口コミ解析システムのための感情語辞書の検討

上谷竜士[1]　　菱田隆彰[2]

概要： 最近では口コミサイトの口コミがユーザの消費行動に影響を及ぼす度合いが強くなっている．口コミサイトに投稿された内容から対象物の印象をより詳細に得るには，対象物に対する全ての口コミを熟読し，客観的な分析を行うことが望ましいが，容易なことではない．我々は，これまでに価格.com，トリップアドバイザー，食べログを対象サイトとし，登録されている口コミを杉本らによって提案した感情語辞書を元に解析を行い可視化するシステムの構築を行った．しかし，口コミサイトは取り扱う対象物の種類によって，ユーザが投稿する内容に偏りが生じる可能性がある．本稿では，複数の傾向の異なる口コミサイトにおいて本手法による分析によって得られる印象の分布を調査し，各サイトに適した表現方法の考察を行う．

A study of an emotional words dictionary for a review analysis system

RYUJI UETANI[1]　　TAKAAKI HISHIDA[2]

1 はじめに

近年 Web 上ではロコミサイトなどある対象物に関しての感想をユーザが投稿し，その内容を他者が閲覧できるサービスの普及が進んでいる．投稿されたロコミは同じ興味を持つユーザや似た購買層のユーザであることから，そのロコミの読者は，その内容から自身が感じる対象物の印象を擬似的に得ることが可能となり，購買意欲や旅行意欲など対象物への直接的な行動に移すきっかけとなる．対象物の印象をより詳細に得るには，対象物に対する全てのロコミを熟読し，客観的な分析を行うことが望ましいが，容易なことではない．特に人気のある対象物は，ロコミされた投稿数が非常に多い場合があり，一つ一つ丁寧に読み解くことは欲しい結果に見合わないほどの手間が必要になってしまう．このような問題に対して我々は，ロコミの中に含まれる感情語に注目し，対象物の印象が一目で把握できる表現方法の提案を行い，ロコミからその印象を集計するシステムの試作を行った．

杉本らは，文献[1]において，ロコミを集約し，含まれる感情語に基づいて分析を行うための自律型データベースシステムの設計を行った．文献[2]では，ロコミサイトに投稿されている感情語の傾向を調査し，文献[3]において口コミの分析に適した感情語の分類を検討し，その感情語辞書を提案した（本稿では杉本式辞書と呼ぶ）．また，松井らは，文献[4]において，杉本式辞書を用いて口コミ内の感情語の分析を行い，価格.com，トリップアドバイザー，食べログのロコミに掲載されている対象物の印象をレーダーチャートで表示するスマートフォンアプリの実装を行った．
これまでの，研究によって幾つかのロコミサイトに対して我々の提案する分析手法によって，ロコミ内の感情語の分布から対象物の印象をグラフとして表現することが可能となった．しかし，ロコミサイトは取り扱う対象物の種類によって，ユーザが投稿する内容に偏りが生じる可能性がある．本稿では，複数の傾向の異なるロコミサイトにおいて本手法による分析によって得られる印象の分布を調査し，各サイトに適した表現方法の考察を行う．

2 対象物の印象の表現方法

関連する研究としては，松浦らによる文献[6]があげられる．松浦らは文芸書を対象としインターネット上の書評に含まれる感情表現から書籍の雰囲気を可視化する手法を提案している．感情表現辞典[7]に含まれる語彙をその書籍の印象の推定に用いることとし，「おもしろい」「いらいら」「リラックス」「つまらない」の 4 種のカテゴリの特徴量の算定を行い，各書評を 2 次元座標や散布図，レーダーチャートによって可視化を行うことで，全体の雰囲気と個々の感情表現の分布が一望できることを示した．

我々が提案する表現手法を以下に示す．松浦らと同様に印象の推定には感情表現辞典に含まれる語彙を用いる．杉本ら[3]が示した提案に従い選定した語彙を 5 つのカテゴリ，「喜び」「楽しさ」「安心」「好き」「驚き」に振り分けて分析用の辞書とする．

