存関係と単体評価文」は、先述した2つの手法を共に実行した際の順序関係数を表す。

本実験では、ユーザーレビューから抽出されたオブジェクト間の順序関係を正解とみなす。正解とされた順序関係を1つ教師データより除外し学習を行い、学習された関数が除外した順序関係を正しく導出できるか一個抜き交差検定 (leave-one-out cross validation) を行う。

評価実験は、それぞれのクエリに対し、以下の6通りの処理を順序関係に施した場合について実施する。

- 拡張なし
- 依存関係に基づく順序関係の拡張
- 依存関係に基づく順序関係の拡張 (信頼度あり)

(注3) : <http://kakaku.com>

表 7 クエリと順序関係数

クエリ	拡張あり			
	拡張なし	依存関係	単体評価文	依存関係と単体評価文
使いやすさ	46	733	2468	3155
持ちやすさ	38	744	1247	1953
撮影しやすさ	17	166	1061	1210
ホールド感の良さ	16	407	391	782
見やすさ	14	231	808	1025
画質の良さ	14	68	446	500
自然さ	13	13	384	384
コンパクトさ	14	142	1261	1389
ノイズの少なさ	12	12	250	250
便利さ	10	568	1576	2134
楽しさ	10	32	1687	1709
操作しやすさ	9	199	851	1041
扱いやすさ	7	107	280	380
握りやすさ	7	209	225	427

- 単体評価文に基づく順序関係の拡張
- 単体評価文に基づく順序関係の拡張（信頼度あり）
- 依存関係と単体評価文に基づく順序関係の拡張（信頼度あり）

Fuzzy Ranking SVM の実装には、オープンソース機械学習ライブラリ scikit-learn^(注4) を用いた。Fuzzy Ranking SVM のカーネルには、RBF カーネルを用いた。文の形態素解析には、形態素解析エンジン MeCab^(注5) を利用した。係り受け解析には、日本語係り受け解析器 CaboCha^(注6) を利用した。

5.4 結果と考察

実験結果を表 8 に示す。表 8 の「ベースライン」はベースラインの精度を表す。「提案手法（拡張なし）」は、順序関係の拡張を行わなかった場合の精度を表す。「提案手法（拡張あり）」のうち、「依存関係」、「単体評価文」は、それぞれ依存関係に基づく拡張、単体評価文に基づく拡張を行った場合の精度を表す。また、末尾に「(信頼度)」と記述されたものは信頼度の計算を行ったものである。

実験の結果、マクロ平均とマイクロ平均の両方において、提案手法がベースラインよりも高い精度を確保した。また、信頼度の計算を行うことで全体的な精度の向上が観測された。特に、依存関係による拡張を行った場合の精度低下は解消され、拡張を行わなかった場合と同等以上の精度を記録した。一方で、単体評価文に基づいた順序関係の拡張を行った場合は、拡張を行わなかった場合と比べて全体的な精度の低下が見られた。

本実験では、信頼度の作用による精度の向上が観測された。依存関係に基づいた拡張においては、13 クエリのうち 8 つについて精度が向上し、逆に精度が低下したものは 1 つに留まった。単体評価文に基づいた拡張においては、13 クエリのうち 6 つについて精度が向上し、3 つについて精度が低下した。信頼度による教師データの重み付けは、有効に作用したと考えられる。

本実験では、単体評価文に基づく拡張は精度の向上を実現しなかった。精度低下の原因として、比較対象の誤りが考えら

る。補完した比較対象が、評価者の想定していたものと異なっていた場合、そこから推測された順序関係は誤っていることになる。本手法では、そのオブジェクトと最も比較されたオブジェクトで比較対象の補完を行った。比較対象の推定指針はオブジェクトのドメインごとに多数考えられる。例えば、カメラドメインにおける暗黙的な比較対象の推定指針としては、価格帯、発売時期、製造メーカーなどが考えられる。今後は、より正確な比較対象の推定、それに伴う信頼度関数の決定に取り組む予定である。

本実験では、ユーザによってレビューに記述されたオブジェクト間の順序関係を正解としている。しかし、関連文書中に記述から得られる経験的属性に基づく順序関係は少ないため、正解セットが不足している問題がある。より正確な評価のため、専門家による評価を実施し正解セットを拡充する必要がある。また、本実験ではカメラカテゴリのオブジェクトに対してのみ評価を行った。他のドメインについての実験も今後行う予定である。

