The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

ソーシャルメディアからの印象抽出に基づく類似エリア判定手法の提案

Proposal of Similar Area Discovery Based on Impression Extracted from Social Media

高間 康史 Yasufumi Takama[*1]　坂元 陽亮 Yosuke Sakamoto[*1]　小林 賢一郎 Kenichiro Kobayashi[*2]　橋本 武彦 Takehiko Hashimoto[*2]

[*1]首都大学東京大学院システムデザイン研究科 Graduate School of System Design, Tokyo Metropolitan University　[*2]株式会社 GA technologies GA Technologies

This paper proposes a method for finding similar areas in terms of impression. Nowadays, social media is one of our important resources, from which we can obtain subjective information, such as the impression of some products and places. Such information is expected to be useful when looking for the places to live. The proposed method constructs an emotional word dictionary and uses it to extract the impression of the area around a station from reviews. The similarity between stations is calculated based on the extracted impression. This paper compares the quality of dictionaries constructed with different methods and shows the accuracy of the similarity judgment.

1. はじめに

本稿では，不動産分野における印象を用いたエリア推薦を目的として，レビューから印象を抽出するための感情語辞書構築手法，及び類似エリア判定手法を提案する．ソーシャルメディアは製品や場所などに対する主観的意見が得られる貴重な情報源であり，居住地を探す際にも有益と考えられる．提案手法では，駅周辺に対して投稿されたレビューから印象を抽出し，その結果に基づき印象の類似する駅を推定する．感情表現辞典に対し，類義語やブログデータなどからブートストラップにより抽出した語を追加して感情語辞書を構築し，辞書とレビューとのマッチングによって印象を計算する．ラベル付きレビューを用いた印象抽出精度評価と，クラウドソーシングによる類似度判定精度評価を通じて，提案手法の有効性を検証する．

2. 関連研究

テキストからの感情・印象抽出は，様々な感情分類手法と抽出手法の組み合わせによる手法が提案されている．山本らは，感情表現辞典 [中村 93] の 10 感情を用いて，文章データセットにおける語の共起関係から感情語辞書を作成し，Twitter に投稿されたツイートからの辞書マッチングによる感情抽出手法を提案している [山本 14]．地域情報の利用に関して，Yao らは周辺地域の安全性が家の価値に影響を与えるとの考えに基づき，犯罪履歴などの統計データに基づく価値推定手法を提案している [Yao16]．

3. 提案手法

3.1 感情語辞書の作成

提案手法で使用する感情語辞書は，感情表現辞典 [中村 93] をベースに構築する．本書で定義されている 10 感情から「恥」を削除,「怒」「昂」を「驚」に併合して，7 感情として感情語辞書を作成する．本辞書は多くの関連研究で利用されているが，出版が 1993 年であり，現代的な文章の多いレビューで使用される語の多くが収録されていないことが想定されるため，類義語およびブログデータなどからブートストラップにより抽出した語を追加することで現代語に対処する．

表 1: 印象と感情の対応関係

印象	感情 (positive)	感情 (negative)
好感度	好	厭
安心度	安	怖
興奮度	驚	
楽しさ	喜	
哀しさ	哀	

類語辞典として日本語 WordNet と Weblio 類語辞典を用い，感情表現辞典中の感情語に対する類義語に，元の感情語と同じ感情ラベルを付与して感情語辞書に追加する．このとき，類義語抽出の精度向上のため，相互類義関係にある語句のみを感情語辞書に追加する．

ブートストラップ [Hearst92] は，少数の語をシードとして利用して，大規模な文章データから同位語などを取得するアプローチである．提案手法では，IDR で提供されている不満調査データセット [*1] と芸能人のブログから抽出したブログデータから，反復回数 1 回のブートストラップにより語を抽出する．