① 対象物に投稿されているロコミについて，すべての文章を形態素に分割する．
② 分割した単語について，カテゴリ分けした 5 つの感情語の辞書に含まれる単語の数をそれぞれ集計し，$C_1, C_2, \dots, C_5$ を得る．
③ 対象物の特徴量 S_i (i=1, … , 5)は

$$S_i = C_i / \max\{ C_i ; i = 1, \dots , 5\}$$

と求める．

求めた特徴量 S_i は，5 つのカテゴリの内，最も含まれている感情語の多いカテゴリの値を 1 (100%)とし，その他の

1 愛知工業大学大学院経営情報科学研究科
2 愛知工業大学情報科学部情報科学科

カテゴリの特徴量はその値との比率で表される．例えば，ある商品の口コミの感情語の集計値が

$$\{C_1, \dots, C_5\} = \{2, 1, 0, 8, 5\}$$

であった場合，その印象を表す特徴量は

$$\{S_1, \dots, S_5\} = \{0.25, 0.125, 0, 1, 0.625\}$$
$$= \{25\%, 12.5\%, 0\%, 100\%, 62.5\%\}$$

となる．この特徴量は図 1 に示すようにレーダーチャートで表現することで可視化することで，その商品の口コミが示す印象を一目で把握することが可能となる．

3 感情語辞書の拡張とその分布

3.1 感情語辞書と口コミ

分析に杉本式辞書に含まれる語彙は，感情表現辞典[7]に含まれる語彙を基に作成しており，基本的に漢字で表現できるものは漢字を使用したもののみを辞書に含めている．しかし，口コミサイトに投稿される文面の多くは砕けた口調であえて漢字を使用しない場合も多い．従って，漢字を使用した語彙のみでは，口コミ内の感情語の取りこぼしが起きている可能性がある．

今回我々は，口コミ内の感情語の取りこぼしを減らすため，辞書の拡張としてひらがなやカタカナのみの表現を加え，抽出できる単語数に変化があるかを比較する．辞書は漢字の表現のみを含んだ既存の辞書（以降，既存辞書と呼ぶ）と，既存の辞書に各語彙のひらがな表現とカタカナ表現を加えた辞書（以降，拡張辞書と呼ぶ）を用意する．

商品に関する口コミサイト価格.com（http://kakaku.com）のレビューページに投稿されている口コミに対して，二つの辞書を用い語彙の抽出を行った結果を表 1 に示す．口コミが 10 件以上投稿されている 7289 件の対象物に対し，抽出できた各カテゴリの単語の総数を表している．全てのカテゴリにおいて拡張辞書の方がより多くの単語を抽出できており，特に安心や驚きのカテゴリの単語に関しては約 4 倍の語数となっている．語彙によっては，口コミ内でひらがなやカタカナで使用される傾向が非常に強いことが，この結果からも分かる．

次に価格.com のレビューページについて，それぞれの解析結果の分布を表 2 に示す. 対象物 7289 件に対する口コミを全て収集して前節に示した表現手法を用い，各対象物の印象を示す特徴量を算出する．

表 2 は全ての対象物について算出された特徴量がどのような値を示したかを集計したもので，各カテゴリで特徴量がどのような値に分布しているかを知ることが出来る．例えば，喜びのカテゴリの特徴量が 100%，つまり最大値を取るような対象物は 2287 件あることを示しており，安心のカテゴリの特徴量が 69-60%を取るような対象物が存在しないことが分かる．既存辞書を用いた場合，好きのカテゴリが最大値になる場合が突出して多く，喜びのカテゴリは最大値を持つ場合か 30%周辺の値を持つ場合の 2 極があり，楽しさのカテゴリは 30%辺り，安心と驚きは 10%辺りを頂点として分布していることが分かる．

同様のサイトに対し，拡張辞書を用いた解析結果を表 3 に示す．ひらがなやカタカナで表記された感情語を抽出することが可能となり，分布に大きな変化が見られた．喜びのカテゴリを最大値にもつものが突出して多くなり，好きのカテゴリは 100%と 50%辺りの 2 つを極としてなだらかに分布するようになった．楽しさ，安心，驚きは，20%辺りを極とした似た分布となったが，既存辞書に比べ広く値が分布するようになった．

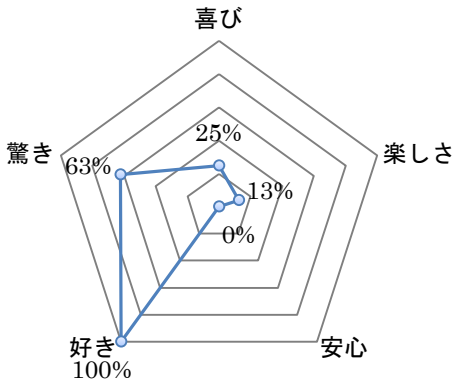

図 1 印象の可視化

表 1　各辞書の抽出単語数の違い

	既存辞書	拡張辞書
喜び	239730	485115
楽しさ	86293	163051
安心	34359	123400
好き	167177	276873
驚き	37921	160287