6. まとめと今後の課題

本研究では、オブジェクトの探索的属性から経験的属性に基づくランキング規則を推定する手法を提案し、経験的属性による検索を可能にした。

本手法では、あらかじめ用意したパターンとのマッチングによって、オブジェクトの関連文書中の比較評価文からオブジェクト間の順序関係を、文の順接関係から属性間の依存関係を収集した。また、関連文書中の単体評価文について暗黙的な比較対象を推定することで、単体評価文からの順序関係抽出を行った。収集した依存関係に基いて順序関係を拡張することで、関連文書中に記述されていない新たな順序関係を導出した。得られた順序関係と依存関係に信頼度を導入し、教師データの重み付けを行った。最後に、得られた信頼度付き順序関係を教師データとして用いて Fuzzy Ranking SVM によるランキング学習を行い、経験的属性に基づくランキング関数を得た。

本稿では、EC サイト価格.com から収集したカメラに関するスペック情報、レビューデータを用いて評価実験を行った。評価実験の結果、信頼度が精度の向上に有効に作用していることを確認した。また、単体評価文に基づいた拡張による精度の低下について言及し、単体評価文における比較対象推定について考察を行った。

今後は、専門家によるオブジェクトの評価に基づいた正解データの収集を行う他、他ドメインを対象とした実験を行い、より多くのデータを対象に本手法の有用性を検証する予定である。

謝 辞

本研究は JSPS 科研費 15H01718, 25240050 の助成を受けたものです。

(注4) : <http://scikit-learn.org/>

(注5) : <https://code.google.com/p/mecab/>

(注6) : <https://code.google.com/p/cabocha/>

表 8 実験結果

クエリ	ベースライン	提案手法（拡張なし）	提案手法（拡張あり）				
			依存関係	依存関係（信頼度）	単体評価文	単体評価文（信頼度）	依存関係と単体評価文（信頼度）
使いやすさ	0.413	0.609	0.413	0.457	0.543	0.652	0.609
持ちやすさ	0.474	0.711	0.658	0.684	0.605	0.605	0.605
撮影しやすさ	0.412	0.558	0.471	0.647	0.529	0.471	0.471
ホールド感の良さ	0.562	0.875	0.625	0.875	0.688	0.812	0.938
見やすさ	0.357	0.571	0.571	0.643	0.714	0.714	0.714
画質の良さ	0.571	0.500	0.500	0.571	0.500	0.571	0.571
自然さ	0.538	0.846	0.846	0.846	0.615	0.615	0.615
コンパクトさ	0.500	0.571	0.643	0.643	0.429	0.571	0.571
ノイズの少なさ	0.417	0.667	0.667	0.667	0.583	0.583	0.583
便利さ	0.700	0.500	0.400	0.500	0.600	0.500	0.500
楽しさ	0.800	0.700	0.500	0.700	0.200	0.400	0.400
操作しやすさ	0.222	0.222	0.667	0.667	0.667	0.667	0.778
扱いやすさ	0.286	0.571	0.429	0.429	0.143	0.429	0.286
握りやすさ	0.714	0.857	1.000	0.857	0.857	0.714	0.714
マクロ平均	0.498	0.626	0.599	0.656	0.548	0.593	0.597
マイクロ平均	0.480	0.639	0.573	0.634	0.559	0.608	0.608

文 献

- [1] Stephen Guo, Aditya Parameswaran, and Hector Garcia-Molina. So who won?: dynamic max discovery with the crowd. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 385–396. ACM, 2012.
- [2] Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 244–251. ACM, 2006.
- [3] Thorsten Joachims. Optimizing search engines using click-through data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142. ACM, 2002.
- [4] Shasha Li, Chin-Yew Lin, Young-In Song, and Zhoujun Li. Comparable entity mining from comparative questions. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 25, No. 7, pp. 1498–1509, 2013.
- [5] Chun-Fu Lin and Sheng-De Wang. Fuzzy support vector machines. *Neural Networks, IEEE Transactions on*, Vol. 13, No. 2, pp. 464–471, 2002.
- [6] Bing Liu, Ming Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351. ACM, 2005.
- [7] Phillip Nelson. Information and consumer behavior. *The Journal of Political Economy*, pp. 311–329, 1970.
- [8] 佐藤敏紀, 奥村学. blog からの比較関係抽出. 情報処理学会 自然言語処理研究会, pp. 7–14, 2007.
- [9] 倉島健, 別所克人, 内山俊郎, 片岡良治. 比較評価情報の抽出とそれに基づくランキング手法の提案. 第 18 回データ工学ワークショップ (DEWS 2007), 2007.
- [10] 内田臣了, 山本岳洋, 加藤誠, 大島裕明, 田中克己. 経験的属性によるオブジェクト検索. 第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM2015), 2015.
- [11] 立石健二, 石黒義英, 福島俊一. インターネットからの評判情報検索. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 101, No. 189, pp. 75–82, 2001.
- [12] 立石健二, 福島俊一, 小林のぞみ, 高橋哲朗, 藤田篤, 乾健太郎, 松本裕治. Web 文書集合からの意見情報抽出と着眼点に基づく要約生成. 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2004, No. 93, pp. 1–8, 2004.