3.2 印象抽出

エリアに関するレビューと辞書とのマッチングによって，エリアの印象を抽出する．提案手法では，感情と印象の対応を表 1 の様に定義する．ここで,「興奮度」,「楽しさ」,「哀しさ」については，対として表現することが難しいと考え，ポジティブな感情のみで定義している．

辞書マッチングでは，レビューを形態素に分割し，各感情語の重みと出現回数の積から印象別のスコア（印象値）をレビューごとに求める．感情語の印象に対する重みは，表 1 のネガティブな感情を持つ感情語は −1，ポジティブな感情を持つ感情語は 1 を初期値とし，係り受け解析による否定形や仮定形，程度副詞の抽出によって重みを更新する．エリアに対する各印象の強さは，該当レビューの印象値の総和により求める．

3.3 エリア推薦

本稿で提案するエリア推薦システムの構成を図 1 に示す．ユーザは入力として，ユーザの好む駅（クエリ駅）と検索範囲を入力する．入力された検索範囲内の各駅とクエリ駅間で印象ベクトルのユークリッド距離を計算し，類似度順に並べたリストをユーザに推薦する．

連絡先: 高間 康史，首都大学東京，ytakama@tmu.ac.jp

*1 https://www.nii.ac.jp/dsc/idr/fuman/fuman_top.html

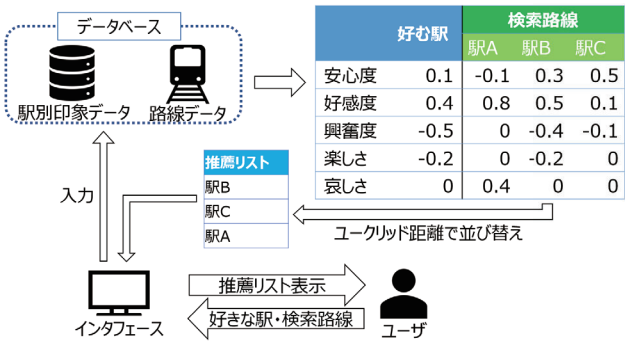

図 1: 提案システムの構成

表 2: 実験に利用した各辞書の説明

辞書名	説明
base	感情表現辞典
wd	日本語 WordNet による base の拡張
wb	Weblio 類語辞典による base の拡張
jw	感情表現辞書 [柴田 17]
+b	ブートストラップによる拡張

4. 評価実験

4.1 印象抽出精度

本実験では，駅に関するレビュー 500 件に印象ラベルを付与して正解データを作成し，提案手法による印象抽出結果と比較を行った．レビューにつき 3 名で，各印象の強さを 7 段階でラベル付けし，多数決で正解データを決定した．多数決で決まらない場合は中央値を採用した．

表 2 に示す各辞書を用いてレビューに対する印象値を計算し，その値に基づきレビューごとに印象を降順に並べた結果と，正解データにおける印象ラベルの順序からスピアマンの順位相関係数を求めた結果を表 3 に示す．

表より，各拡張手法による語彙追加によって相関を示すレビュー数が増加していることがわかる．特に，wb+b では全体の約 50%のレビューが相関を示している．この結果から，提案手法による辞書拡張は，印象抽出精度の向上に有効といえる．一方で，逆相関を示すレビューが約 30%存在するため，否定形・仮定形の抽出部分の改善や，感情と印象の対応関係の再検討が必要と考える．

4.2 類似度判定精度

印象に基づく類似駅の判定精度を評価する実験を行った．正解データはクラウドソーシングを利用して作成した．各ターゲット駅に対し，提案手法による印象が類似する上位 10 駅と，それ以外の駅 10 駅を提示して，類似すると思う駅を任意数選択してもらった．回答は 97 人から得られた．ターゲット駅及び選択肢となる駅は，東京都内でレビュー数の多い駅，あるいは乗客数の多い駅の中から抽出し，印象抽出の結果に基づき「原宿」「新橋」「水道橋」「代々木」「目黒」「蒲田」の 6 駅をターゲット駅として採用した．また，感情語辞書は前節に示した結果が最も良好な wb+b を用いた．