表 2　既存辞書による特徴量の分布（価格.com レビューページ）

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	2287	68	127	173	650	105	620	828	1504	927	0	0
楽しさ	2	2	0	2	6	15	102	1334	5793	23	10	0
安心	0	0	0	2	0	1	10	4	21	1095	5642	514
好き	5060	143	259	408	420	457	131	122	279	6	4	0
驚き	0	0	0	0	0	2	8	21	18	2042	4837	361

表 3　拡張辞書による特徴量の分布（価格.com レビューページ）

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	6193	140	212	177	151	143	88	66	58	41	8	12
楽しさ	236	48	98	110	207	377	524	859	1568	1795	885	582
安心	130	41	76	102	198	315	475	852	1728	2025	851	496
好き	817	164	353	470	563	880	877	967	917	743	314	224
驚き	160	22	80	117	273	516	741	1267	1958	1379	433	343

表 4　誤判定される感情語の例

登録した単語	元の感情語	操作
どう	動	削除
アンド	安堵	削除
かんじ	莞爾	削除
とうぜん	陶然	削除
楽（らく）	楽	楽しさ→安心へ移動

抽出された単語を確認すると「たのしい」や「ウレシイ」などの表現が多く抽出されており，感情を柔らかく抽象的にあらわすためにひらがなやカタカナが多く用いられていることが予想される．拡張辞書を用いることでロコミにより適した感情語抽出が可能であることが示されたと考えられる．

3.2 誤判定に関する調査

拡張辞書がロコミの分析により適していることは示されたが，ひらがなやカタカナで表現された単語は漢字を含んだ単語に比べ，より多様な意味の別の単語として読み取ることが可能であり，分解されたロコミ文章の中で全く意味の異なる形態素をより多く誤判定してしまう可能性がある．拡張辞書を用いた解析結果を確認したところ，幾つかの単語について誤って抽出している場合があることが分かった．

表 4 に誤って判定された感情語の中で特に出現数が多い単語五つを例として示す．「どう」に関しては最も出現数が多く主に「どうでしょうか」，「どうしたらいいか分からない」など副詞となる「どう」が抽出される場合が目立った．「アンド」では“&”をカタカナで表される場合に誤判定が発生していた．「かんじ」や「とうぜん」も「感じ」「当然」のひらがな表現を誤判定する場合が多いことが分かった．また，「楽」に関してはひらがな表現ではないが，ロコミ内では　「こんなに楽なんだ」，「使えば楽ですね」のような使い方が多く，「楽しい」に関する感情と言うよりは「安心」を示す感情語として出現している場合が多く見受けられた．

このような誤判定を避けるため，本来の意味として検出されることのほとんどないひらがなやカタカナ表現の単語は拡張辞書から削除し，ロコミにおいては使用される意図の異なる単語についてはカテゴリの振替を行った修正辞書を構成する．新たに構成した修正辞書による解析結果を表 5 に示す．誤判定する単語を辞書から削除したため，抽出される単語の総数は減ったが，より正確な感情語抽出が可能となったと考えられる．一部の対象物の特徴量に影響はあるものの，全体的な分布については大きな変化が見られなかった．

3.3 偏りの大きな分布に対する表現の検討

表 5 の分布によると，価格.com の対象物のほとんどは，喜びのカテゴリの特徴量が最大値のグラフとなる．可視化を前提とした場合，対象物の個性が際立つように表現されることが望ましいが，特定のサイトに対して特徴量の分布を操作するような辞書を用いることは好ましくない．本節では，一部のカテゴリの感情語の抽出量が際立って多く，相対的に他のカテゴリの特徴が現れにくくなるような状況に対して適切な表現方法について検討を行う．

価格.com のレビューページのロコミの特徴を調査した．喜びのカテゴリとして抽出される感情語のうち，抽出数の多い五つの単語を表 6 に示す．特に「満足」という単語の抽出数が突出しており，次に多い「感動」に比べ約 10 倍の量となっている．対象としているサイトは商品に対する感想が投稿されるため，その使用感や満足度を表す文章が多いことが原因であるが，この「満足」という 1 単語に

表 5　修正辞書による特徴量の分布（価格.com レビューページ）

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	6270	139	183	175	128	130	78	62	66	39	7	12
楽しさ	162	24	50	57	83	171	269	507	1119	1779	1809	1259
安心	179	42	70	99	201	320	486	862	1725	2001	808	496
好き	809	170	288	433	561	823	873	950	952	794	368	268
驚き	90	20	26	44	96	202	353	730	1674	2354	974	726

表 6　喜びのカテゴリ内の頻出感情語

感情語	出現数
満足	356376
感動	34540
嬉しい	26972
明るい	24725
すっきり	16074

よって他の感情語の量の違いによって得られるはずの個々の対象物の特徴が可視化の際に少ない変化量として表現されてしまい，個性が埋没してしまうことになる．サイトによって口コミ投稿者が頻繁に使用する用語の中に感情語の一部が含まれ,その感情語は一般的な単語となってしまい，文章の特徴を表す単語として機能しなくなるのである．

このような場合，一般化した感情語は多様な個性を表す用語としてではなく，個別の指標として用いることで適切な表現が可能となると考えられる．例えば，価格.com のレビューページの場合，満足という単語の個数は別に数え商品の満足度を表す指標とし，満足という単語を除いた辞書を用いてカテゴリの特徴量とすることで，その商品の個性の表現とするのである．

表 7 に「満足」「まんぞく」「マンゾク」の三つの単語を喜びのカテゴリの辞書から除いて特徴量を求めた結果を示す．喜びのカテゴリの感情語の抽出数が他のカテゴリに近い値となったことで，表 5 に比べ，どのカテゴリも分布がなだらかになり，対象物の個性の違いが分かりやすくなったと考えられる．

3.4 傾向の異なる口コミサイトの分布の違い

前節の調査から，取り扱う対象物によって，口コミ内で使用される感情語の一部は他の感情語と同じように機能しないことが分かった．扱う対象物が異なれば，同様に特徴量が偏る別の単語が存在することが考えられる.本節では，扱う対象物の異なる幾つかのサイトについて，分布の違いを調査する．今回調査したサイトは以下の 3 つのサイトである．

- トリップアドバイザー（https://www.tripadvisor.jp/）
- 4travel（http://4travel.jp/）
- 食べログ（http://tabelog.com/）

トリップアドバイザーと 4travel は旅行に関する情報を取り扱うサイトでありよく似た傾向のサイトと考えられる．食べログは飲食店に関する情報を扱っている．各サイトを解析した結果を　表 8, 9, 10 に示す．解析には表 5 と同様に修正辞書を使用した．

表 8, 9 を比較するとわかるように，トリップアドバイザーと 4travel については，非常によく似た分布となっている。また，観光に関する情報が対象となるため，楽しさのカテゴリが最大値を取る対象物が多いが，全体としては特定のカテゴリの特徴量が突出することもなく，偏りの少ない分布と言える．今回構築した修正辞書は，旅行サイトの口コミ分析に対して特別な単語の処置を行わなくとも有効に機能すると考えられる．

表 10 に示した食べログの分布は,これまで調査したサイトとはまた異なる傾向を示した．喜びと好きのカテゴリに何かを最大値に持つ対象物が多く，他の三つのカテゴリはどれも 20%辺りを中心に低い値をとるものが多いことがわかる．

表 7　「満足」を除いた辞書による価格.com レビューページの特徴量の分布

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	1907	225	483	563	724	843	617	624	546	342	117	298
楽しさ	416	56	128	194	309	518	520	858	1245	1248	587	1210
安心	1258	85	269	311	498	730	646	915	1047	757	281	492
好き	3970	163	383	353	436	489	298	335	326	213	63	260
驚き	633	68	205	241	461	720	732	1033	1270	894	324	708

表 8　トリップアドバイザーの特徴量の分布

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	2679	70	236	290	504	753	268	523	518	239	52	1157
楽しさ	3128	69	238	282	438	727	257	449	414	193	30	1064
安心	748	9	57	105	237	552	240	534	737	763	365	2942
好き	2024	46	180	243	478	743	373	578	663	379	80	1502
驚き	820	11	76	97	275	599	305	617	824	740	228	2697

表 9　4travel の特徴量の分布

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	511	25	71	81	139	207	105	165	143	88	23	335
楽しさ	931	21	58	62	117	174	65	89	75	49	11	241
安心	182	2	21	29	44	133	62	119	214	250	135	702
好き	486	10	57	85	126	197	116	160	166	101	18	371
驚き	205	4	19	29	69	158	93	180	252	225	84	575