表 3: 各辞書を用いた実験結果による相関別レビュー件数

辞書	base	wd	wd+b	wb	wb+b	jw	jw+b
相関	175	240	248	238	**266**	162	186
無相関	**275**	98	136	130	118	122	104
逆相関	50	107	116	132	116	**216**	210

表 4 に原宿・代々木・水道橋の適合率・再現率を求めた結果を示す．選択人数が N 人以上の駅を正解データ，提案手法の推薦リストは上位 L 件として評価を行っている．結果から，原宿駅では $N = 15$ の場合を除いて，L に関わらず適合率 0.5 以上となることがわかる．また，$L = 10$ の時に再現率は大きくなるが，適合率はそれほど低下しないことがわかる．蒲田駅でも，同様の結果が得られた．

代々木駅では $N = 5$ の場合，L に関わらず適合率 0.6 以上となることがわかる．再現率を見ると，L に関わらず，$N = 10$ の場合に最小となっていることから，多数の人が類似するとした駅と，少数の人だけが類似するとした駅の両方を推薦できているといえる．新橋，目黒駅でも代々木駅と同様の結果が得られた．

水道橋駅では $N = 5$ の場合を除いて，適合率 0.2 以下，再現率 0.3 以下と小さい値になっている．これらの結果から，水道橋駅以外では，ある程度良い精度で推薦ができていると考える．また，選択人数が少ない駅は，実際にその駅を利用していた人のみが類似すると判断した駅である可能性が考えられる．その場合，これらの駅は推薦された人にとって，意外かつ有用な推薦になる可能性があると考える．

表 4: 各ターゲット駅の適合率・再現率

		原宿		代々木		水道橋	
L	N	P	R	P	R	P	R
3	5	1.00	0.21	0.67	0.12	0.33	0.06
3	10	1.00	0.43	0.33	0.11	0.00	0.00
3	15	0.33	0.25	0.33	0.20	0.00	0.00
5	5	1.00	0.36	0.80	0.24	0.60	0.18
5	10	0.60	0.43	0.20	0.11	0.00	0.00
5	15	0.20	0.25	0.20	0.20	0.00	0.00
10	5	0.90	0.64	0.90	0.53	0.80	0.47
10	10	0.50	0.71	0.30	0.33	0.20	0.25
10	15	0.30	0.75	0.20	0.40	0.00	0.00

5. おわりに

本稿では，エリアに関するレビューから印象を抽出するための感情語辞書構築手法，及び類似エリア判定手法を提案した．印象抽出精度の評価実験から，提案する各辞書拡張手法によって印象抽出精度が向上することを示した．また，類似駅判定についても，ある程度精度の良い推薦が可能であることを示した．今後は，エリアに関する統計データに基づく類似度判定との特性比較や統合を検討する予定である．

謝辞

本研究の一部は JSPS 科研費 16K1253500 の助成を受けたものです．

参考文献

[中村 93] 中村明, 感情表現辞典, 東京堂 出版, 1993.

[山本 14] 山本湧輝, 熊本忠彦, 灘本明代, Twitter 特有表現を考慮したツイートの多次元感情抽出手法の提案, 情報処理学会関西支部大会, G-01, 2014.

[Yao16] Z. Yao, Y. Fu, B. Liu, H. Xiong, The Impact of Community Safety on House Ranking, SIAM Int'l Conf. on Data Mining, pp. 459–467, 2016.

[Hearst92] M. A. Hearst, Automatic acquisition of hyponyms from large text corpora, COLING'92, Vol. 2, pp. 539–545, 1992.

[柴田 17] 柴田大作, 若宮翔子, 伊藤薫, 荒牧英治, クラウドソーシングによる日本語感情表現辞書の構築, NLP2017, pp. 771–774, 2017.