表 10　食べログの特徴量の分布

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	11243	1257	2686	2945	3410	3834	2894	2555	1891	894	171	359
楽しさ	941	106	475	650	1301	2367	2910	4793	7226	6594	2880	3896
安心	495	42	267	405	864	1694	2231	4390	7989	8206	3521	4035
好き	23044	1073	2132	1848	1841	1689	856	713	549	198	21	175
驚き	164	0	63	89	279	546	613	1549	4412	9913	7813	8698

表 11　食べログの特徴量の分布（3 カテゴリ化）

	100%	99-90%	89-80%	79-70%	69-60%	59-50%	49-40%	39-30%	29-20%	19-10%	9-1%	0%
喜び	9580	2626	3721	3978	4337	4029	2718	1887	930	273	53	7
好き	18677	2370	3239	2896	2677	2126	1121	657	268	90	12	6
くつろぎ	7910	2423	4026	4462	4817	4528	2996	1935	800	198	40	4

特徴量を可視化した場合，二つのカテゴリのみが突出した似た形状のグラフが多くなるだろう．抽出した感情語を確認したところ極端に抽出数の多い単語も見受けられなかった．このような場合の表現方法としては，一部のカテゴリを統合し，抽出数の差を少なくすることで無駄のない表現に変えることが可能になると考えられる．

例えば，楽しさと安心，驚きのカテゴリを一つのカテゴリ，ここではくつろぎのカテゴリと定義する．三つのカテゴリに含まれる感情語の抽出数の合計をくつろぎカテゴリの抽出数として特徴量の計算を行う．その結果を表 11 に示す．好きのカテゴリを最大値となる対象物は少し多いが，三つのカテゴリそれぞれ様々な値をなだらかに取るような結果となった．それぞれのカテゴリが対象物の印象を表す指標として有効に機能する表現となったと言えるだろう．

4 まとめ

本研究では，口コミを分析し口コミサイトで扱われている対象物の印象をよりわかりやい形で表現するために必要な，我々が提案した分析用の辞書について，異なる傾向を持つ幾つかの口コミサイトに対して有効に機能するかを検証した．

辞書の一部の語彙は，口コミ内ではより砕けた表現として，ひらがなやカタカナで表記されることが多いことや，一部の感情語や使われ方によって分類が異なる場合があることが確認され，それらの問題点を考慮することで，既存の辞書よりも口コミの分析に適した辞書を構築することができた．また，一部の感情語は，サイトの特性によって，一般的な単語に近くなり，対象物の個性を表す語として他の単語と同列に扱うことが難しいことも分かった．そのような単語は事前に別処理を行うことで，対象物の特徴を現す要素として使用するとともに，対象物の印象をより明確にすることが可能となった．今後は，これらの処置を組入れた，新たな解析システムを構築し，実用性の検証などを行いたいと考えている．

参考文献

[1] 杉本祐介，土井千章，中川智尋，太田賢，稲村浩，水野忠則，菱田隆彰：口コミデータを活用するデータベースシステムの実現, 情報処理学会研究報告. モバイルコンピューティングとユビキタス通信（MBL），2014-MBL-70(44), pp 1-6, 2014.

[2] 杉本祐介，水野忠則，菱田隆彰：口コミに含まれる感情語を利用した観光地分類の検討，マルチメディア、分散協調とモバイルシンポジウム 2014（DICOMO 2014），pp. 1345-1350, 2014.

[3] 杉本祐介，佐藤太一，土井千章，中川智尋，太田賢，稲村浩，内藤克浩，水野忠則，菱田隆彰：口コミを利用したレコメンドに適した感情語の分類方法の検討，モバイルコンピューティングとユビキタス通信（MBL），2015-MBL-74(50), pp 1-6, 2015.

[4] 松井瑠偉人，上谷竜士，梶克彦，内藤克浩，水野忠則，菱田隆彰：感情を活用した口コミ解析システムの実装，情報学ワークショップ WiNF2015 講演論文集，pp.10-14，2015.

[5] 松浦有容，渥美幸雄：感情表現による書評情報の可視化手法の提案と実装，専修大学情報科学研究所所報 (78)，pp.11-28，2012.

[6] 中村明：感情表現辞典，東京堂出版，1993.